Tight worst-case loss bounds for predicting with expert advice

David Haussler*
Jyrki Kivinen†
Manfred K. Warmuth‡

UCSC-CRL-94-36
November 3, 1994
(Revised December 8, 1994)

Baskin Center for
Computer Engineering & Information Sciences
University of California, Santa Cruz
Santa Cruz, CA 95064 USA

ABSTRACT

We consider on-line algorithms for predicting binary or continuous-valued outcomes, when the algorithm has available the predictions made by N experts. For a sequence of trials, we compute total losses for both the algorithm and the experts under a loss function. At the end of the trial sequence, we compare the total loss of the algorithm to the total loss of the best expert, i.e., the expert with the least loss on the particular trial sequence. Vovk has introduced a simple algorithm for this prediction problem and proved that for a large class of loss functions, with binary outcomes the total loss of the algorithm exceeds the total loss of the best expert at most by the amount $c \ln N$, where c is a constant determined by the loss function. This upper bound does not depend on any assumptions on how the experts' predictions or the outcomes are generated, and the trial sequence can be arbitrarily long. We give a straightforward alternative method for finding the correct value c and show by a lower bound that for this value of c, the upper bound is asymptotically tight. The lower bound is based on a probabilistic adversary argument. The class of loss functions for which the $c \ln N$ upper bound holds includes the square loss, the logarithmic loss, and the Hellinger loss. We also consider another class of loss functions, including the absolute loss, for which we have an $\Omega\left(\sqrt{\ell \log N}\right)$ lower bound, where ℓ is the number of trials. We show that for the square and logarithmic loss functions, Vovk's algorithm achieves the same worst-case upper bounds with continuous-valued outcomes as with binary outcomes. For the absolute loss, we show how bounds earlier achieved for binary outcomes can be achieved with continuous-valued outcomes using a slightly more complicated algorithm.

Keywords: worst-case loss bounds, on-line learning, learning theory

---

*Supported by NSF grant IRI-9123692; e-mail haussler@cse.ucsc.edu.

†Funded by the Academy of Finland; e-mail kivinen@cse.ucsc.edu

‡Supported by NSF grant IRI-9123692; e-mail manfred@cse.ucsc.edu.

1 Introduction

Consider an on-line prediction problem in which the prediction algorithm is to predict a sequence of outcomes y_t, $t = 1, \ldots, \ell$. In the usual learning approach, the algorithm is provided with *instances* z_t. At trial t, the algorithm sees the instance z_t, must then give its *prediction* $\hat{y}_t$ of the outcome, and finally sees the actual outcome y_t. The algorithm is charged a loss if its prediction differs from the actual outcome, and its goal is to minimize its total loss over a sequence of ℓ trials. To make the algorithm's task feasible, some sort of relationship is assumed to exist between the instance z_t and the outcome y_t.

The on-line prediction problem considered in this paper is somewhat different from the one just described. Assume that there are N experts $\mathcal{E}_i$, $i = 1, \ldots, N$, each trying to predict the outcomes y_t as best they can. Let $x_{t,i}$ be the prediction of the ith expert $\mathcal{E}_i$ about the tth outcome. We make no assumptions about how the experts' predictions $x_{t,i}$ are generated. Perhaps the experts are different on-line learning algorithms that use the instances z_t to predict y_t, or perhaps each expert is a human with access to some private information not available to the other experts. We give as input to our algorithm at trial t the *prediction vector* $\mathbf{x}_t$ that consists of the predictions of the experts at that trial. The algorithm does not see the data used by the experts to generate their predictions, and is thus entirely dependent on the quality of the expert advice contained in the prediction vector. Therefore, to predict nearly as well as the best expert is a reasonable goal for the algorithm.

Formally, an on-line prediction algorithm is for us an algorithm that generates at trial t its prediction $\hat{y}_t$ based on the prediction vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t$ and the earlier outcomes $y_1, \ldots, y_{t-1}$. We take the predictions of both the algorithm and the experts, as well as the outcomes, to be real numbers in $[0, 1]$. The performance of a learning algorithm is measured using a *loss function* L, which is a mapping from $[0, 1] \times [0, 1]$ to $[0, \infty)$; sometimes also the value ∞ is allowed. The square loss, L_{sq}, defined by $L_{\mathrm{sq}}(p, q) = (p - q)^2$, is a typical loss function. At trial t, a learning algorithm A suffers a loss $L(y_t, \hat{y}_t)$. Over the whole trial sequence $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$, the algorithm attempts to achieve a small *total loss* $\mathrm{Loss}_L(A, S) = \sum_{t=1}^{\ell} L(y_t, \hat{y}_t)$. Similarly, the total loss of the ith expert over the trial sequence is given by $\mathrm{Loss}_L(\mathcal{E}_i, S) = \sum_{t=1}^{\ell} L(y_t, x_{t,i})$. Then $\min_{1 \le i \le N} \mathrm{Loss}_L(\mathcal{E}_i, S)$ gives the loss of the *best expert* on the particular sequence S. As explained, we require the algorithm to predict almost as well as the best expert. Specifically, we require that the *additional loss* $\mathrm{Loss}_L(A, S) - \min_{1 \le i \le N} \mathrm{Loss}_L(\mathcal{E}_i, S)$ is small for all sequences S. We do not make assumptions about how the experts' predictions are generated, or how the outcomes y_t relate to the prediction vectors $\mathbf{x}_t$. The only allowance we make for the algorithm is that it can make a large loss if none of the experts is good. Our framework for on-line prediction is based on the work of Vovk [16, 17] and Cesa-Bianchi et al. [2]. Similar frameworks have also been considered by Cover [6], Dawid [7], Feder et al. [9, 14, 19], and Mycielski [15]. See Chung [5] for recent related results.

In this paper, we start by considering the case in which the outcomes are binary, i.e., $y_t \in \{0, 1\}$ for all t. The predictions $\hat{y}_t$ of the algorithm and $x_{t,i}$ of the experts are still allowed to range continuously from 0 to 1. Thus, the algorithm could predict with $\hat{y}_t$ close to $1/2$ to avoid committing itself too strongly to either possible outcome $y_t = 0$ or $y_t = 1$. We later see how the results can be generalized for continuous-valued outcomes $y_t \in [0, 1]$. Cesa-Bianchi et al. [3] have considered the case in which both the outcomes and the predictions of the experts and the algorithm are required to be binary.

It turns out that for a large class of loss functions, such as the square loss, logarithmic loss, and absolute loss, the worst-case upper bounds for the additional loss are the same both

for binary and continuous-valued outcomes. Further, for the square and logarithmic loss, the algorithm for binary outcomes works for continuous outcomes, as well.

We are interested in what bounds for the worst-case additional loss are possible for different loss functions. Vovk [16] introduced a general on-line prediction algorithm that is applicable for all loss functions when the outcomes are binary. Vovk's analysis allows for a more general setting than the one we consider; for instance, the predictions may be restricted to some discrete set. For the case with continuous-valued predictions, which we consider here, Vovk proved for a large class of loss functions bounds of the form

$$\mathrm{Loss}_L(A,S) - \min_{1\le i\le N} \mathrm{Loss}_L(\mathcal{E}_i,S) \le c_L \ln N \quad , \tag{1.1}$$

where c_L is a positive constant determined by the loss function L. For instance, for the square loss Vovk's algorithm achieves the bound with $c_L = 1/2$ [16], and for logarithmic loss with $c_L = 1$ [8, 16]. Note that the bound (1.1) for the additional loss is independent of the length ℓ of the trial sequence S. On the other hand, for the absolute loss L_{abs} given by $L_{\mathrm{abs}}(y_t,\hat{y}_t) = |y_t - \hat{y}_t|$ Cesa-Bianchi et al. [2] have shown that bounds of the form (1.1) are not obtainable, but the best possible algorithm has a worst-case bound of the form $\mathrm{Loss}_L(A,S) - \min_{1\le i\le N} \mathrm{Loss}_L(\mathcal{E}_i,S) = \Theta\left(\sqrt{\ell \log N}\right)$. Slightly weaker results for the absolute loss were obtained already by Littlestone and Warmuth [13].

In this paper, we give a simplified version of Vovk's analysis in the case that the predictions can range continuously in $[0,1]$. This gives a straightforward method for obtaining the value c_L in (1.1). The value c_L itself is the same as implied by Vovk's results. Further, we see that our method gives optimal values for the constant c_L. That is, we show that if c_L is chosen appropriately, we have not only the upper bound (1.1) for all trial sequences S, but also for some trial sequence S the lower bound

$$\mathrm{Loss}_L(A,S) - \min_{1\le i\le N} \mathrm{Loss}_L(\mathcal{E}_i,S) \ge (c_L - o(1)) \ln N \quad , \tag{1.2}$$

where $o(1)$ is a quantity that approaches 0 as N and ℓ approach ∞. Hence, for the class of loss functions that satisfies our conditions, we have an asymptotically tight bound for the worst-case additional loss.

The conditions the loss function must satisfy for the bounds (1.1) and (1.2) to hold are natural and can easily be seen to be satisfied by most usual loss functions, except for the absolute loss. We also define another class of loss functions, including the absolute loss, for which we can prove the lower bound

$$\mathrm{Loss}_L(A,S) - \min_{1\le i\le N} \mathrm{Loss}_L(\mathcal{E}_i,S) = \Omega\left(\sqrt{\ell \log N}\right) \quad .$$

Hence, for the loss functions in this class, an upper bound like (1.1), with no dependence on the length ℓ of the trial sequence, cannot be achieved.

It is possible to construct loss functions that are in neither of our classes, and for which we thus do not know any bounds. It is an open problem to provide upper and lower bounds that would apply to all loss functions.

The asymptotically tight loss bounds are given in Subsection 3.1 together with a discussion of the condition the loss function must satisfy for the bounds to be applicable. Subsection 3.2 restates Vovk's algorithm and upper bound proof simplified for our purposes. The lower bound proof, given in Subsection 3.3, is based on generating the trial sequence by a simple randomized adversary and showing that already the expected loss of the algorithm tightly approaches the

upper bound implied in (1.1) for the worst-case loss. Thus, in a sense we see that in our particular setting, the average case is almost as difficult as the worst case. The proof technique with a randomized adversary was used by Cesa-Bianchi et al. [2] in the special case of the absolute loss.

Finally, in Subsection 4.1 we show that for certain loss functions, such as the square and logarithmic loss, Vovk's algorithm achieves the same worst-case loss bound even if the outcomes are allowed to be continuous-valued. For the absolute loss, the worst-case bounds proven for binary outcomes [16, 2] can be achieved with continuous-valued outcomes by using a slightly more complicated algorithm, as we show in Subsection 4.2.

2 On-line prediction and loss bounds

We consider the performance of an on-line learning algorithm A over a sequence $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$ of ℓ trials. The sequence S is an *N-expert* trial sequence if the tth *prediction vector* $\mathbf{x}_t$ is in $[0,1]^N$ for $t = 1, \ldots, \ell$. We consider both *binary outcomes*, with the *outcomes* y_t either 0 or 1, and *continuous-valued outcomes*, with y_t any real number from the interval $[0,1]$. At trial t, the algorithm A produces its prediction $\hat{y}_t \in [0,1]$ as a function of the prediction vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t$ and the outcomes $y_1, \ldots, y_{t-1}$. The main algorithms considered in this paper make their predictions $\hat{y}_t$ independently of the length ℓ of the whole trial sequence, but in some situations we also consider how the algorithms can be fine-tuned if ℓ is known in advance.

The performance of the learner at trial t is measured by $L(y_t, \hat{y}_t)$, where L is a *loss function* with the range $[0, \infty)$, or sometimes $[0, \infty]$. For binary outcomes $y_t \in \{0, 1\}$ it suffices to consider the functions L_0 and L_1 defined by $L_0(y, \hat{y}) = L(0, \hat{y})$ and $L_1(\hat{y}) = L(1, \hat{y})$.

Example 2.1: The *relative entropy loss* L_{ent} is defined by $L_{\mathrm{ent}}(y, \hat{y}) = y \ln \frac{y}{\hat{y}} + (1-y) \ln \frac{1-y}{1-\hat{y}}$. By the usual convention $0 \ln 0 = 0$, this gives $L_0(\hat{y}) = -\ln(1-\hat{y})$ and $L_1(\hat{y}) = -\ln \hat{y}$ for $L = L_{\mathrm{ent}}$. In the binary case $y \in \{0, 1\}$, the relative entropy loss is better known as the *logarithmic loss*.

The *square loss* L_{sq} is defined by $L_{\mathrm{sq}}(y, \hat{y}) = (y - \hat{y})^2$. Hence, for $L = L_{\mathrm{sq}}$, we have $L_0(\hat{y}) = \hat{y}^2$ and $L_1(\hat{y}) = (1 - \hat{y})^2$.

The Hellinger loss L_{H} is given by $L_{\mathrm{H}}(y, \hat{y}) = \frac{1}{2}\left(\left(\sqrt{1-y} - \sqrt{1-\hat{y}}\right)^2 + \left(\sqrt{y} - \sqrt{\hat{y}}\right)^2\right)$. Hence, for $L = L_{\mathrm{H}}$ we have $L_0(\hat{y}) = 1 - \sqrt{1-\hat{y}}$ and $L_1(\hat{y}) = 1 - \sqrt{\hat{y}}$.

The absolute loss L_{abs} is given by $L_{\mathrm{abs}}(y, \hat{y}) = |y - \hat{y}|$, and we have $L_0(\hat{y}) = \hat{y}$ and $L_1(\hat{y}) = 1 - \hat{y}$ for $L = L_{\mathrm{abs}}$. □

It is worth noting some properties of the loss functions of Example 2.1, since these will be important later. In each case, the function L_0 is increasing and L_1 decreasing in $[0,1]$, so the loss $L(y, \hat{y})$ increases as the prediction $\hat{y}$ moves away from the outcome y. The functions L_0 and L_1 are differentiable, and by the previous remark, $L_0'(z) \geq 0$ and $L_1'(z) \leq 0$ for all z. Except for the absolute loss, the second derivatives $L_0''(z)$ and $L_1''(z)$ are positive for all z, which means that errors become progressively more expensive as the difference between the prediction and outcome increases.

