Tight Worst-Case Loss Bounds for Predicting with Expert Advice

David Haussler, Jyrki Kivinen, and Manfred K. Warmuth

Computer and Information Sciences, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

Abstract. We consider on-line algorithms for predicting binary 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.

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. [1]. Similar frameworks have also been considered by Cover [5], Dawid [6], Feder et al. [8, 14, 18], and Mycielski [15]. See Chung [4] for recent related results.

In this paper, we consider the special case in which the outcomes are restricted to be 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 $y_t = 1$. Many of the results can be generalized for continuous-valued outcomes $y_t \in [0, 1]$ [10]. Cesa-Bianchi et al. [2] have considered the case in which both the outcomes and the predictions of the experts and the algorithm are required to be binary.

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

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

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 $c_L = 1$ [7, 16]. Note that the bound (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. [1] have shown that bounds of the form (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 already obtained 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). 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) for all trial sequences S, but also for some trial sequence S the lower bound

$$\mathrm{Loss}_L(A,S) - \min_{1\le i\le N} \mathrm{Loss}_L(\mathcal{E}_i,S) \ge (c_L - o(1)) \ln N \ , \qquad (2)$$

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) and (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

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

Hence, for the loss functions in this class, an upper bound like (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 Sect. 3 together with a discussion of the condition the loss function must satisfy for the bounds to be applicable. Sect. 4 restates Vovk's algorithm and upper bound proof simplified for our purposes. The lower bound proof, sketched in Sect. 5, 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) 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. [1] in the special case of the absolute loss.

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 $t = 1, \ldots, \ell$. Here we consider only *binary outcomes*, with the *outcomes* y_t either 0 or 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 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 it is possible to fine-tune the algorithms if ℓ is known in advance [1].

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 $L_0(\hat{y}) = L(0, \hat{y})$ and $L_1(\hat{y}) = L(1, \hat{y})$.

Example 1. The *relative entropy loss* L_{ent} is defined by $L_{\text{ent}}(y, \hat{y}) = y \ln \frac{y}{\hat{y}} + (1-y) \ln \frac{1-y}{1-\hat{y}}$. 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 $L = L_{\text{ent}}$. 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 $L_{\text{sq}}(y, \hat{y}) = (y - \hat{y})^2$. Hence, for $L = L_{\text{sq}}$, 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 $L_{\text{H}}(y, \hat{y}) = \frac{1}{2}\left(\left(\sqrt{1-y} - \sqrt{1-\hat{y}}\right)^2 + \left(\sqrt{y} - \sqrt{\hat{y}}\right)^2\right)$. Hence, for $L = L_{\text{H}}$ 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 $L_{\text{abs}}(y, \hat{y}) = |y - \hat{y}|$, and we have $L_0(\hat{y}) = \hat{y}$ and $L_1(\hat{y}) = 1 - \hat{y}$ for $L = L_{\text{abs}}$.

It is worth noting some properties of the loss functions of Example 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 $\text{Loss}_L(A, S) = \sum_{t=1}^{\ell} L(y_t, \hat{y}_t)$ as the loss of the algorithm and $\text{Loss}_L(\mathcal{E}_i, S) = \sum_{t=1}^{\ell} L(y_t, x_{t,i})$ as the loss of the ith expert on the sequence S. We define

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

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

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

be the worst case amount of additional loss for A. 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. 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. For these loss functions there are algorithms that satisfy $V_{L,A}(N,\ell) \leq \frac{1}{2}\ln N$ and $V_{L,A}(N,\ell) \leq \ln N$, respectively [16, 7]. 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 $V_{L,A}(N,\ell) = \Omega\left(\sqrt{\ell \ln N}\right)$ [1]. Many of the upper bounds hold even with continuous-valued outcomes $y_t \in [0,1]$ [10].

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.

3 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

$$S(z) = L_0'(z)L_1''(z) - L_1'(z)L_0''(z) \tag{3}$$

and a function R by

$$R(z) = \frac{L_0'(z)L_1'(z)^2 - L_1'(z)L_0'(z)^2}{S(z)} \ . \tag{4}$$

We then define a constant c_L by

$$c_L = \sup_{0<z<1} R(z) \ . \tag{5}$$

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 $V_{L,A}(N,\ell) \leq c_L \ln N$ and show that the bound is asymptotically tight.

Theorem 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 (5) is finite and $S(z)$ defined in (3) is positive for $0 < z < 1$. Then there is an on-line prediction algorithm A for which*

$$V_{L,A}(N,\ell) \leq c_L \ln N \ . \tag{6}$$

Further, we have

$$V_L(N,\ell) \geq (c_L - o(1)) \ln N \ , \tag{7}$$

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

The algorithm A that obtains the bound (6), 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 Sect. 4. The lower bound (7) is based on a probabilistic proof that is sketched in Sect. 5. The lower bound also holds for algorithms that get knowledge of ℓ beforehand.

Example 2. Consider the loss functions of Example 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 (4) 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. [1]. 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

$$(1-q)L_0'(z) + qL_1'(z) = 0 \ . \tag{8}$$

If $1 - q \neq 0$ and $L_1'(z) \neq 0$, this implies $q/(1-q) = -L_0'(z)/L_1'(z)$. 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 (8) holds, and hence z cannot be a Bayes-optimal prediction for more than one bias. If $(1-q)L_0''(z) + qL_1''(z) > 0$ holds in addition to (8), then z is a local minimum point. There may be one or more Bayes-optimal predictions for a given bias.

Example 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 (8) shows that the unique Bayes-optimal prediction z for a bias $0 < q < 1$ is given by $z = 1/(1 + (1/q - 1)^2)$.

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.

The following lemma gives the connection between $S(z)$ and Bayes-optimality (proof omitted).

Lemma 2. *If $S(z) > 0$ for all z, then for all biases $0 \leq q \leq 1$ there is a unique Bayes-optimal prediction z. If for all biases q the Bayes-optimal prediction is unique, then $S(z) \geq 0$ for all z, and there is no interval $[a,b]$ with $a < b$ such that $S(z) = 0$ for all $z \in [a,b]$.*

In Sect. 5 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. *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. *If $S(z) = 0$ for some $0 < z < 1$, we have $V_L(N,\ell) = \Omega\left(\ell^{1/2-\alpha}\sqrt{\log N}\right)$ 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 $a \leq z \leq b$, we have $V_L(N,\ell) = \Omega\left(\sqrt{\ell \log N}\right)$.*

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 4. 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 $R(z) = (\alpha/(\alpha+1))(z^{-\alpha}(1-z) + (1-z)^{-\alpha}z)$. 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 $z = 1/(1 + (1/q - 1)^{1/(\alpha+1)})$, as can be seen by a straightforward calculation.

4 The Algorithm and the Upper Bound Proof

We consider an algorithm first introduced by Vovk [16]. We give a brief summary of the algorithm in the special case where L_0 is continuous and increasing and L_1 continuous and decreasing. For a more detailed exposition, and examples for various loss functions, see also Kivinen and Warmuth [10].

A continuous and increasing function L_0 has a continuous and increasing inverse defined in $[L_0(0), L_0(1)]$. We say that a function L_0^{-1} defined in $[L_0(0), \infty]$ is a *generalized inverse* of L_0 if $L_0^{-1}(L_0(\hat{y})) = \hat{y}$ for $0 \leq \hat{y} \leq 1$ and, additionally, $L_0^{-1}(z) \geq 1$ for $z \geq L_0(1)$. We further assume that $L_0(0) = 0$. Then with the constraint $0 \leq \hat{y} \leq 1$, $\hat{y} \leq L_0^{-1}(z)$ is equivalent with $L_0(\hat{y}) \leq z$ for $z \geq 0$. Similarly, assuming that L_1 is continuous and decreasing and satisfies $L_1(1) = 0$, we call L_1^{-1} a generalized inverse of L_1 if $L_1^{-1}(L_1(\hat{y})) = \hat{y}$ holds for $0 \leq \hat{y} \leq 1$ and, additionally, $L_1^{-1}(z) \leq 0$ for $z \geq L_1(0)$.

For instance, if L is the square loss L_{sq}, we have the generalized inverses $L_0^{-1}(z) = \sqrt{z}$ and $L_1^{-1}(z) = 1 - \sqrt{z}$ for $0 \leq z \leq 1$. For the relative entropy loss L_{ent} we have $L_0^{-1}(z) = 1 - e^{-z}$ and $L_1^{-1}(z) = e^{-z}$. For the absolute loss L_{abs} we have $L_0^{-1}(z) = z$ and $L_1^{-1}(z) = 1 - z$.