Consider now a loss function L and an on-line prediction algorithm A. Let $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$ be an N-expert trial sequence, and let the prediction of the algorithm A at trial t of the sequence S be $\hat{y}_t$. We then have $\mathrm{Loss}_L(A, S) = \sum_{t=1}^{\ell} L(y_t, \hat{y}_t)$ as the loss of the algorithm and $\mathrm{Loss}_L(\mathcal{E}_i, S) = \sum_{t=1}^{\ell} L(y_t, x_{t,i})$ as the loss of the ith expert on the sequence S. We define

$$V_{L,A}(S) = \mathrm{Loss}_L(A,S) - \min_{1 \le i \le N} \mathrm{Loss}_L(\mathcal{E}_i, S)$$

to be the *additional loss* of the algorithm, i.e., the amount by which the loss of the algorithm exceeds the loss of the best expert. We let

$$V_{L,A}(N,\ell) = \sup \left\{ V_{L,A}(((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_\ell,y_\ell))) \mid \mathbf{x}_t \in [0,1]^N, y_t \in \{0,1\} \right\}$$

be the worst case amount of additional loss for A, when the outcomes in an N-expert trial of length ℓ are restricted to be binary. Finally, we let $V_L(N,\ell) = \inf_A V_{L,A}(N,\ell)$ be the best additional loss obtainable by an on-line prediction algorithm A. The goal of this paper is to study for general loss functions L what are the lowest additional losses $V_{L,A}(N,\ell)$ that can be obtained by an on-line prediction algorithm, and to generalize the results for continuous-valued outcomes $y_t \in [0,1]$. We are particularly interested in whether $V_{L,A}(N,\ell)$ can have an upper bound that is independent on ℓ. Such bounds have previously been proven for square loss and logarithmic loss when the outcomes are binary. For these loss functions there are algorithms that satisfy $V_{L,A}(N,\ell) \le \frac{1}{2}\ln N$ and $V_{L,A}(N,\ell) \le \ln N$, respectively [16, 8]. On the other hand, for the absolute loss it is known that no upper bound of this form exists, but the algorithm A that minimizes $V_{L,A}(N,\ell)$ has $V_{L,A}(N,\ell) = \Omega\left(\sqrt{\ell \ln N}\right)$ [2]. One of our results provides a formula from which the best possible upper bound for $V_{L,A}(N,\ell)$ can be obtained for a wide class of loss functions L. For example, we obtain $V_{L,A}(N,\ell) \le 2^{-1/2}\ln N$ if L is the Hellinger loss.

It is implicit in the definitions that any lower bounds for $V_L(N,\ell)$ hold even for algorithms that know the length ℓ of the trial sequence beforehand. Most of our upper bounds for $V_{L,A}(N,\ell)$ are achieved by an algorithm A that depends only on the loss function, not on ℓ. The exception is the upper bound for the absolute loss, as will be discussed in Example 3.14.

Our upper bounds for $V_{L,A}(N,\ell)$ are not based on probabilistic assumptions, but we use probabilistic techniques in the lower bound proofs. We use $\mathrm{E}[X]$ and $\mathrm{Var}[X]$ to denote the expected value and variance of a random variable X. If we want to emphasize the underlying probability measure P, we write $\mathrm{E}_{x\in P}[X(x)]$ and $\mathrm{Var}_{x\in P}[X(x)]$. The probability of an event φ according to a probability measure P is denoted by $\Pr_{x\in P}[\varphi(x)]$.

We use $\mathbf{N}_+$ to denote the set $\{1,2,3,\ldots\}$ of the positive integers and $\mathbf{R}$ to denote the set of real numbers.

3 Binary outcomes

3.1 Main results

The proofs of our upper and lower bounds require that the loss function satisfies certain constraints. We first state the main result with all the necessary restrictions and then discuss the meaning of these restrictions. First, given loss functions L_0 and L_1 that are twice differentiable, we define a function S by

$$S(z) = L_0'(z)L_1''(z) - L_1'(z)L_0''(z) \tag{3.1}$$

and a function R by

$$R(z) = \frac{L_0'(z)L_1'(z)^2 - L_1'(z)L_0'(z)^2}{S(z)} \ . \tag{3.2}$$

We then define a constant c_L by

$$c_L = \sup_{0<z<1} R(z) \ . \tag{3.3}$$

Our main result concerns the case where c_L is finite. When c_L is finite and the loss function satisfies certain other conditions, we can prove an upper bound $V_{L,A}(N,\ell) \le c_L \ln N$ and show that the bound is asymptotically tight.

Theorem 3.1: *Let L be a loss function such that $L_0(0) = L_1(1) = 0$, L_0 and L_1 are twice differentiable in $(0,1)$, and $L_0'(z) > 0$ and $L_1'(z) < 0$ for $0 < z < 1$. Assume that the constant c_L defined in (3.3) is finite and $S(z)$ defined in (3.1) is positive for $0 < z < 1$. Then there is an on-line prediction algorithm A for which*

$$V_{L,A}(N,\ell) \le c_L \ln N \quad . \tag{3.4}$$

Further, we have

$$V_L(N,\ell) \ge (c_L - o(1)) \ln N \quad , \tag{3.5}$$

where $o(1)$ denotes a quantity that approaches 0 as ℓ and N approach ∞.

The algorithm A that obtains the bound (3.4), as well as the proof of the bound, are already given by Vovk [16]. The algorithm makes its predictions independently of the length ℓ of the trial sequence. We give the algorithm and a simplified proof in Subsection 3.2. The lower bound (3.5) is based on a probabilistic proof that is given in Subsection 3.3. The lower bound holds also for algorithms that get knowledge of ℓ beforehand.

Example 3.2: Consider the loss functions of Example 2.1. For the logarithmic loss, $R(z)$ is identically 1, and therefore $c_L = 1$. For the square loss, we have $R(z) = 2z - 2z^2$, and hence $c_L = 1/2$. For the Hellinger loss, we have $R(z) = z\sqrt{1-z} + (1-z)\sqrt{z}$, and it is straightforward to show that $R(z)$ is maximized for $z = 1/2$. Hence, $c_L = 2^{-1/2}$. For the absolute loss, the denominator of $R(z)$ is identically 0, so $c_L = \infty$. □

If the function R defined in (3.2) is unbounded in $(0,1)$, and hence the value c_L is infinite, we do not have good general bounds for the achievable additional losses $V_{L,A}$. The special case of absolute loss was considered by Cesa-Bianchi et al. [2]. They show that for the optimal algorithm A we have $V_{L,A}(N,\ell) = \Theta\left(\sqrt{\ell \ln N}\right)$. For the absolute loss, the value c_L is infinite because the denominator $S(z)$ is 0 for all z. For the logarithmic loss, the square loss, and the Hellinger loss, the value $S(z)$ is positive for all z. As we shall soon explain, the sign of $S(z)$ is intimately connected with the uniqueness of the Bayes-optimal prediction in a certain probabilistic prediction game.

Let Q be a probability measure on $\{0,1\}$, with $\Pr_{y\in Q}[y=1] = q$. For a prediction $z \in [0,1]$, the *expected loss* for probability measure Q, or for *bias* q, is $\mathrm{E}_{y\in Q}[L(y,z)] = (1-q)L_0(z) + qL_1(z)$. Here we define $0 \cdot \infty = 0$. For example, for the logarithmic loss we have $L_0(1) = \infty$, but the expected loss for prediction 1 is defined to be 0 for bias 1. For other biases it would be infinite. A prediction z is *Bayes-optimal* for bias q if it minimizes the expected loss. Note that since we assume L_0 and L_1 to be continuous in a closed interval, the expected loss always has a minimum value at some z. This holds even if we allow infinite losses. If L_0 is increasing and L_1 decreasing, then the prediction 0 is Bayes-optimal for bias 0 and the prediction 1 for bias 1. If a value $0 < z < 1$ is a local extremum point for the expected loss, then

$$(1-q)L_0'(z) + qL_1'(z) = 0 \quad . \tag{3.6}$$

If $1 - q \neq 0$ and $L_1'(z) \neq 0$, this implies

$$\frac{q}{1-q} = -\frac{L_0'(z)}{L_1'(z)} \quad . \tag{3.7}$$

More generally, if either $L_0'(z)$ or $L_1'(z)$ is nonzero for a given value $z \in (0,1)$, then there is a unique value $q \in (0,1)$ for which (3.6) holds, and hence z cannot be a Bayes-optimal prediction for more than one bias. If

$$(1-q)L_0''(z) + qL_1''(z) > 0 \tag{3.8}$$

holds in addition to (3.6), then z is a local minimum point. There may be one or more Bayes-optimal predictions for a given bias.

Example 3.3: For the logarithmic and square losses, it is easy to show that $z = q$ is the unique Bayes-optimal prediction for bias q.

For the Hellinger loss, solving (3.7) shows that the unique Bayes-optimal prediction z for a bias $0 < q < 1$ is given by

$$z = \frac{1}{1 + \left(\frac{1-q}{q}\right)^2} \quad .$$

For the absolute loss, $z = 0$ is the unique Bayes-optimal prediction for biases $q < 1/2$ and $z = 1$ for biases $q > 1/2$. For the bias $q = 1/2$, any prediction is Bayes-optimal. □

Lemma 3.4: *If $S(z) > 0$ for all z, then for all biases $0 \le q \le 1$ there is a unique Bayes-optimal prediction z. If for all biases q the Bayes-optimal prediction is unique, then $S(z) \ge 0$ for all z, and there is no interval $[a,b]$ with $a < b$ such that $S(z) = 0$ for all $z \in [a,b]$.*

The proof of Lemma 3.4 is given in Subsection 3.3. In Subsection 3.3 we also prove the following lower bounds, which show that if the denominator S is not always strictly positive, the gap $V_{L,A}(N,\ell)$ cannot have an upper bound that is independent of ℓ.

Theorem 3.5: *Let L be a loss function such that L_0 and L_1 are twice differentiable in $(0,1)$, and $L_0'(z) > 0$ and $L_1'(z) < 0$ for all z. Let S be as in (3.1).*

1. *If $S(z) = 0$ for some $0 < z < 1$, we have*

$$V_L(N,\ell) = \Omega\left(\ell^{1/2-\alpha}\sqrt{\log N}\right) \tag{3.9}$$

   for all $\alpha > 0$.

2. *If $S(z) < 0$ for some $0 < z < 1$, or there are values $a < b$ such that $S(z) = 0$ for all $a \le z \le b$, we have*

$$V_L(N,\ell) = \Omega\left(\sqrt{\ell \log N}\right) \quad . \tag{3.10}$$

This generalizes the results of Cesa-Bianchi et al. [2] for the absolute loss.

Finally, it is possible to construct loss functions L for which the value c_L is infinite, but the denominator $S(z)$ is positive for all z. For such loss functions the results of this paper have no implications whatsoever.

Example 3.6: Define a loss function by $L_0(z) = (1-z)^{-\alpha} - 1$ and $L_1(z) = z^{-\alpha} - 1$ for some positive value α. We then have

$$R(z) = \frac{\alpha}{\alpha+1}\left(z^{-\alpha}(1-z) + (1-z)^{-\alpha}z\right) \quad .$$

Therefore, $R(z)$ approaches ∞ as z approaches 0 or 1, and c_L is infinite. Hence, our results give no upper bound for $V_L(N,\ell)$. However, the denominator $S(z)$ is given by

$$S(z) = \alpha^2(\alpha+1)(z(1-z))^{-\alpha-2}$$

and is hence strictly positive for $0 < z < 1$. Therefore, we have no lower bound, either. For this loss function it is an open problem to define the value $V_L(N,\ell)$.

Since $S(z)$ is positive, we know that the Bayes-optimal prediction z for each bias q is unique. Specifically, we have

$$z = \frac{1}{1 + \left(\frac{1-q}{q}\right)^{1/(\alpha+1)}} \ ,$$

as can be seen by a straightforward calculation. □

3.2 The algorithm and the upper bound

We consider an algorithm first introduced by Vovk. The algorithm has two positive real valued parameters c and η. We first introduce the algorithm in a somewhat open form, leaving the parameters c and η unspecified and defining the prediction $\hat{y}_t$ only by giving a condition it must satisfy. For the moment we also leave open the possibility that there is no prediction that satisfies the condition, in which case we say that the algorithm fails. The parameter c is can vaguely be characterized as a measure for the error allowed for the algorithm. The smaller the value c, the tighter upper bound we get for the additional loss assuming that the algorithm does not fail. Hence, for applying the algorithm we need to find the least value c for which the algorithm is guaranteed to never fail when the *learning rate* η is chosen suitably.

It turns out that for a loss function L that satisfies the assumptions of Theorem 3.1, the suitable choice is $c = c_L$ and $\eta = 1/c$. This gives a bound $V_{L,A}(N, \ell) \leq c_L \ln N$. The main part of the proof is in showing that for any choice $c \geq c_L$ the algorithm is guaranteed not to fail for $\eta = 1/c$. We also give a more direct way of choosing a prediction $\hat{y}_t$ that satisfies the required conditions, provided that such a prediction exists. Examples show that the seemingly complicated conditions for $\hat{y}_t$ are actually quite simple for the usual loss functions.

The algorithm uses an N-dimensional weight vector $\mathbf{w}_t = (w_{t,1}, \ldots, w_{t,N})$ as its internal state. The weight $w_{t,i}$ is always nonnegative and summarizes the performance of the ith expert in previous trials. At the end of the tth trial we have $-\ln w_{t,i} = \eta \mathrm{Loss}_L(\mathcal{E}_i, S_t)$, where S_t consists of the first t trials of S. Note that the weights $w_{t,i}$ are invariant under permutations of the trial sequence S_t. The predictions $\hat{y}_t$ of the algorithm are independent of the total length ℓ of the trial sequence.

Algoritm 3.7 (The Generic Algorithm): Let L be a loss function and c and η be any positive constants.

Initialization: Set the weights to some initial values $w_{1,i} > 0$.

Prediction: Let $v_{t,i} = w_{t,i}/W_t$, where $W_t = \sum_{i=1}^{N} w_{t,i}$. At the beginning of trial t, compute for $y = 0$ and $y = 1$ the value

$$\Delta(y) = -c \ln \sum_{i=1}^{N} v_{t,i} e^{-\eta L(y, x_{t,i})} \ . \tag{3.11}$$

On receiving the tth input $\mathbf{x}_t$, predict with any value $\hat{y}_t$ that satisfies for $y = 0$ and $y = 1$ the condition

$$L(y, \hat{y}_t) \leq \Delta(y) \ . \tag{3.12}$$

It no such value $\hat{y}_t$ exists, the algorithm fails.

Update: After receiving the tth outcome y_t, let

$$w_{t+1,i} = w_{t,i} e^{-\eta L(y_t, x_{t,i})}. \tag{3.13}$$

To understand the algorithm, note that by (3.11) and (3.13) we can write $\Delta(y_t) = U_{t+1} - U_t$, where $U_t = -c \ln W_t$. Hence, we can consider $-c \ln W_t$ as a potential function, and the condition $L(y_t, \hat{y}_t) \leq \Delta(y_t)$ means that at each trial, the increase of the potential must be at least as large as the loss of the algorithm. The basic idea of proving the upper bound for the loss of the Generic Algorithm is based on relating the total potential increase $U_{\ell+1} - U_1$ to the total loss of the best expert. The following upper bound was already given by Vovk [16].

Theorem 3.8: *Let L be any loss function. Let $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$ be an N-expert trial sequence in which the outcomes $y_t \in \{0, 1\}$ are binary. Assume that during this trial sequence, the Generic Algorithm 3.7 with parameters c and η does not fail but produces at each trial t a prediction $\hat{y}_t$. Then for all i the total loss satisfies*

$$\mathrm{Loss}_L(A, S) \leq -c \ln \frac{W_{\ell+1}}{W_1} \leq -c \ln \frac{w_{1,i}}{W_1} + c\eta \mathrm{Loss}_L(\mathcal{E}_i, S) \ . \tag{3.14}$$

Proof The condition (3.12) for $y = y_t$ together with (3.11) and (3.13) implies

$$L(y_t, \hat{y}_t) \leq -c \ln \frac{W_{t+1}}{W_t}$$

and hence

$$\sum_{t=1}^{\ell} L(y_t, \hat{y}_t) \leq -c \ln \frac{W_{\ell+1}}{W_1} \leq -c \ln \frac{w_{\ell+1,i}}{W_1}$$

for all i. Finally, by (3.13) we get

$$\frac{w_{\ell+1,i}}{W_1} = \frac{w_{1,i}}{W_1} \prod_{t=1}^{\ell} e^{-\eta L(y_t, x_{t,i})} \ ,$$

and the theorem follows. □

For given values c and η, we say that the loss function L is (c, η)-*realizable* if the condition (3.12) for $y = 0$ and $y = 1$ can always be satisfied by a suitable choice of $\hat{y}_t$. To prove the upper bound of Theorem 3.1, it now suffices to show that a loss function L that satisfies the assumptions of Theorem 3.1 is $(c, 1/c)$-realizable for $c = c_L$. The result then follows from Theorem 3.8 by setting $w_{1,i} = 1$ for all i. The rest of this subsection gives our formulation of Vovk's [16] proof for these results.

We first devolop an equivalent version of condition (3.12). Write $\Delta_0 = \Delta(0)$ and $\Delta_1 = \Delta(1)$, so the condition (3.12) for $y \in \{0, 1\}$ can be expressed as $L_0(\hat{y}_t) \leq \Delta_0$ and $L_1(\hat{y}_t) \leq \Delta_1$. To obtain explicit bounds for $\hat{y}_t$ from these conditions, we need to have some notion of an inverse for L_0 and L_1. Assume that L_0 is continuous and strictly increasing and L_1 is continuous and strictly decreasing in $[0, 1]$, which is implied by the assumptions of Theorem 3.1. Then L_0 has a continuous strictly increasing inverse L_0^{-1} that is defined in $[L_0(0), L_0(1)]$, and L_1 has a continuous strictly decreasing inverse L_1^{-1} that is defined in $[L_1(1), L_1(0)]$.

Consider first the case with $\Delta_0 \in [L_0(0), L_0(1)]$ and $\Delta_1 \in [L_1(1), L_1(0)]$. Then the values $L_0^{-1}(\Delta_0)$ and $L_1^{-1}(\Delta_1)$ are defined, and (3.12) for $y \in \{0, 1\}$ can be equivalently written as

$$L_1^{-1}(\Delta_1) \leq \hat{y}_t \leq L_0^{-1}(\Delta_0) \ . \tag{3.15}$$

A prediction $\hat{y}_t$ that satisfies (3.15) can be found if and only if

$$L_1^{-1}(\Delta_1) \leq L_0^{-1}(\Delta_0) \ . \tag{3.16}$$

If (3.16) holds, the prediction $\hat{y}_t$ can be chosen to be an arbitrary value between the bounds $L_1^{-1}(\Delta_1)$ and $L_0^{-1}(\Delta_0)$. For instance their mean $(L_1^{-1}(\Delta_1) + L_0^{-1}(\Delta_0))/2$ is a valid choice for $\hat{y}_t$.

Consider now the possibility that the value Δ_0 or Δ_1 is outside of the range of L_0 or L_1, respectively. If, for instance, Δ_0 is larger than $L_0(1)$, then the condition $L_0(\hat{y}_t) \leq \Delta_0$ in (3.12) holds for all $\hat{y}_t$. Thus, the equivalence between (3.12) and (3.15) will be maintained for all nonnegative Δ_0 if the inverse L_0^{-1} is extended in such a way that the condition $\hat{y}_t \leq L_0^{-1}(\Delta_0)$ holds for all $\hat{y}_t \in [0,1]$ when $\Delta_0 > L_0(1)$. Hence, we say that L_0^{-1} is a *generalized inverse* of L_0 if $L_0^{-1}(L_0(\hat{y})) = \hat{y}$ for all $\hat{y} \in [0,1]$ and $L_0^{-1}(\Delta_0) \geq 1$ whenever $\Delta_0 \geq L_0(1)$. Similarly, L_1^{-1} is a generalized inverse of L_1 if $L_1^{-1}(L_1(\hat{y})) = \hat{y}$ for all $\hat{y} \in [0,1]$ and $L_1^{-1}(\Delta_1) \leq 0$ whenever $\Delta_1 \geq L_1(0)$.

For instance, if L is the square loss $L_{\rm sq}$, we have the generalized inverses $L_0^{-1}(z) = \sqrt{z}$ and $L_1^{-1}(z) = 1 - \sqrt{z}$ for $0 \leq z \leq 1$, so (3.16) becomes

$$\sqrt{\Delta_0} + \sqrt{\Delta_1} \geq 1 \ .$$

For the relative entropy loss $L_{\rm ent}$ we have $L_0^{-1}(z) = 1 - e^{-z}$ and $L_1^{-1}(z) = e^{-z}$, so we get

$$e^{-\Delta_0} + e^{-\Delta_1} \leq 1 \ .$$

For the absolute loss $L_{\rm abs}$ we have $L_0^{-1}(z) = z$ and $L_1^{-1}(z) = 1 - z$, so we need to have

$$\Delta_0 + \Delta_1 \geq 1 \ .$$

Our definitions of generalized inverses let us show the equivalence between (3.15) and (3.12) for all values of Δ_0 and Δ_1.

Lemma 3.9: *Assume that L is a loss function such that $L_0(0) = L_1(1) = 0$, L_0 is continuous and strictly increasing in $[0,1]$, and L_1 is continuous and strictly decreasing in $[0,1]$. For any generalized inverses L_0^{-1} and L_1^{-1}, the condition (3.15) is equivalent to (3.12) for $y \in \{0,1\}$.*

Proof If $\Delta_0 \notin [0, L_0(1)]$, then both $L_0(\hat{y}_t) \leq \Delta_0$ and $\hat{y}_t \leq L_0^{-1}(\Delta_0)$ hold for all $\hat{y}_t \in [0,1]$. If $\Delta_1 \notin [0, L_1(0)]$, then both $L_1(\hat{y}_t) \leq \Delta_1$ and $L_1^{-1}(\Delta_1) \leq \hat{y}_t$ hold for all $\hat{y}_t \in [0,1]$. Hence, we may assume that Δ_0 is in the range of L_0 and Δ_1 is in the range of L_1. In this case (3.12) and (3.15) are equivalent because L_0 is strictly increasing and L_1 strictly decreasing. □

We are now ready to show that if in Algorithm 3.7 we use a value c such that $c \geq c_L$, where c_L is as defined in (3.3), and set $\eta = 1/c$, then the algorithm never fails.

Lemma 3.10: *Let L be any loss function such that L_0 and L_1 are twice continuously differentiable, $L_0(0) = L_1(1) = 0$, and $L_0'(z) > 0$ and $L_1'(z) < 0$ hold for $0 < z < 1$. Assume that the value c_L defined in (3.3) is finite, and $S(z)$ defined in (3.1) is positive for all z. Then for all $\mathbf{w}_t$ and $\mathbf{x}_t$ such that $0 \leq x_{t,i} \leq 1$ and $w_{t,i} \geq 0$ for $1 \leq i \leq N$, condition (3.16) holds whenever $c \geq c_L$ and $\eta = 1/c$.*

Proof For $0 \leq z \leq 1$, define $p(z) = \exp(-L_0(z)/c)$ and $q(z) = \exp(-L_1(z)/c)$, and for r in the range of p define

$$f(r) = \exp(-L_1(L_0^{-1}(-c \ln r))/c) \ . \tag{3.17}$$

Note that $f(p(z)) = q(z)$.

First, assume that $f''(p(z)) \leq 0$ holds for $0 \leq z \leq 1$. We are later going to show that this is in fact true if $c \geq c_L$. Let $r_i = p(x_{t,i})$ and $s_i = q(x_{t,i}) = f(r_i)$ for $i = 1, \ldots, N$. Then for $\eta = 1/c$ we have $\Delta_0 = -c \ln(\sum_i v_{t,i} r_i)$ and $\Delta_1 = -c \ln(\sum_i v_{t,i} s_i)$. The assumption $f''(r) \leq 0$ implies $\sum_i v_{t,i} s_i = \sum_i v_{t,i} f(r_i) \leq f(\sum_i v_{t,i} r_i)$. We get

$$\begin{aligned}
\Delta_1 &= -c \ln \left(\sum_{i=1}^{N} v_{t,i} s_i \right) \\
&\geq -c \ln \left(f \left(\sum_{i=1}^{N} v_{t,i} r_i \right) \right) \\
&= L_1 \left(L_0^{-1} \left(-c \ln \left(\sum_{i=1}^{N} v_{t,i} r_i \right) \right) \right) \\
&= L_1(L_0^{-1}(\Delta_0)) \ ,
\end{aligned}$$

from which condition (3.16) follows since L_1^{-1} is decreasing.

We now show that our assumptions on L_0 and L_1 imply that for $c \geq c_L$, the function f has a nonpositive second derivative in the range of q. We have $f(p(z)) = q(z)$ and thus $f'(p(z)) = q'(z)/p'(z)$. Differentiating further, we obtain $f''(p(z))p'(z) = (q''(z)p'(z) - q'(z)p''(z))/p'(z)^2$. Since $p'(z) = -L_0'(z)p(z)/c < 0$, we have $f''(p(z)) \leq 0$ if and only if $q''(z)p'(z) - q'(z)p''(z) \geq 0$. By substituting $p'(z) = -L_0'(z)p(z)/c$ and $p''(z) = (-L_0''(z)/c + (L_0'(z))^2/c^2)p(z)$, and using similar expressions for $q'(z)$ and $q''(z)$, we see that $f''(p(z)) \leq 0$ if and only if

$$\left(-L_0'(z)L_1'(z)^2 + L_1'(z)L_0'(z)^2 + c \left(L_0'(z)L_1''(z) - L_1'(z)L_0''(z) \right) \right) \frac{p(z)q(z)}{c^3} \geq 0.$$

Finally, since our assumptions imply $L_0'(z)L_1''(z) - L_1'(z)L_0''(z) > 0$, we conclude that $f''(p(z)) \leq 0$ holds if and only if $c \geq R(z)$. Hence, $c \geq c_L$ is a necessary and sufficient condition for having $f''(p(z)) \leq 0$ for all z. □

Note that above argument shows that the nonpositivity of $f''(r)$ is also a necessary condition. If $f''(r)$ is positive on some interval, by placing all the values $x_{t,i}$ in this interval but not making them equal we get $\sum_i v_{t,i} f(r_i) > f(\sum_i v_{t,i} r_i)$ and, hence, $L_1^{-1}(\Delta_1) > L_0^{-1}(\Delta_0)$.

In particular, we see that since the Generic Algorithm 3.7 does not fail with the paramaters $c = c_L$ and $\eta = 1/c_L$, we get the upper bound claimed in Theorem 3.1 by applying Theorem 3.8 with the initial weights $w_{1,i} = 1$ for all i.

Theorem 3.11: *Let L be a loss function for which the constant c_L is finite. Let A be the Generic Algorithm 3.7 with the parameters $c = c_L$, $\eta = 1/c_L$, and the initial weights $w_{1,i} = 1$ for all i. Then for all N and ℓ the additional loss of the algorithm satisfies*

$$V_{L,A}(N, \ell) \leq c_L \ln N \ .$$

We are now ready to write the Generic Algorithm 3.7 in a more explicit form for particular loss functions.