Algorithm 4 (The Generic Algorithm) Let L_0 be continuous and increasing and L_1 continuous and decreasing, with $L_0(0) = L_1(1) = 0$. Let L_0^{-1} and L_1^{-1} be generalized inverses of L_0 and L_1. Let c and η be arbitrary 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 $\Delta(0)$ and $\Delta(1)$ where

$$\Delta(y) = -c \ln \sum_{i=1}^{N} v_{t,i} e^{-\eta L(y, x_{t,i})} \ . \tag{9}$$

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

$$L_1^{-1}(\Delta(1)) \leq \hat{y} \leq L_0^{-1}(\Delta(0)) \ . \tag{10}$$

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

Update: After receiving the tth outcome y_t, let

$$w_{t+1,i} = w_{t,i} e^{-\eta L(y_t, x_{t,i})} . \tag{11}$$

Given values c and η, we say that the loss function L is (c, η)-*realizable* if it never fails with these parameter values, i.e., if for these values of c and η the inequality $L_1^{-1}(\Delta(1)) \leq L_0^{-1}(\Delta(0))$ holds for all possible $\mathbf{w}_t$ and $\mathbf{x}_t$. The main technical problem in the analysis of the algorithm is finding for a given loss function L values η and c for which L is (c, η)-realizable. For now, assume that such values of c and η are given.

To understand the algorithm, note that by (9) and (11) we can write $\Delta(y_t) = U_{t+1} - U_t$, where $U_t = -c \ln W_t$. The condition (10) implies $L(y_t, \hat{y}_t) \leq \Delta(y_t)$

both for $y_t = 0$ and $y_t = 1$. Hence, we can consider $-c \ln W_t$ as a potential function, and the condition (10) means that at each trial, the increase of the potential must be at least as large as the loss of the algorithm. Adding these bounds for $t = 1, \ldots, \ell$ gives us $\text{Loss}_L(A, S) \leq -c \ln (W_{\ell+1}/W_1)$. Finally, by noting that for $1 \leq i \leq N$ we have $W_{\ell+1} \geq w_{\ell+1,i} = w_{1,i} \exp(\text{Loss}_L(\mathcal{E}_i, S))$ we obtain the following basic bound [16].

Theorem 5. *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. Assume that during this trial sequence, the Generic Algorithm 4 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*

$$\text{Loss}_L(A, S) \leq -c \ln \frac{W_{\ell+1}}{W_1} \leq -c \ln \frac{w_{1,i}}{W_1} + c\eta \text{Loss}_L(\mathcal{E}_i, S) \ . \tag{12}$$

We now need to choose values c and η for which L is (c, η)-realizable and the right-hand side of (12) is as small as possible. Our main goal is to have $\eta = 1/c$, which makes the ratio $\text{Loss}_L(A, S)/\text{Loss}_L(\mathcal{E}_i, S)$ approach 1 as the losses increase. The second goal is to make c as small as possible while keeping $\eta = 1/c$.

Lemma 6. *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 (5) is finite, and $S(z)$ defined in (3) is positive for all z. Then the loss function L is $(c, 1/c)$-realizable if and only if $c \geq c_L$.*

Proof sketch. Define $p(z) = \exp(-L_0(z)/c)$ and $q(z) = \exp(-L_1(z)/c)$ for $0 \leq z \leq 1$, and define $f(r) = \exp(-L_1(L_0^{-1}(-c \ln r))/c)$ for r in the range of p. Note that $f(p(z)) = q(z)$. A straightforward computation shows that $f''(p(z)) \leq 0$ holds if and only if $c \geq R(z)$. Hence, f is concave if and only if $c \geq c_L$. Assume now that this is the case.

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 $\sum_i v_{t,i} s_i = \sum_i v_{t,i} f(r_i) \leq f(\sum_i v_{t,i} r_i)$ and, hence, $-c \ln \sum_i v_{t,i} s_i \geq -c \ln f(\sum_i v_{t,i} r_i)$. This is equivalent with $\Delta(1) \geq L_1(L_0^{-1}(\Delta(0)))$, from which $L_1^{-1}(\Delta(1)) \leq L_0^{-1}(\Delta(0))$ follows since L_1^{-1} is decreasing. □

In particular, we see that since the Generic Algorithm 4 does not fail with the parameters $c = c_L$ and $\eta = 1/c_L$, we get the upper bound claimed in Theorem 1 by applying Theorem 5 with the initial weights $w_{1,i} = 1$ for all i.