Example 3.12: If L is the logarithmic loss, we have $c_L = 1$ and can therefore take $c = \eta = 1$ in the Generic Algorithm 3.7. After simple manipulations we get $\Delta_0 = -\ln(1 - p_t)$ and $\Delta_1 = -\ln p_t$, where $p_t = \sum_i v_{t,i} x_{t,i}$ is the weighted average of the experts' predictions. Hence,

$$L_0^{-1}(\Delta_0) = L_1^{-1}(\Delta_1) = p_t \ ,$$

and $\hat{y}_t = p_t$ is the only prediction for which (3.12) holds for $y \in \{0, 1\}$ with this choice of c and η. The loss bound we obtain was previously shown by De Santis et al. [8] and Vovk [16]. □

Example 3.13: Let L be the square loss. Vovk [16] has shown that the square loss is $(1/2, 2)$-realizable. Here the result follows from Lemma 3.10 and Example 3.2. The note after the proof of Lemma 3.10 further implies that the square loss is not $(c, 1/c)$-realizable for any $c < 1/2$. Hence, we take $c = 1/2$ and $\eta = 2$ in the Generic Algorithm 3.7 for the square loss. The condition (3.12) for $y \in \{0, 1\}$ now becomes

$$1 - \left(-\frac{\ln \sum_{i=1}^{N} v_{t,i} e^{-2(1-x_{t,i})^2}}{2} \right)^{1/2} \leq \hat{y}_t \leq \left(-\frac{\ln \sum_{i=1}^{N} v_{t,i} e^{-2x_{t,i}^2}}{2} \right)^{1/2} . \tag{3.18}$$

By numerically substituting random values for $\mathbf{v}_t$ and $\mathbf{x}_t$ we see that the seemingly natural choice $\hat{y}_t = \sum_i v_{t,i} x_{t,i}$ usually does not satisfy (3.18). More generally, there is no function f such that choosing $\hat{y}_t = f\left(\sum_i v_{t,i} x_{t,i}\right)$ would guarantee (3.18) to hold. To see this, consider $N = 2$ and set first $\mathbf{x}_t = (0, 7/10)$ and $\mathbf{v}_t = (2/7, 5/7)$. Then $\sum_i v_{t,i} x_{t,i} = 1/2$, and evaluating the left-hand side of (3.18) with these values of $\mathbf{x}_t$ and $\mathbf{v}_t$ yields a bound $0.52 < f(1/2)$. On the other hand, we also have $\sum_i v_{t,i} x_{t,i} = 1/2$ when $\mathbf{x}_t = (3/10, 1)$ and $\mathbf{v}_t = (5/7, 2/7)$, and evaluating the right-hand side of (3.18) with these values gives the contradictory condition $f(1/2) < 0.48$. Hence, the algorithm needs more information than is provided by merely the weighted average of the experts' predictions.

It can be proved that in the more restricted case that all the experts' predictions $x_{t,i}$ are in $\{0, 1\}$, we can guarantee (3.15) for the square loss with $c = 1/\eta \approx 0.41$ instead of $c = 0.5$. This gives a slightly improved bound. However, restricting the experts to predict with binary values while allowing the algorithm to predict with countinuous values does not seem a natural setting. □

Example 3.14: Take L to be the absolute loss. As now $c_L = \infty$, we know that the absolute loss is not $(c, 1/c)$-realizable for any c. We therefore let $\eta > 0$ be arbitrary, and see for which values c the absolute loss is (c, η)-realizable.

By using the bound $e^{-\eta x} \leq 1 - (1 - e^{-\eta})x$ that holds for all $x \in [0, 1]$, we obtain

$$\begin{aligned}
& L_0^{-1}(\Delta_0) - L_1^{-1}(\Delta_1) \\
&= -c \ln \sum_{i=1}^{N} v_{t,i} e^{-\eta x_{t,i}} - \left(1 + c \ln \sum_{i=1}^{N} v_{t,i} e^{-\eta(1-x_{t,i})} \right) \\
&\geq c \left(-\ln \sum_{i=1}^{N} v_{t,i}(1 - (1 - e^{-\eta}) x_{t,i}) - \ln \sum_{i=1}^{N} v_{t,i}(1 - (1 - e^{-\eta})(1 - x_{t,i})) \right) - 1 \\
&= c \left(-\ln(1 - p_t + p_t e^{-\eta}) - \ln(p_t + (1 - p_t) e^{-\eta}) \right) - 1
\end{aligned}$$

where $p_t = \sum_i v_{t,i} x_{t,i}$. By Jensen's inequality, this is positive for $c \geq (2 \ln \frac{2}{1+e^{-\eta}})^{-1}$, and the prediction condition (3.12) for $y \in \{0, 1\}$ becomes

$$1 + \frac{\ln \sum_{i=1}^{N} v_{t,i} e^{-\eta(1-x_{t,i})}}{2 \ln \frac{2}{1+e^{-\eta}}} \leq \hat{y}_t \leq -\frac{\ln \sum_{i=1}^{N} v_{t,i} e^{-\eta x_{t,i}}}{2 \ln \frac{2}{1+e^{-\eta}}} . \tag{3.19}$$

Cesa-Bianchi et al. [2] have noted that (3.19) always holds if we choose

$$\hat{y}_t = \frac{\ln(1 - p_t + p_t e^{-\eta})}{\ln(1 - p_t + p_t e^{-\eta}) + \ln((1 - p_t) e^{-\eta} + p_t)} ,$$

but does not in general hold for $\hat{y}_t = p_t$. Hence, the weighted average of the experts' prediction provides sufficient information for the prediction, but cannot be used directly.

The bound obtained by applying Theorem 3.8 for the absolute loss with the choice $c = \left(2 \ln \frac{2}{1+e^{-\eta}}\right)^{-1}$, namely

$$\mathrm{Loss}_L(A, S) \leq \frac{-\ln \frac{w_{1,i}}{W_1} + \eta \mathrm{Loss}_L(\mathcal{E}_i, S)}{2 \ln \frac{2}{1+e^{-\eta}}} \quad , \tag{3.20}$$

was first proven by Vovk [16]. We would like to choose the learning rate η in such a way that the loss bound on the right-hand side of (3.20) is minimized. This tuning of the learning rate is discussed in detail by Cesa-Bianchi et al. [2, 3]. Here we just cite some of the basic results. If all the initial weights $w_{1,i}$ are 1 and η is chosen to be $\ln h\left(\sqrt{2(\ln N)/\ell}\right)$ where $h(z) = 1 + 2z + z^2/\ln 2$, the Generic Algorithm 3.7 for absolute loss satisfies

$$V_{L,A}(N, \ell) \leq \sqrt{\frac{\ell \ln(N+1)}{2}} + \frac{\log_2(N+1)}{2} \quad .$$

Note that here it is necessary to know ℓ before the first trial in order to choose the learning rate η appropriately. Similar results can be obtained by basing the choise of η on an upper bound for the loss $\min_i \mathrm{Loss}_L(\mathcal{E}_i, S)$ of the best expert instead of on ℓ.

Finally, we consider the variations of the Generic Algorithm given by Cesa-Bianchi et al. [2] for the special case of the absolute loss. Instead of the update (3.13), we write more generally $w_{t+1,i} = \alpha_{t,i} w_{t,i}$ and $\Delta(y) = -c \ln \sum_{i=1}^{N} v_{t,i} \alpha_{t,i}$, and consider choices for the factors $\alpha_{t,i}$ in addition to the choice $\alpha_{t,i} = e^{-\eta |y_t - x_{t,i}|}$ of the Generic Algorithm. First, note that if $-\ln \alpha_{t,i} \leq \eta |y_t - x_{t,i}|$, the proof of Theorem 3.8 can easily be generalized to yield the same loss bound. Second, note that the proof given for the inequality $L_1^{-1}(\Delta_1) \leq L_0^{-1}(\Delta_0)$ is valid assuming $\alpha_{t,i} \leq 1 - (1 - e^{-\eta}) x_{t,i}$. Hence, the algorithm works and gives the same worst-case loss bound for any choice

$$e^{-\eta |y_t - x_{t,i}|} \leq \alpha_{t,i} \leq 1 - (1 - e^{-\eta}) x_{t,i} \quad . \tag{3.21}$$

Interestingly enough, the weights obtained using $\alpha_{t,i} = 1 - (1 - e^{-\eta}) x_{t,i}$ have a Bayesian interpretation [2]. □

3.3 Lower bounds

This subsection contains proofs of the lower bounds for $V_L(N, \ell)$ stated in Theorems 3.1 and 3.5 in Subsection 3.1. The lower bounds hold even for algorithms that receive ℓ as input begore the first trial. Theorem 3.16 shows how a probability measure for the experts and outcomes leads to a lower bound for $V_L(N, \ell)$ for large N and ℓ. The proof of Theorem 3.16 is based on Lemma 3.15, which shows that we can change the order of taking expectations and going to the limit with certain random variable sequences. The lower bound in Theorem 3.16 is in terms of certain characteristics of the probability measures, and is interesting only if the probability measures are chosen carefully. Lemma 3.17 shows a particular way of choosing the probability measures, when a prediction b is the unique Bayes-optimal prediction for a bias q. Lemmas 3.18 show a way to choose the probability measures in Theorem 3.16 if the Bayes-optimal prediction is not unique. Finally, we combine the results by showing that either each prediction z can be made to be the unique Bayes-optimal prediction by choosing a suitable bias, in which case Lemma 3.17 yields a lower bound for $V_L(N, \ell)$ in terms of c_L, or else there is a bias for which two distinct Bayes-optimal prediction exist and Lemma 3.18 yields a lower bound $V_L(N, \ell) = \Omega\left(\sqrt{\ell \log N}\right)$.

We begin with a technical lemma.

Lemma 3.15: *Let P be a probability measure in X and Q a probability measure in Y. For $\ell \in \mathbf{N}_+$ and $y \in Y$, let $U^y_{1\ell}, \ldots, U^y_{N\ell}$ be N independent identically distributed random variables such that $\mathrm{E}_{x\in P}[U^y_{i\ell}(x)] = 0$ and $\mathrm{Var}_{x\in P}[U^y_{i\ell}(x)] = 1$. Assume that there are independent identically distributed random variables $F_1, \ldots, F_N$ such that the sequence $U^y_{i1}, U^y_{i2}, \ldots$ converges in distribution to F_i for all i and y. Further, let $r_1, r_2, \ldots$ be functions on Y such that $\lim_{\ell\to\infty} r_\ell(y) = 1$ holds with probability 1 for y drawn according to Q, and $|r_\ell(y)| \le B$ holds for all y for some constant B. Then*

$$\lim_{\ell\to\infty} \mathrm{E}_{y\in Q}\left[r_\ell(y)\mathrm{E}_{x\in P}\left[\min_{1\le i\le N} U^y_{i\ell}(x)\right]\right] = \mathrm{E}\left[\min_{1\le i\le N} F_i\right] \quad .$$

Proof Write $U^y_{*\ell} = \min_{1\le i\le N} U^y_{i\ell}$ and $F_* = \min_{1\le i\le N} F_i$. We first show that for all y, the sequence $U^y_{*1}, U^y_{*2}, \ldots$ converges in distribution to F_*. For all $a \in \mathbf{R}$ we have

$$\begin{aligned}
\Pr[F_* \le a] &= 1 - \prod_{i=1}^{N}(1 - \Pr[F_i \le a]) \\
&= 1 - \prod_{i=1}^{N}(1 - \lim_{\ell\to\infty} \Pr[U^y_{i\ell} \le a]) \\
&= \lim_{\ell\to\infty}(1 - \prod_{i=1}^{N}(1 - \Pr[U^y_{i\ell} \le a])) \\
&= \lim_{\ell\to\infty} \Pr[U^y_{*\ell} \le a] \quad ,
\end{aligned}$$

which proves the claim.

Next we see that

$$\mathrm{E}_{x\in P}\left[\left|U^y_{*\ell}(x)\right|^{1+p}\right] \le 2N \tag{3.22}$$

holds for all y when $p = 0$ or $p = 1$. To see this, first note that for all $A \subseteq \mathbf{R}$, if $U^y_{*\ell}(x) \in A$ then $U^y_{i\ell}(x) \in A$ for at least one value i. As the distribution of $U^y_{i\ell}$ does not depend on i, this implies $\Pr_{x\in P}[U^y_{*\ell}(x) \in A] \le N \Pr_{x\in P}[U^y_{1\ell}(x) \in A]$ if A is measurable. This implies

$$\begin{aligned}
\mathrm{E}_{x\in P}\left[\left|U^y_{*\ell}(x)\right|^{1+p}\right] &\le N\mathrm{E}_{x\in P}\left[\left|U^y_{1\ell}(x)\right|^{1+p}\right] \\
&= N\int \left|U^y_{1\ell}\right|^{1+p} dP \\
&\le N\left(1 + \int_{|U^y_{1\ell}|\ge 1} \left|U^y_{1\ell}\right|^{1+p} dP\right) \\
&\le N\left(1 + \mathrm{E}_{x\in P}\left[\left(U^y_{1\ell}(x)\right)^2\right]\right) \\
&= 2N \quad .
\end{aligned}$$

As the sequence $U^y_{*1}, U^y_{*2}, \ldots$ converges in distribution to F_*, the bound (3.22) with $p = 1$ guarantees [1, Corollary, p. 292] $\lim_{\ell\to\infty} \mathrm{E}_{x\in P}\left[U^y_{*\ell}(x)\right] = \mathrm{E}\left[F_*\right]$ for all y and, therefore, $\lim_{\ell\to\infty} r_\ell(y)\mathrm{E}_{x\in P}\left[U^y_{*\ell}(x)\right] = \mathrm{E}\left[F_*\right]$ with probability 1 for y drawn from Q. The bound (3.22) with $p = 0$ implies $\left|r_\ell(y)\mathrm{E}_{x\in P}\left[U^y_{*\ell}(x)\right]\right| \le 2BN$, and the bounded convergence theorem [1, Thm. 16.5, p. 180]

$$\lim_{\ell\to\infty} \mathrm{E}_{y\in Q}\left[r_\ell(y)\mathrm{E}_{x\in P}\left[U^y_{*\ell}(x)\right]\right] = \mathrm{E}\left[F_*\right] \quad ,$$

as claimed. □

Theorem 3.16 shows how a probability measure for the experts and outcomes leads to a lower bound for $V_L(N, \ell)$ for large N and ℓ.

Theorem 3.16: *Let P be a probability measure on $[0,1]$ and Q a probability measure on $\{0,1\}$. Assume that for $y = 0$ and $y = 1$, the condition $\Pr_{x\in P}[L(y,x) > K] = 0$ holds for some constant K. Let b be a Bayes-optimal prediction for Q. Let $\tau = \mathrm{E}_{y\in Q, x\in P}[L(y,x)]$ and $\sigma^2 = \mathrm{E}_{y\in Q}\left[\mathrm{Var}_{x\in P}\left[L(y,x)\right]\right]$. Assume that for $y = 0$ and $y = 1$ the variance $\mathrm{Var}_{x\in P}\left[L(y,x)\right]$ is strictly positive. Then for all $\varepsilon > 0$ there is an ℓ_ε such that for all $\ell \geq \ell_\varepsilon$ we have*

$$V_L(N,\ell) \geq \ell \mathrm{E}_{y\in Q}[L(y,b)] - \ell\tau + (a_N - \varepsilon)\sigma\sqrt{\ell \ln N} \ , \tag{3.23}$$

where $\lim_{N\to\infty} a_N = \sqrt{2}$.

Proof Given $\mathbf{x} \in [0,1]^{N\times\ell}$ and $\mathbf{y} \in \{0,1\}^\ell$, we define an N-expert trial sequence of length ℓ by $\langle \mathbf{x}, \mathbf{y}\rangle = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$. For an on-line prediction algorithm A, consider $V_{L,A}(\langle \mathbf{x}, \mathbf{y}\rangle)$ as a random variable, with $\mathbf{x}$ and $\mathbf{y}$ drawn from the product measures $P^{N\times\ell}$ and Q^ℓ, respectively. The expected value of a random variable is clearly a lower bound for the supremum. Combining this with the linearity of expectation, we get

$$\begin{aligned} V_{L,A}(N,\ell) &\geq \mathrm{E}_{\mathbf{x}\in P^{N\times\ell}}\mathrm{E}_{\mathbf{y}\in Q^\ell} V_{L,A}(\langle \mathbf{x}, \mathbf{y}\rangle) \\ &= \sum_{j=1}^{\ell} \mathrm{E}_{y\in Q}[L(y,\hat{y}_t)] - \mathrm{E}_{\mathbf{x}\in P^{N\times\ell}}\mathrm{E}_{\mathbf{y}\in Q^\ell}\left[\min_{1\leq i\leq N} \mathrm{Loss}_L(\mathcal{E}_i, \langle \mathbf{x}, \mathbf{y}\rangle)\right] \\ &\geq \ell \mathrm{E}_{y\in Q}[L(y,b)] - \mathrm{E}_{\mathbf{x}\in P^{N\times\ell}}\mathrm{E}_{\mathbf{y}\in Q^\ell}\left[\min_{1\leq i\leq N} \mathrm{Loss}_L(\mathcal{E}_i, \langle \mathbf{x}, \mathbf{y}\rangle)\right] \ . \end{aligned}$$

Since this holds for any A, we obtain (3.23) if we can prove that

$$\mathrm{E}_{\mathbf{x}\in P^{N\times\ell}}\mathrm{E}_{\mathbf{y}\in Q^\ell}\left[\min_{1\leq i\leq N} \mathrm{Loss}_L(\mathcal{E}_i, \langle \mathbf{x}, \mathbf{y}\rangle)\right] \leq \ell\tau - (a_N - \varepsilon)\sigma\sqrt{\ell \ln N} \ . \tag{3.24}$$

Let $q = \Pr_{y\in Q}[y = 1]$. Then

$$\tau = (1-q)\mathrm{E}_{x\in P}[L_0(x)] + q\mathrm{E}_{x\in P}[L_1(x)]$$

and

$$\sigma^2 = (1-q)\mathrm{Var}_{x\in P}[L_0(x)] + q\mathrm{Var}_{x\in P}[L_1(x)] \ .$$

Given a sequence $\mathbf{y} \in \{0,1\}^\infty$ and $\ell \in \mathbf{N}_+$, define

$$\hat{q}_\ell(\mathbf{y}) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i \ .$$

We also let

$$\hat{\tau}_\ell(\mathbf{y}) = (1 - \hat{q}_\ell(\mathbf{y}))\mathrm{E}_{x\in P}[L_0(x)] + \hat{q}_\ell(\mathbf{y})\mathrm{E}_{x\in P}[L_1(x)]$$

and

$$\hat{\sigma}_\ell(\mathbf{y})^2 = (1 - \hat{q}_\ell(\mathbf{y}))\mathrm{Var}_{x\in P}[L_0(x)] + \hat{q}_\ell(\mathbf{y})\mathrm{Var}_{x\in P}[L_1(x)]$$

be the estimates obtained for τ and σ^2 by using $\hat{q}_\ell(\mathbf{y})$ instead of the true probability q.

For $\mathbf{x} \in [0,1]^{N\times\infty}$ and $\mathbf{y} \in \{0,1\}^\infty$, let $T_{ij}^{\mathbf{y}}(\mathbf{x}) = L(y_j, x_{ij})$ be the loss of expert i at trial j, if $\mathbf{x}$ is the sequence of experts' predictions and $\mathbf{y}$ the sequence of outcomes. We consider $T_{ij}^{\mathbf{y}}$ as a random variable on the domain $[0,1]^{N\times\infty}$. We now define for $i = 1, \ldots, N$ and $\ell = 1, 2, \ldots$ the random variable $S_{i\ell}$ in the domain $[0,1]^{N\times\infty} \times \{0,1\}^\infty$ by $S_{i\ell}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\ell} L(y_j, x_{ij})$ to denote the loss of expert i in the first ℓ trials. We also define for a given sequence $\mathbf{y} \in \{0,1\}^\infty$

the random variable $S^{\mathbf{y}}_{i\ell}$ by $S^{\mathbf{y}}_{i\ell}(\mathbf{x}) = S_{i\ell}(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{\ell} T^{\mathbf{y}}_{ij}$. The underlying probability measures for these random variables are the product measures defined by P and Q, so for a fixed $\mathbf{y}$ the random variables $T^{\mathbf{y}}_{ij}$ and $T^{\mathbf{y}}_{i'j'}$ are independent for $(i,j) \neq (i',j')$. To study the distribution of $S^{\mathbf{y}}_{i\ell}$, we define a suitably normalized random variable $U^{\mathbf{y}}_{i\ell}$. Let now

$$U^{\mathbf{y}}_{i\ell} = \frac{S^{\mathbf{y}}_{i\ell} - \sum_{j=1}^{\ell} \mathrm{E}[T^{\mathbf{y}}_{ij}]}{\sqrt{\sum_{j=1}^{\ell} \mathrm{Var}[T^{\mathbf{y}}_{ij}]}} \ . \tag{3.25}$$

Then $\mathrm{E}[U^{\mathbf{y}}_{i\ell}] = 0$ and $\mathrm{Var}[U^{\mathbf{y}}_{i\ell}] = 1$. Further, since we have assumed that $\Pr[|T^{\mathbf{y}}_{ij}|] > K) = 0$, the Lindeberg form of the central limit theorem implies that each sequence $U^{\mathbf{y}}_{i1}, U^{\mathbf{y}}_{i2}, \ldots$ converges in distribution to a standard normal random variable.

We now apply Lemma 3.15 to the random variables $U^{\mathbf{y}}_{i\ell}$. Then the random variables F_i in Lemma 3.15 have standard normal distribution. By a standard result [10], their minimum F_* has expectation $\mathrm{E}[F_*] = -a_N\sqrt{\ln N}$, where $\lim_{N\to\infty} a_N = \sqrt{2}$. We take $r_\ell(\mathbf{y}) = \hat{\sigma}_\ell(\mathbf{y})/\sigma$. Then $|r_\ell(\mathbf{y})| \leq K/\sigma$, and by the strong law of large numbers we have $\lim_{\ell\to\infty} r_\ell(\mathbf{y}) = 1$ for almost all $\mathbf{y}$. Lemma 3.15 now implies

$$\lim_{\ell\to\infty} \mathrm{E}_{\mathbf{y}\in Q^\infty}\left[\frac{\hat{\sigma}_\ell(\mathbf{y})}{\sigma}\mathrm{E}_{\mathbf{x}\in P^{N\times\infty}}\left[\min_{1\leq i\leq N} U^{\mathbf{y}}_{i\ell}\right]\right] = -a_N\sqrt{\ln N} \ . \tag{3.26}$$

By partitioning the summations in (3.25) into two parts according to whether $y_i = 0$ or $y_i = 1$, we can write

$$U^{\mathbf{y}}_{i\ell} = \frac{S^{\mathbf{y}}_{i\ell} - \ell((1-\hat{q}_\ell(\mathbf{y}))\mathrm{E}_{x\in P}[L_0(x)] + \hat{q}_\ell(\mathbf{y})\mathrm{E}_{x\in P}[L_1(x)])}{\sqrt{\ell((1-\hat{q}_\ell(\mathbf{y}))\mathrm{Var}_{x\in P}[L_0(x)] + \hat{q}_\ell(\mathbf{y})\mathrm{Var}_{x\in P}[L_1(x)])}} = \frac{S^{\mathbf{y}}_{i\ell} - \ell\hat{\tau}_\ell(\mathbf{y})}{\hat{\sigma}_\ell(\mathbf{y})\sqrt{\ell}} \ .$$

By substituting this into (3.26), we obtain

$$\lim_{\ell\to\infty} \frac{\mathrm{E}_{\mathbf{y}\in Q^\infty}\left[\mathrm{E}_{\mathbf{x}\in P^{N\times\infty}}\left[\min_{1\leq i\leq N} S^{\mathbf{y}}_{i\ell}(\mathbf{x}) - \ell\hat{\tau}_\ell(\mathbf{y})\right]\right]}{\sigma\sqrt{\ell}} = -a_N\sqrt{\ln N} \ .$$

Therefore, for all $\varepsilon > 0$ there is a value ℓ_ε such that for all $\ell \geq \ell_\varepsilon$ we have

$$\begin{aligned}
&\mathrm{E}_{\mathbf{y}\in Q^\infty}\left[\mathrm{E}_{\mathbf{x}\in P^{N\times\infty}}\left[\min_{1\leq i\leq N} S_{i\ell}(\mathbf{x},\mathbf{y}) - \ell\hat{\tau}_\ell(\mathbf{y})\right]\right] \\
&\quad = \quad \mathrm{E}_{\mathbf{y}\in Q^\ell}\left[\mathrm{E}_{\mathbf{x}\in P^{N\times\ell}}\left[\min_{1\leq i\leq N} \mathrm{Loss}_L(\mathcal{E}_i, \langle \mathbf{x},\mathbf{y}\rangle)\right]\right] - \ell\tau \\
&\quad \leq \quad -(a_N - \varepsilon)\sigma\sqrt{\ell\ln N} \ .
\end{aligned}$$

This implies (3.24), as desired. □

We now see how Theorem 3.16 implies a lower bound for $V_L(N,\ell)$ when the probability measure P for the experts is chosen suitably.

Lemma 3.17: *Let L be a loss function such that L_0 and L_1 are twice differentiable, and $L_0'(z) > 0$ and $L_1'(z) < 0$ hold for $0 < z < 1$. Assume that $b \in (0,1)$ is a Bayes-optimal prediction for bias $q \in (0,1)$.*

1. *If $(1-q)L_0''(b) + qL_1''(b) > 0$, then*

$$V_L(N,\ell) \geq (R(b) - o(1))\ln N \ ,$$

where $R(b)$ is as in (3.2) and $o(1)$ denotes a quantity that approaches 0 as ℓ and N approach ∞.

2. *If* $(1-q)L_0''(b) + qL_1''(b) = 0$, *then for all* $\alpha > 0$ *we have* $V_L(N,\ell) = \Omega\left(\ell^{1/2-\alpha}\sqrt{\ln N}\right)$.

Proof Let Q be the probability measure on $\{0,1\}$ for which $\Pr_{y\in Q}[y=1] = q$. Let A be an arbitrary on-line prediction algorithm. For any probability measure P on $[0,1]$ and for any $\varepsilon > 0$, we have by Theorem 3.16 for sufficiently large ℓ the bound

$$V_{L,A}(N,\ell) \geq \ell(\mathrm{E}_{y\in Q}[L(y,b)] - \tau) + (a_N - \varepsilon)\sigma\sqrt{\ell \ln N} \ , \tag{3.27}$$

where $\lim_{N\to\infty} a_N = \sqrt{2}$. For some positive parameter h, define P to give $x = b - h$ with probability $1/2$ and $x = b + h$ with probability $1/2$. We can expand

$$L_0(b \pm h) = L_0(b) \pm L_0'(b)h + \frac{L_0''(b)}{2}h^2 + o(h^2) \ ,$$

where $o(h^2)$ denotes a quantity $f(h)$ such that $\lim_{h\to 0}(f(h)/h^2) = 0$, and similarly for L_1. We now substitute these expansions into the various quantities in (3.27). First, note that $E_{x\in P}[L_0(x)] = L_0(b) + h^2 L_0''(b)/2 + o(h^2)$, so

$$\mathrm{Var}_{x\in P}[L_0(x)] = \mathrm{E}_{x\in P}\left[(L_0(x) - E_{x\in P}L_0(x))^2\right] = L_0'(b)^2h^2 + o(h^2) \ .$$

Similarly, $\mathrm{Var}_{x\in P}[L_1(x)] = L_1'(b)^2h^2 + o(h^2)$, and

$$\sigma^2 = h^2((1-q)L_0'(b)^2 + qL_1'(b)^2) + o(h^2) \ .$$

We also have

$$\tau = (1-q)(L_0(b) + h^2L_0''(b)/2) + q(L_1(b) + h^2L_1''(b)/2) + o(h^2) \ ,$$

so

$$\mathrm{E}_{y\in Q}[L(y,b)] - \tau = -\frac{h^2}{2}((1-q)L_0''(b) + qL_1''(b)) - o(h^2) \ .$$

Hence, $V_{L,A}(N,\ell) \geq \ell(rh - sh^2) - o(h^2)$, where

$$r = (a_N - \varepsilon)\sqrt{\frac{\ln N}{\ell}}\sqrt{(1-q)L_0'(b)^2 + qL_1'(b)^2}$$

and

$$s = \frac{(1-q)L_0''(b) + qL_1''(b)}{2} \ .$$

We first consider the case $s > 0$, which gives the first part of the theorem. The main part $\ell(rh - sh^2)$ is maximized by choosing $h = r/(2s) = \Theta\left(\sqrt{(\ln N)/\ell}\right)$. For this value of h, we get

$$\begin{aligned} V_{L,A}(N,\ell) &\geq \ell\frac{r^2}{4s} + o((\ln N)/\ell) \\ &= \frac{(a_N-\varepsilon)^2}{2}\frac{(1-q)L_0'(b)^2 + qL_1'(b)^2}{(1-q)L_0''(b) + qL_1''(b)}\ln N - o((\ln N)/\ell) \ . \end{aligned}$$

Application of (3.7) now gives the claimed result, since $\lim_{N\to\infty} a_N^2/2 = 1$.