Theorem 7. *Let L be a loss function for which the constant c_L is finite and the value $S(z)$ positive for all z. Let A be the Generic Algorithm 4 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 $V_{L,A}(N, \ell) \leq c_L \ln N$.*

5 Lower Bounds Proofs

This section contains proofs of the lower bounds for $V_L(N,\ell)$ stated in Theorems 1 and 3 in Sect. 3. The lower bounds hold even for algorithms that receive ℓ as input before the first trial. Theorem 9 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 9 is based on Lemma 8, 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 9 is in terms of certain characteristics of the probability measures, and is interesting only if the probability measures are chosen carefully. Lemma 10 shows a particular way of choosing the probability measures, when a prediction b is the unique Bayes-optimal prediction for a bias q. Lemma 11 show a way to choose the probability measures in Theorem 9 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 10 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 11 yields a lower bound $V_L(N,\ell) = \Omega\left(\sqrt{\ell \log N}\right)$.

We begin with a technical lemma that shows that under the conditions that arise in our main proof, certain expectations converge as we desire.

Lemma 8. *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*

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

Theorem 9 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 9. *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 $\ell \ge \ell_\varepsilon$ we have*

$$V_L(N,\ell) \ge \ell \mathrm{E}_{y\in Q}[L(y,b)] - \ell\tau + (a_N - \varepsilon)\sigma\sqrt{\ell \ln N} \ , \qquad (13)$$

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 (13) if we can prove that

$$\mathrm{E}_{\mathbf{x} \in P^{N \times \ell}} \mathrm{E}_{\mathbf{y} \in Q^{\ell}} \left[\min_{1 \leq i \leq N} \mathrm{Loss}_L(\mathcal{E}_i, \langle \mathbf{x}, \mathbf{y} \rangle) \right] \leq \ell\tau - (a_N - \varepsilon)\sigma\sqrt{\ell \ln N} . \quad (14)$$

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 $S_{i\ell}^{\mathbf{y}}$ by $S_{i\ell}^{\mathbf{y}}(\mathbf{x}) = S_{i\ell}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\ell} T_{ij}^{\mathbf{y}}(\mathbf{x})$. 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 $T_{ij}^{\mathbf{y}}$ and $T_{i'j'}^{\mathbf{y}}$ are independent for $(i,j) \neq (i',j')$. To study the distribution of $S_{i\ell}^{\mathbf{y}}$, we define a suitably normalized random variable $U_{i\ell}^{\mathbf{y}}$. Let

$$U_{i\ell}^{\mathbf{y}} = \frac{S_{i\ell}^{\mathbf{y}} - \sum_{j=1}^{\ell} \mathrm{E}[T_{ij}^{\mathbf{y}}]}{\sqrt{\sum_{j=1}^{\ell} \mathrm{Var}[T_{ij}^{\mathbf{y}}]}} . \quad (15)$$

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

We now apply Lemma 8 to the random variables $U_{i\ell}^{\mathbf{y}}$. Then the random variables F_i in Lemma 8 have standard normal distribution. By a standard result [9], 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 8 now implies

$$\lim_{\ell\to\infty} \mathrm{E}_{\mathbf{y}\in Q^\infty}\left[\frac{\hat{\sigma}_\ell(\mathbf{y})}{\sigma}\mathrm{E}_{\mathbf{x}\in P^{N\times\infty}}\left[\min_{1\leq i\leq N} U_{i\ell}^{\mathbf{y}}\right]\right] = -a_N\sqrt{\ln N} \tag{16}$$

By partitioning the summations in (15) into two parts according to whether $y_i = 0$ or $y_i = 1$, we can write

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

By substituting this into (16), we obtain

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

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

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

This implies (14), as desired. □

We now see how Theorem 9 implies a lower bound for $V_L(N,\ell)$ when the probability measure P for the experts is chosen suitably.

Lemma 10. *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 (4) 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 sketch. We apply Lemma 9 with Q and P such that $\Pr_{y\in Q}[y = 1] = q$ and $\Pr_{x\in P}[x = b - h] = \Pr_{x\in P}[x = b + h] = 1/2$, for some small h. We use second order Taylor approximations of L_0 and L_1 around b to get approximate values for the quantities τ and σ. The quantities q and $1-q$ can be stated in terms of $L_0'(b)$ and $L_1'(b)$ by applying (8). We then choose h that maximizes the resulting approximation for the right-hand side of (13). □

If for some bias q we have two Bayes-optimal predictions b_1 and b_2, applying Theorem 9 to the probability measure P with $\Pr_{x\in P}[x = b_1] = \Pr_{x\in P}[x = b_2] = 1/2$ gives the following result.

Lemma 11. *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 $\sigma^2 = \frac{1-q}{4}(L_0(b_1) - L_0(b_2))^2 + \frac{q}{4}(L_1(b_1) - L_1(b_2))^2$.*

Note that for the absolute loss, we can apply Lemma 11 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. [1].