Consider now the case $s = 0$, which gives the second part of the theorem. We need a sequence $h_1, h_2, \ldots$ with $\lim_{\ell\to\infty} h_\ell = 0$. To obtain the actual bound claimed here, we choose $h_\ell = \ell^{-\alpha}$. Slightly different results can be obtained by choosing different values h_ℓ. We now have $V_{L,A}(N,\ell) \geq a\ell^{1/2-\alpha}\sqrt{\ln N} - o(\ell^{-2\alpha})$, where

$$a = (a_N - \varepsilon)\sqrt{(1-q)L_0'(b)^2 + qL_1'(b)^2} > 0 \ .$$

□

Lemma 3.18: *Let L be a loss function such that L_0 is strictly increasing and L_1 strictly decreasing. Assume that for bias q there are two distinct Bayes-optimal predictions b_1 and b_2. Then for all $\varepsilon > 0$ there is an ℓ_ε such that for all $\ell \geq \ell_\varepsilon$ we have*

$$V_L(N,\ell) \geq (a_N - \varepsilon)\sigma\sqrt{\ell \ln N} \ ,$$

where $\lim_{N\to\infty} a_N = \sqrt{2}$ and

$$\sigma^2 = \frac{1-q}{4}\left(L_0(b_1) - L_0(b_2)\right)^2 + \frac{q}{4}\left(L_1(b_1) - L_1(b_2)\right)^2 \ . \tag{3.28}$$

Proof Let b_1 and b_2 be two distinct Bayes-optimal predictions for some probability measure Q on $\{0,1\}$. As L_0 and L_1 are strictly monotone, the bias of Q cannot be 0 or 1. We define a probability measure P by $\Pr_{x\in P}[x = b_1] = \Pr_{x\in P}[x = b_2] = 1/2$, and apply Theorem 3.16. Then $\tau = \mathrm{E}_{y\in Q}[L(y,b_1)] = \mathrm{E}_{y\in Q}[L(y,b_2)]$. Further, we get

$$\begin{aligned}
\mathrm{Var}_{x\in P}[L(0,x)] &= \mathrm{E}_{x\in P}\left[L(0,x)^2\right] - \mathrm{E}_{x\in P}[L(0,x)]^2 \\
&= \frac{1}{2}L_0(b_1)^2 + \frac{1}{2}L_0(b_2)^2 - \left(\frac{1}{2}L_0(b_1) + \frac{1}{2}L_0(b_2)\right)^2 \\
&= \frac{1}{4}\left(L_0(b_1) - L_0(b_2)\right)^2 \ ,
\end{aligned}$$

and similarly $\mathrm{Var}_{x\in P}[L(0,x)] = \frac{1}{4}\left(L_1(b_1) - L_1(b_2)\right)^2$. Hence, σ is as given in (3.28). The results now follows from Theorem 3.16 with either $b = b_1$ or $b = b_2$. □

Note that for the absolute loss, we can apply Lemma 3.18 with $q = 1/2$, $b_1 = 0$, and $b_1 = 1$. This gives $\sigma = 1/2$, and hence $V_L(N,\ell) \geq (1 - o(1))\sqrt{(\ell \ln N)/2}$, which is the result obtained by Cesa-Bianchi et al. [2].

Lemma 3.19: *If a prediction $z \in (0,1)$ is not Bayes-optimal for any bias $q \in [0,1]$, then there are two predictions b_1 and b_2 with $b_1 < z < b_2$ such that for some bias q both b_1 and b_2 are Bayes-optimal.*

Proof Consider a prediction $z \in (0,1)$ that is not Bayes-optimal for any bias. Let R_1 be the set of biases q for which there is a Bayes-optimal prediction $b < z$, and let R_2 be the set of biases q for which there is a Bayes-optimal prediction $b > z$. If we can show $R_1 \cap R_2 \neq \emptyset$, we are done. Since z is never Bayes-optimal, we have $R_1 \cup R_2 = [0,1]$. Hence, if both R_1 and R_2 are closed, their intersection cannot be empty.

Suppose that R_1 is not closed. Let $p_1, p_2, \ldots$ be a monotone sequence of points in R_1 that converges to a point $p \notin R_1$. Let $b_n < z$ be a Bayes-optimal prediction for bias p_n, $n = 0, 1, \ldots$. The sequence $b_1, b_2, \ldots$ is also monotone and converges to some limit $b \leq z$. Let b' be a Bayes-optimal prediction for bias p. As $p \notin R_1$, we have $b' > z$. Define $F(q,x) = (1-q)L_0(x) + qL_1(x)$. Since b_n is Bayes-optimal for bias p_n, we have $F(p_n,b_n) \leq F(p_n,b')$ for all n. Since F is continuous, this implies $F(p,b) \leq F(p,b')$. As b' is Bayes-optimal for bias p, so is b. Thus $p \in R_1$, contradiction. Similar argument works if we assume R_2 to be not closed. □

Proof of Lemma 3.4 Since we assume L_0 to be strictly increasing and L_1 to be strictly decreasing, 0 is the unique Bayes-optimal prediction for the bias 0 and 1 is the unique Bayes-optimal prediction for the bias 1.

Assume first that b_1 and b_2 are two Bayes-optimal predictions for some bias $0 < q < 1$, with $b_1 < b_2$. Thus, the expected loss $f(z) = (1-q)L_0(z) + qL_1(z)$ has local minima at $z = b_1$ and $z = b_2$, and therefore $f(z)$ has a local maximum at some value a with $b_1 < a < b_2$. We then have $f'(a) = 0$ and $f''(a) \leq 0$. The condition $f'(a) = 0$ implies $q/(1-q) = -L_0'(a)/L_1'(a)$, which substituted into $f''(a) \leq 0$ gives $S(a) \leq 0$.

Assume now that for every bias q there is a unique Bayes-optimal prediction. Then Lemma 3.19 implies that for all z there is a bias q for which z is Bayes-optimal, and we know that this bias q must be unique. Let $B(z)$ denote the bias for which z is the Bayes-optimal prediction. We know that B is strictly increasing. Let $f(z) = -L_0'(z)/L_1'(z)$. We then have $f(z) = g(B(z))$ where $g(q) = q/(1-q)$. Since g and B are strictly increasing, so is f, and therefore the derivative $f'(z)$ cannot be negative, and cannot be zero on any continuous interval. As

$$f'(z) = \frac{L_0'(z)L_1''(z) - L_1'(z)L_0''(z)}{L_1'(z)^2} = \frac{S(z)}{L_1'(z)^2} \quad ,$$

the claim follows. □

The lower bounds in Theorem 3.1 and Theorem 3.5 follow directly from the following theorem.

Theorem 3.20: *Let L be a loss function such that L_0 and L_1 are twice differentiable, and $L_1'(z) > 0$ and $L_1'(z) < 0$ hold for all $0 < z < 1$. Let $S(z)$ be as in (3.1).*

1. *If $S(z) > 0$ for $0 < z < 1$, then $V_L(N,\ell) \geq (c_L - o(1)) \ln N$, where c_L is as in (3.3).*
2. *If $S(z) = 0$ for some $0 < z < 1$, then $V_L(N,\ell) = \Omega\left(\ell^{1/2-\alpha}\sqrt{\ln N}\right)$ for all $\alpha > 0$.*
3. *If $S(z) < 0$ for some $0 < z < 1$, or $S(z) = 0$ for all the values z in some continuous interval, then $V_L(N,\ell) = \Omega\left(\sqrt{\ell \ln N}\right)$.*

Proof If for some bias there are two distinct Bayes-optimal predictions, we have by Lemma 3.18 the bound $V_L(N,\ell) = \Omega\left(\sqrt{\ell \ln N}\right)$, which is the strongest of the bounds claimed here. Thus, we only need to consider the case in which for each bias there is at most one Bayes-optimal prediction. By Lemma 3.19, we then have for all predictions z a bias such that z is Bayes-optimal. By Lemma 3.4, the value $S(z)$ is always nonnegative and cannot be zero on any continuous interval.

Recall that when z is Bayes-optimal for q, the condition (3.6) implies $(1-q)L_0''(z) + qL_1''(z) = S(z)$. If $S(z) = 0$, then applying Lemma 3.17 (2) with the bias q that makes z Bayes-optimal gives the bound $V_L(N,\ell) = \Omega\left(\ell^{1/2-\alpha}\sqrt{\ln N}\right)$ for all $\alpha > 0$. If $S(z) > 0$ for all z, Lemma 3.17 (1) gives $V_L(N,\ell) \geq (R(z) - o(1)) \ln N$ for all z, from which $V_L(N,\ell) \geq (c_L - o(1)) \ln N$ follows. □

3.4 Alternative lower bound methods

First notice that for the logarithmic loss, there is a simple argument that shows the lower bound $V_L(N,\ell) \geq \ln N$ for $N = 2^k$ and $\ell \geq k$.

Example 3.21: For arbitrary positive integer k, let $N = 2^k$ and $\ell = k$. Let A be an arbitrary on-line prediction algorithm. For the trials $t = 1, \ldots, \ell$ we choose binary prediction vectors $\mathbf{x}_t \in \{0,1\}^N$ in such a way that the set of the experts' prediction sequences $\{(x_{1,i}, \ldots, x_{t,i}) \mid 1 \leq i \leq N\}$ contains all the $2^\ell = N$ possible binary sequences of length ℓ. The outcomes y_t are chosen by an adversary in such a way that $y_t = 0$ if the prediction $\hat{y}_t$ of the algorithm A satisfies $\hat{y}_t \geq 1/2$, and $y_t = 1$ otherwise. Then at each trial the algorithm incurs

loss at least $\ln 2$, and the total loss of the algorithm will $\ell \ln 2 = \ln N$. One expert will have total loss 0, so we obtain $V_{L,A}(N, \ell) \geq \ln N$. This matches exactly the upper bound for $V_{L,A}(N, \ell)$ given in Theorem 3.1 and Example 3.2 when A is the Generic Algorithm 3.7.

Another way of thinking of this lower bound argument is as follows. At the first trial, half of the experts predict 0 and half of the experts predict 1. After the trial, those that made a mistake are eliminated, and those that were correct remain. At subsequent trials, half of the remaining experts predict 0 and half predict 1. Thus, at trial t there are $N/2^{t-1}$ experts remaining, each with cumulative loss 0, while the rest of the experts have cumulative loss ∞ and have been eliminated. □

Note that by considering a single trial this easily gives for the logarithmic loss the bound $V_L(2, 1) \geq \ln 2$. The general lower bound $V_L(N, \ell) \geq \ln N$ for the logarithmic loss, when $N = 2^k$ and $\ell \geq k$, can also be obtained by applying the following Theorem 3.23 to this lower bound for $V_L(2, 1)$. Theorem 3.23 is proven using the following lemma.

Lemma 3.22: *Assume that for all on-line prediction algorithms A' there is an N-expert trial sequence S' of length ℓ' such that $V_{L,A'}(S') \geq a$, and that for all on-line prediction algorithms A'' there is a two-expert trial sequence S'' of length ℓ'' such that $V_{L,A''}(S'') \geq b$. Then for all on-line prediction algorithms A there is a $2N$-expert trial sequence S of length $\ell' + \ell''$ such that $V_{L,A}(S) \geq a + b$.*

Proof A $2N$-expert *coupled* trial sequence is a sequence in which each instance $\mathbf{x}_t$ has the property $x_{t,i} = x_{t,N+i}$ for $1 \leq i \leq N$. A $2N$-expert *simple* trial sequence is a sequence where each instance $\mathbf{x}_t$ has the property $x_{t,1} = x_{t,2} = \cdots = x_{t,N}$ and $x_{t,N+1} = x_{t,N+2} = \cdots = x_{t,2N}$. Note that $2N$-expert coupled trial sequences are essentially N-expert trial sequences and $2N$-expert simple trial sequences are essentially two-expert trial sequences.

Since we assumed that for all prediction algorithms A' there is an N-expert trial sequence S' of length ℓ' such that $V_{L,A'}(S') \geq a$, it follows that for all on-line prediction algorithms A there is a $2N$-expert coupled trial sequence S_1 of length ℓ' such that $V_{L,A}(S_1) \geq a$. Similarly, since we assumed that for all prediction algorithms A'' there is a two-expert trial sequence S'' of length ℓ'' such that $V_{L,A''}(S'') \geq b$, it follows that for all on-line prediction algorithms A there is a $2N$-expert simple trial sequence S_2 of length ℓ'' such that $V_{L,A}(S_2) \geq b$.

Let now A be an arbitrary on-line prediction algorithm for trial sequences of length $\ell' + \ell''$. Given a trial sequence S' of length ℓ', let $A(S')$ denote the algorithm for trial sequences of length ℓ'' that simulates the algorithm A but processes the trial sequence S' before the first actual trial. Our assumptions imply that there is a $2N$-expert coupled trial sequence S_1 of length ℓ' for which $V_{L,A}(S_1) \geq a$, and that there is a $2N$-expert simple trial sequence S_2 of length ℓ'' for which $V_{L,A(S_1)}(S_2) \geq b$. Let S be the $2N$-expert trial sequence of length $\ell' + \ell''$ that is obtained by concatenating S_1 and S_2.