The following lemma implies that if some $b \in (0,1)$ is not Bayes-optimal for any bias and hence cannot be applied in Lemma 10 to give a lower bound, we obtain a lower bound by applying Lemma 11.

Lemma 12. *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.*

The lower bounds in Theorem 1 and Theorem 3 follow directly from the following theorem.

Theorem 13. *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. *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 (5).*
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 11 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 12, we then have for all predictions z a bias such that z is Bayes-optimal. By Lemma 2, 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 (8) implies $(1 - q)L_0''(z) + qL_1''(z) = S(z)$. If $S(z) = 0$, then applying Lemma 10 (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 10 (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. □

6 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 [10] 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 [4] address some of these problems.

In this paper we restricted the predictions of the experts to lie between zero and one. More general ranges can be allowed if simple scaling methods are applied [10]. It would be interesting 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.

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, 3, 11]. The only worst-case loss bounds for the latter case that have been obtained are 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

David Haussler and Manfred K. Warmuth have been supported by NSF grant IRI-9123692. Jyrki Kivinen has been funded by Emil Aaltonen Foundation, University of Helsinki, and the Academy of Finland.

References

1. Cesa-Bianchi, N., Freund, Y., Helmbold, D. P., Haussler, D., Schapire, R. E., Warmuth, M. K.: 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.
2. Cesa-Bianchi, N., Freund, Y., Helmbold, D. P., Warmuth, M. K.: On-line prediction and conversion strategies. In *Computational Learning Theory: EuroCOLT '93*, pages 205–216. Oxford University Press, Oxford, UK, 1994.
3. Cesa-Bianchi, N., Long, P., Warmuth, M. K.: 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.

4. Chung, T. H.: Approximate methods for sequential decision making using expert advice. In *Proc. 7th ACM Workshop on Computational Learning Theory*, pages 183–189. ACM Press, New York, NY, 1994.
5. Cover, T.: Behavior of sequential predictors of binary sequences. In *Proc. 4th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 263–272. Publishing House of the Czechoslovak Academy of Sciences, 1965.
6. Dawid, A. P.: Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian Statistics* (to appear).
7. DeSantis, A., Markowsky, G., Wegman, M. N.: Learning probabilistic prediction functions. In *Proc. 29th IEEE Symposium on Foundations of Computer Science*, pages 110–119. IEEE Computer Society Press, Los Alamitos, CA, 1988.
8. Feder, M., Merhav, N., Gutman, M.: Universal prediction of individual sequences. *IEEE Transactions on Information Theory* **38** (1992) 1258–1270.
9. Galambos, J.: *The Asymptotic Theory of Extreme Order Statistics*. R. E. Krieger, Malabar, FL, 1987. Second Edition.
10. Kivinen, J., Warmuth, M. K.: Using experts for predicting continuous outcomes. In *Computational Learning Theory: EuroCOLT '93*, pages 109–120. Oxford University Press, Oxford, UK, 1994.
11. Kivinen, J., Warmuth, M. K.: Exponentiated gradient versus gradient descent for linear predictors. Technical Report UCSC-CRL-94-16, Univ. of Calif. Computer Research Lab, Santa Cruz, CA, June 1994.
12. Littlestone, N., Long, P. M., Warmuth, M. K.: On-line learning of linear functions. In *Proc. 23rd ACM Symposium on Theory of Computing*, pages 465–475. ACM Press, New York, NY, 1991.
13. Littlestone, N., Warmuth, M. K.: The weighted majority algorithm. *Information and Computation* **108** (1994) 212–261.
14. Merhav, N., Feder, M.: Universal sequential learning and decisions from individual data sequences. In *Proc. 5th ACM Workshop on Computational Learning Theory*, pages 413–427. ACM Press, New York, NY, 1992.
15. Mycielski, J.: A learning algorithm for linear operators. *Proceedings of the American Mathematical Society* **103** (1988) 547–550.
16. Vovk, V.: Aggregating strategies. In *Proc. 3rd Workshop on Computational Learning Theory*, pages 371–383. Morgan Kaufmann, San Mateo, CA, 1990.
17. Vovk, V.: Universal forecasting algorithms. *Information and Computation* **96** (1992) 245–277.
18. Weinberger, M. J., Merhav, N., Feder, M.: Optimal sequential probability assignment for individual sequences. *IEEE Transactions on Information Theory* **40** (1994) 384–396.