To complete the proof, we show that $\mathrm{Loss}_L(A, S) - \mathrm{Loss}_L(\mathcal{E}_i, S) \geq a + b$ holds for some $1 \leq i \leq 2N$. Note that $\mathrm{Loss}_L(A, S) = \mathrm{Loss}_L(A, S_1) + \mathrm{Loss}_L(A(S_1), S_2)$. We know that $\mathrm{Loss}_L(A, S_1) \geq \mathrm{Loss}_L(\mathcal{E}_i, S_1) + a$ holds for some $1 \leq i \leq 2N$. Since S_1 is a coupled trial sequence, this implies that for some $1 \leq k \leq N$ we have $\mathrm{Loss}_L(A, S_1) \geq \mathrm{Loss}_L(\mathcal{E}_i, S_1) + a$ both for $i = k$ and for $i = N + k$. We also know that $\mathrm{Loss}_L(A(S_1), S_2) \geq \mathrm{Loss}_L(\mathcal{E}_j, S_2) + b$ holds for some $1 \leq j \leq 2N$. Since S_2 is a simple trial sequence, this implies that $\mathrm{Loss}_L(A(S_1), S_2) \geq \mathrm{Loss}_L(\mathcal{E}_j, S_2) + b$ holds for all $1 \leq j \leq N$ or for all $N + 1 \leq j \leq 2N$. Hence, we have $\mathrm{Loss}_L(A, S_1) \geq \mathrm{Loss}_L(\mathcal{E}_j, S_1) + a$ and $\mathrm{Loss}_L(A(S_1), S_2) \geq \mathrm{Loss}_L(\mathcal{E}_j, S_2) + b$ for $j = k$ or for $j = N + k$, which proves the claim. □

Again, the proof of Lemma 3.22 remains valid if the algorithms are allowed to know the length of the trial sequence beforehand. An obvious induction based on Lemma 3.22 gives the following result.

Theorem 3.23: *For any loss function L and positive integer k, we have $V_L(2^k, k\ell) \geq kV_L(2,\ell)$.*

In particular, if $\lim_{\ell\to\infty} V_L(2,\ell) \geq c\ln 2$ for some constant c, then for $N = 2^k$, Theorem 3.23 implies $\lim_{\ell\to\infty} V_L(N,\ell) \geq c\log_2 N \ln 2 = c\ln N$. Hence, if we were able to prove $\lim_{\ell\to\infty} V_L(2,\ell) \geq c_L \ln 2$ for the constant c_L defined in (3.3), we would again obtain the asymptotic lower bound $V_L(N,\ell) \geq (c_L - o(1))\ln N$ stated in Theorem 3.1. However, this new bound would be stronger because the term $o(1)$ approaches 0 as ℓ approaches ∞ for all N of the form $N = 2^k$, whereas in the bound of Theorem 3.1 the term $o(1)$ is stated to approach 0 only when both N and ℓ approach ∞.

To obtain the lower bound $V_L(N,\ell) \geq (1/2 - o(1))\ln N$ given in Theorem 3.1 and Example 3.2 for the square loss by applying Theorem 3.23, we would need to show

$$\lim_{\ell\to\infty} V_L(2,\ell) = \frac{\ln 2}{2} \quad . \tag{3.29}$$

We conjecture that (3.29) indeed is true. We have numerically obtained lower bounds such as $V_L(2,500) \geq 0.3456$, while $(\ln 2)/2 \approx 0.3466$. (Obviously $V_L(2,\ell)$ is an increasing function of ℓ, and $V_L(2,\ell) \leq (\ln 2)/2$ by the upper bound of Theorem 3.1 and Example 3.2.) These numerical results are based on a recurrence we have not been able to solve in a closed form. Note that for the square loss, the simple construction used for the logarithmic loss does not yield an optimal lower bound. If we have $\ell = 1$ and $N = 2$, with $\mathbf{x}_1 = (0,1)$, we have $V_{L,A}((\mathbf{x}_1, y_1)) \leq 1/4 = 0.25$ for the algorithm A that predicts $1/2$, and this bound falls short of the required $(\ln 2)/2 \approx 0.3466$.

The preceding remarks show that for the logarithmic loss we have $\lim_{\ell\to\infty} V_L(2^k,\ell) = k\lim_{\ell\to\infty} V_L(2,\ell)$. It is an interesting open question to see which loss functions L have this property. Theorem 3.23 gives $\lim_{\ell\to\infty} V_L(2^k,\ell) \geq k\lim_{\ell\to\infty} V_L(2,\ell)$ for all loss functions. To show equality it is sufficient to show $\lim_{\ell\to\infty} V_L(2,\ell) \geq c_L\ln 2$, and our conjecture is that this is true for the square loss.

4 Continuous-valued outcomes

4.1 Applying the Generic Algorithm

We now show that under certain assumptions, The Generic Algorithm 3.7 also works for continuous-valued outcomes $y_t \in [0,1]$. These assumptions hold for the square and relative entropy loss, but not for the absolute loss, which will be considered in Subsection 4.2. We also consider the more general situation where the values $x_{t,i}$ and y_t are not in the range $[0,1]$.

Lemma 4.1: *Assume that for all $y, a, b \in [0,1]$, the function g defined by $g(y,a,b) = L(y,a)/c - \eta L(y,b)$ satisfies*

$$\frac{\partial^2 g(y,a,b)}{\partial y^2} + \left(\frac{\partial g(y,a,b)}{\partial y}\right)^2 \geq 0 \quad . \tag{4.1}$$

If (3.12) holds for binary values $y \in \{0,1\}$, then it holds for all values $y \in [0,1]$.

Proof We write (3.12) as $(L(y,\hat{y}_t) - \Delta(y))/c \le 0$. By exponentiating both sides and applying (3.11), this becomes

$$e^{L(y,\hat{y}_t)/c} \sum_{i=1}^{N} v_{t,i} e^{-\eta L(y,x_{t,i})} \le 1 \quad . \tag{4.2}$$

Let us denote the left-hand side of (4.2) by $f(y)$. Then

$$f(y) = \sum_{i=1}^{N} v_{t,i} e^{g(y,\hat{y}_t,x_{t,i})} \quad ,$$

so for the second derivative of F we get

$$\frac{\partial^2 f(y)}{\partial y^2} = \sum_{i=1}^{N} v_{t,i} \left(\frac{\partial^2 g(y,\hat{y}_t,x_{t,i})}{\partial y^2} + \left(\frac{\partial g(y,\hat{y}_t,x_{t,i})}{\partial y} \right)^2 \right) e^{g(y,\hat{y}_t,x_{t,i})} \quad .$$

As our assumption implies this to be nonnegative, the maximum value of f for y in the interval $[0,1]$ occurs for $y = 0$ or $y = 1$. Since (3.12) is equivalent to $f(y) \le 1$ for $y \in \{0,1\}$, this proves our claim. □

Theorem 4.2: *Let L be a loss function for which the constant c_L is finite and the condition (4.1) holds for $c = c_L$ and $\eta = 1/c_L$. Let A be the Generic Algorithm 3.7 with the parameters $c = c_L$, $\eta = 1/c_L$, and the initial weights $w_{1,i} = 1$ for all i. Let $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$ be a trial sequence for which $\mathbf{x}_t \in [0,1]^N$ and $y_t \in [0,1]$ hold for all t. Then the algorithm does not fail during the trial sequence, and its additional loss satisfies*

$$V_{L,A}(S) \le c_L \ln N \quad .$$

Proof First note that by Lemma 3.10, the algorithm A does not fail. By Lemma 4.1, the predictions $\hat{y}_t$ of the algorithm satisfy $L(y_t,\hat{y}_t) \le \Delta(y_t)$. We then proceed as in the proof of Theorem 3.8, and obtain the claimed bound by choosing $w_{1,i} = 1$ for all i. □

Example 4.3: Let L be the relative entropy loss L_{ent}. We have

$$\frac{\partial L(y,z)}{\partial y} = \ln y - \ln(1-y) - \ln z + \ln(1-z) \quad ,$$

so the second derivative $\partial^2 L(y,z)/\partial y^2 = 1/y + 1/(1-y)$ does not depend on z. Hence, if $c = 1/\eta$, the second derivative of the function g of Lemma 4.1 is 0, and (4.1) holds. Recall that $c_L = 1$ for the relative entropy loss. Hence, by Theorem 4.2, if A is the Generic Algorithm 3.7 with $c = \eta = 1$, we have $V_{L,A}(S) \le \ln N$ for any N-expert trial sequence S even if the outcomes $y_t \in [0,1]$ are continuous-valued. □

Example 4.4: Let L be the square loss L_{sq}. As the second derivative $\partial^2 L(y,z)/\partial y^2$ is constant, the second derivative of the function g of Lemma 4.1 is 0 whenever $c = 1/\eta$, and hence (4.1) trivially holds. Since $c_L = 1/2$, we let A be the Generic Algorithm 3.7 with $c = 1/2$ and $\eta = 2$. Then by Theorem 4.2 we have $V_{L,A}(S) \le \frac{1}{2} \ln N$ even if the trial sequence S contains continuous-valued outcomes.

Consider now the more general case that at trial t, the experts' predictions $x_{t,i}$ and the outcome y_t are in a known range $[s_t, s_t + r_t]$. Let $x'_{t,i} = (x_{t,i} - s_t)/r_t$ and $y'_t = (y_t - s_t)/r_t$, and let $\hat{y}'_t$ be the prediction of the Generic Algorithm when it is given these scaled inputs $x'_{t,i}$ and

outcomes y'_t. Then Theorem 3.8 applies to this scaled sequence of trials. For an algorithm that predicts with $\hat{y}_t = s_t + r_t\hat{y}'_t$ we then have the following loss bound, if we choose $\eta = 2$ and the initial weights to be equal:

$$\sum_{i=1}^{\ell}\left(\frac{y_t - \hat{y}_t}{r_t}\right)^2 \le \min_{1\le i\le N}\sum_{i=1}^{\ell}\left(\frac{y_t - x_{t,i}}{r_t}\right)^2 + \frac{\ln N}{2} \quad . \tag{4.3}$$

We can consider (4.3) as giving a loss bound similar to (3.14), but with a loss function that changes dynamically as the ranges of $x_{t,i}$ and y_t vary. Note that achieving this bound requires that s_t and r_t are known before the prediction $\hat{y}_t$ is to be made. This is the case, for instance, if the outcome y_t is assumed to be within the range defined by the smallest and largest expert prediction at trial t. Another special case is that before the first trial, we know that $x_{t,i}$ and y_t will always be in some range $[S, S+R]$. We can then take $r_t = R$ for all t, and (4.3) is equivalent with

$$\sum_{i=1}^{\ell}(y_t - \hat{y}_t)^2 \le \min_{1\le i\le N}\sum_{i=1}^{\ell}(y_t - x_{t,i})^2 + \frac{R^2\ln N}{2} \quad .$$

Note that if the range of y_t is not bounded, loss bounds of the above form cannot be attained. To see that, let $N = 2$, and consider a one-trial sequence in which the first prediction vector is $(-R/2, R/2)$. The outcome is chosen by an adversary to be either $y_1 = R/2 + \sqrt{K}$ or $y_1 = -R/2 - \sqrt{K}$, depending on whether the algorithm's prediction was negative or not. Then the loss of the best expert is K, and the loss of the algorithm is at least $\left(R/2 + \sqrt{K}\right)^2 = K + R\sqrt{K} + R^2/4$. Thus, if we let K grow, the additional loss of the algorithm grows as $\Omega\left(\sqrt{K}\right)$. □

Since the absolute loss L_{abs} does not even have a first derivative everywhere, the technique of Lemma 4.1 does not give any results for this loss function. In the next subsection we devise a new algorithm particularly for this problem.

4.2 The Vee Algorithm

We now show how the loss bounds obtained for the absolute loss with binary outcomes can also be achieved when the outcomes are continuous-valued. The results of this section were obtained independently by Vovk [18].

We call our algorithm the Vee Algorithm. In choosing the prediction it is now necessary to explicitly also consider other outcomes than just $y = 0$ and $y = 1$. We will show that the prediction can still be computed in time $O(N \log N)$.

Algoritm 4.5 (The Vee Algorithm): As the Generic Algorithm 3.7, except that we have fixed the loss function to be the absolute loss, the parameter c to be $\left(2\ln\frac{2}{1+e^{-\eta}}\right)^{-1}$, and predicting is done as follows:

Prediction: On receiving the tth input $\mathbf{x}_t$, let $Y = \{0, 1, x_{t,1}, \ldots, x_{t,N}\}$ and $v_{t,i} = w_{t,i}/W_t$. Predict with any value $\hat{y}_t$ that satisfies the condition

$$\max_{y\in Y}\{y - \Delta(y)\} \le \hat{y}_t \le \min_{y\in Y}\{y + \Delta(y)\} \quad , \tag{4.4}$$

where

$$\Delta(y) = -\frac{\ln(\sum_{i=1}^{N} v_{t,i}e^{-\eta|y - x_{t,i}|})}{2\ln\frac{2}{1+e^{-\eta}}} \quad .$$

It is easy to see how the prediction $\hat{y}_t$ can be obtained in time $O(N)$ once the values

$$s(y) = \sum_{i=1}^{N} v_{t,i} e^{-\eta |y - x_{t,i}|}$$

have been obtained for all the $N + 2$ choices of y. Let $\mathbf{x}'_t$ be a vector that contains the components of the prediction vector $\mathbf{x}_t$ sorted into an ascending order. Thus, $x'_{t,i} \leq x'_{t,i+1}$ for $1 \leq i \leq N - 1$. The vector $\mathbf{x}'_t$ can be obtained in time $O(N \log N)$. We show how all the sums $s(y)$ for $y \in \{ 0, x_{t,1}, \ldots, x_{t,N}, 1 \}$ can be obtained in time $O(N)$ given the sorted prediction vector $\mathbf{x}'_t$. To unify notation, write $x'_{t,0} = 0$ and $x_{t,N+1} = 1$. Note that for $0 \leq j \leq N + 1$ we can write $s(x'_{t,j}) = a_j + b_j$ where

$$a_j = \sum_{i=1}^{j} v_{t,j} e^{-\eta (x'_{t,j} - x'_{t,i})}$$

and

$$b_j = \sum_{i=j+1}^{N} v_{t,j} e^{-\eta (x'_{t,i} - x'_{t,j})} \quad .$$

We have $a_0 = 0$, and b_0 can be computed in time $O(N)$. Further, given a_j and b_j we obtain a_{j+1} and b_{j+1} in time $O(1)$ by

$$a_{j+1} = e^{-\eta (x'_{t,j+1} - x'_{t,j})} a_j + v_{t,j+1}$$

and

$$b_{j+1} = e^{-\eta (x'_{t,j} - x'_{t,j+1})} \left(b_j - v_{t,j+1} e^{-\eta (x'_{t,j+1} - x_{t,j})} \right) \quad .$$

Hence, the prediction $\hat{y}_t$, if it exists, can be found in total time $O(N \log N)$.

We see in Lemma 4.6 that there always is a prediction $\hat{y}_t$ that satisfies (4.4) and that (4.4) implies $|y - \hat{y}_t| \leq \Delta(y)$ for *all* $y \in [0, 1]$ and not merely for $y \in \{ 0, 1 \}$, which was the requirement in the Generic Algorithm. Hence, we now get for continuous-valued outcomes $y_t \in [0, 1]$ the bound (3.20) that was previously obtained for binary outcomes $y_t \in \{ 0, 1 \}$. Note that if (3.20) holds for $y_t \in [0, 1]$, it actually holds for all y_t, provided we still have $x_{t,i} \in [0, 1]$. This is because moving y_t outside the range of the experts' predictions increases every $|y_t - x_{t,i}|$ as much as it increases $|y_t - \hat{y}_t|$, and the coefficient $\eta / (2 \ln \frac{2}{1 + e^{-\eta}})$ that appears in front of $|y_t - x_{t,i}|$ in (3.20) is greater than 1. Again, the parameter η can be tuned as mentioned in Example 3.14, and the scaling method of Example 4.4 can be used if the values $x_{t,i}$ are not in the range $[0, 1]$.

For the absolute loss, (3.12) has a simple geometric interpretation. Figure 4.1 gives an example of the graphs of the left-hand side $|y - \hat{y}|$ and the right-hand side $\Delta(y)$ as functions of y, fixing $\hat{y} = 0.58$ and $\mathbf{x} = (0.33, 0.83, 0.97, 0.52)$. The left-hand side of the inequality is represented by a vee-curve with its tip at $(\hat{y}, 0)$. The graph of Δ has a nondifferentiable tip at each value $y = x_i$. The condition (3.12) states that the vee-curve must be below the graph of Δ at y. For continuous-valued outcomes we wish (3.12) to hold for $y \in [0, 1]$ and hence the vee-curve to be below the graph of Δ everywhere. If we were to move the tip of the vee to the left of 0.51, the right arm of the vee would intersect the Δ-curve at the value $y = 0.97$. Hence, the value of the maximum on the left-hand side of (4.4) is roughly 0.51. Similarly, the minimum on the right-hand side is about 0.63, since moving the tip of the vee over this value would make its left arm intersect the Δ-curve at $y = 0.33$. For binary outcomes we only required (3.12) to hold for $y = 0$ and $y = 1$, which gives the weaker condition that the vee-curve must be below the graph of Δ at the endpoints.

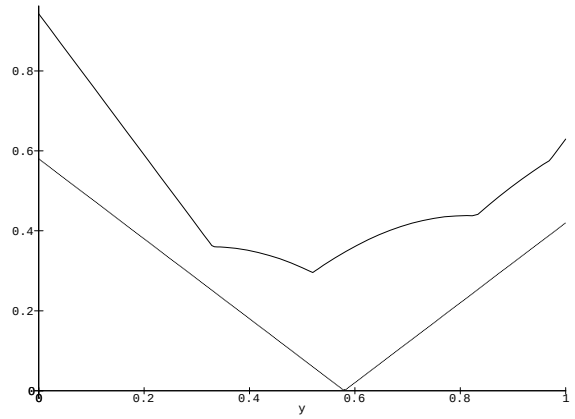

Figure 4.1: Example graphs of the functions Δ (above) and L_{abs} (below).

For binary outcomes, the loss bound (3.20) was previously shown for a whole family of algorithms defined by a number of different prediction and update factors $\alpha_{t,i}$ [2], as was briefly explained in Example 3.14. In the continuous case we have less freedom. Suppose we were to use $\alpha_{t,i} = 1 - (1 - e^{-\eta})x_{t,i}$, and let $N = 1$, $\mathbf{x} = (0)$, and $\eta = 1$. We would then have $\Delta(0) = 0$ and $1 - \Delta(1) \approx -0.316$, and hence the condition (4.4) would not hold for any $\hat{y}_t$. The Algorithm WMC [13] does work for the continuous case, and is allowed to use any update that satisfies (3.21). However, its worst case bound has $1 - e^{-\eta}$ in the denominator instead of $2 \ln \frac{2}{1+e^{-\eta}}$, and hence it is slightly worse than the bounds given here.

As we noticed in Example 3.14, for binary outcomes it was possible to choose the prediction $\hat{y}_t$ as a function of the weighted average of the experts' predictions. If the outcomes are allowed to be continuous-valued, this is not possible any more. To see that there is no function f such that $\hat{y}_t = f(\sum_i v_{t,i} x_{t,i})$ guarantees (4.4) to hold, we consider two cases. First, let $\mathbf{v}_t = (0.3, 0.7)$ and $x = (0, 1)$, so $\sum_i v_{t,i} x_{t,i} = 0.7$. For the value $\eta = 1$, the left-hand side of (4.4) is approximately 0.72, and we obtain a constraint $0.72 \leq f(0.7)$ for f. On the other hand, considering $\mathbf{v}_t = (1, 0)$ and $\mathbf{x}_t = (0.7, 0)$ on the right-hand side of (4.4) gives a contradictory constraint $f(0.7) \leq 0.70$.

We now show that a prediction that satisfies (4.4) always exists and satisfies the conditions of Theorem 3.8.

Lemma 4.6: *Let $\mathbf{v}_t \in [0,1]^N$ with $\sum_i v_{t,i} = 1$ and $\mathbf{x}_t \in [0,1]^N$, and let $\eta > 0$. Then a prediction $\hat{y}_t$ that satisfies (4.4) exists. Further, (4.4) implies $|y - \hat{y}_t| \leq \Delta(y)$ for all $y \in [0,1]$.*

Proof We prove the existence of $\hat{y}_t$ by showing that

$$y - \Delta(y) \leq z + \Delta(z) \tag{4.5}$$

holds for all y, z, $\mathbf{v}_t$, and $\mathbf{x}_t$. Define

$$g(\mathbf{v}, \mathbf{x}, y, z) = \sum_{i=1}^{N} \sum_{j=1}^{N} v_i v_j \exp\left(-\eta(|y - x_i| + |z - x_j|) + (y - z) 2 \ln(2/(1 + e^{-\eta}))\right) \quad . \tag{4.6}$$

Then (4.5) is equivalent to $g(\mathbf{v}_t, \mathbf{x}_t, y, z) \leq 1$. The second derivative $\partial^2 g(\mathbf{v}, \mathbf{x}, y, z)/\partial x_i^2$ is defined and positive if $x_i \notin \{0, y, z, 1\}$. Thus it suffices to show $g(\mathbf{v}, \mathbf{x}, y, z) \leq 1$ for $N = 4$ and $\mathbf{x} = \mathbf{x}_a = (0, y, z, 1)$. In this restricted case the second derivative $\partial^2 g(\mathbf{v}, \mathbf{x}_a, y, z)/\partial z^2$ is positive if $z \notin \{0, y, 1\}$. Furthermore, since $\Delta(z) \geq 0$, (4.5) trivially holds if $z \geq y$. Thus it suffices to show (4.5) for $z = 0$, $y > 0$ and $\mathbf{x} = \mathbf{x}_b = (0, y, 0, 1)$. Finally, since the second derivative $\partial^2 g(\mathbf{v}, \mathbf{x}_b, y, 0)/\partial y^2$ is positive, we are left with the case $z = 0$, $y = 1$ and $\mathbf{x} \in \{0, 1\}^N$. In this case, the original inequality (4.5) can be rewritten as

$$\frac{\ln((1-p)e^{-\eta} + p) + \ln(1 - p + pe^{-\eta})}{2} \leq \ln \frac{1 + e^{-\eta}}{2}$$

where $r = \sum_i v_i x_i$. This holds for all $0 \leq p \leq 1$ because the function ln is concave.

A similar argument based on second derivatives shows that for $y \in [0, 1]$, the value $y - \Delta(y)$ obtains its maximum and the value $y + \Delta(y)$ its minimum when $y \in \{0, 1, x_{t,1}, \ldots, x_{t,N}\}$. □

Lemma 4.6 immediately implies the following result.

Theorem 4.7: *Let $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$ be a trial sequence with $\mathbf{x}_t \in [0, 1]^N$ and $y_t \in [0, 1]$ for all t. Let L be the absolute loss and A be the Vee Algorithm 4.5. We then have*

$$\mathrm{Loss}_L(A, S) \leq \frac{-\ln \frac{w_{1,i}}{W_1} + \eta \mathrm{Loss}_L(\mathcal{E}_i, S)}{2 \ln \frac{2}{1+e^{-\eta}}}$$

for all i.

5 Further work

One of the most challenging open problems is to give tight bounds for the additional loss of the prediction algorithm compared to the loss of the best expert for even more general classes of loss functions than those considered in this paper. When the outcomes y_t are binary, it might be possible to produce such bounds for arbitrary loss functions. The next challenge is to extend the results for continuous-valued outcomes to more general loss functions. Another direction worth exploring is to let outcomes be discrete valued with more than two choices. The recent results of Chung [5] address some of these problems.

In this paper we restricted the predictions of the experts to lie between zero and one, except in specific examples where we have indicated how scaling tricks can be used. It would be nice to do a thorough investigation of how scaling the range of the variables affects the results. Bounding some norm of the prediction vector might also lead to interesting problems. Restricting the range of the predictions of individual experts is related to bounding the infinity norm of the prediction vectors.

It would be interesting to see whether the alternative update rules defined by (3.21) for the absolute loss work for other loss functions. As we have seen, it is sometimes possible to obtain the prediction as a function of the weighted average of the experts' predictions. We would like to know exactly when this simplification is possible without weakening our bounds, or with weakening them only slightly.

In this paper we have given bounds of the additional loss of our algorithms over the loss of the best expert. A more challenging problem is to bound the additional loss of the algorithms over the best linear combination of experts [12, 4, 11]. The only worst-case loss bounds for the latter case that have been obtained were for the square loss function. Hopefully, some of the

results of the present paper can be generalized to the linear combination case. An intermediate case worth exploring is the case of bounding the additional loss of the algorithm compared with the best "stretched" expert, i.e., an original expert multiplied by some positive constant.

Acknowledgments

We would like thank David P. Helmbold for his significant help in developing the Vee Algorithm.

References

[1] Patrick Billingsley. *Probability and Measure*. Wiley, New York, NY, 1986. Second Edition.

[2] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. E. Schapire, and M. K. Warmuth. How to use expert advice. Technical Report UCSC-CRL-94-33, Univ. of Calif. Computer Research Lab, Santa Cruz, CA, 1994. An extended abstract appeared in STOC '93.

[3] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, and M. Warmuth. On-line prediction and conversion strategies. In *Computational Learning Theory: Eurocolt '93*, volume New Series Number 53 of *The Institute of Mathematics and its Applications Conference Series*, pages 205–216, Oxford, 1994. Oxford University Press.

[4] N. Cesa-Bianchi, P. Long, and M. Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. Technical Report UCSC-CRL-93-36, Univ. of Calif. Computer Research Lab, Santa Cruz, CA, 1993. An extended abstract appeared in COLT '93.

[5] T. H. Chung. Approximate methods for sequential decision making using expert advice. In *Proc. 7th Annu. ACM Workshop on Comput. Learning Theory*, pages 183–189. ACM Press, New York, NY, 1994.

[6] T. Cover. Behavior of sequential predictors of binary sequences. In *Proceedings of the 4th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 263–272. Publishing House of the Czechoslovak Academy of Sciences, 1965.

[7] A. P. Dawid. Prequential analysis, stochastic complexity and bayesian inference. *Bayesian Statistics*. To appear.

[8] A. DeSantis, G. Markowsky, and M. N. Wegman. Learning probabilistic prediction functions. In *Proc. 29th Annu. IEEE Sympos. Found. Comput. Sci.*, pages 110–119. IEEE Computer Society Press, Los Alamitos, CA, 1988.

[9] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.

[10] Janos Galambos. *The Asymptotic Theory of Extreme Order Statistics*. R. E. Krieger, Malabar, FL, 1987. Second Edition.

[11] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. Technical Report UCSC-CRL-94-16, University of California, Santa Cruz, Computer Research Laboratory, June 1994.

[12] N. Littlestone, P. M. Long, and M. K. Warmuth. On-line learning of linear functions. In *Proc. of the 23rd Symposium on Theory of Computing*, pages 465–475. ACM Press, New York, NY, 1991.

[13] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[14] N. Merhav and M. Feder. Universal sequential learning and decisions from individual data sequences. In *Proc. 5th Annu. Workshop on Comput. Learning Theory*, pages 413–427. ACM Press, New York, NY, 1992.

[15] J. Mycielski. A learning algorithm for linear operators. *Proceedings of the American Mathematical Society*, 103(2):547–550, 1988.

[16] V. Vovk. Aggregating strategies. In *Proc. 3rd Annu. Workshop on Comput. Learning Theory*, pages 371–383. Morgan Kaufmann, 1990.

[17] V. Vovk. Universal forecasting algorithms. *Inform. Comput.*, 96(2):245–277, 1992.

[18] V. Vovk. Unpublished manuscript, October 1994.

[19] M. J. Weinberger, N. Merhav, and M. Feder. Optimal sequential probability assignment for individual sequences. *IEEE Transactions on Information Theory*, 40(2):384–396, March 1994.