Journal of Computer and System Sciences **64**, 76–102 (2002)
doi:10.1006/jcss.2001.1798, available online at http://www.idealibrary.com on IDEAL®

Relative Expected Instantaneous Loss Bounds

Jürgen Forster and Manfred K. Warmuth

Department of Computer Science, University of California, Santa Cruz, Santa Cruz, California 95064
E-mail: forster@lmi.ruhr-uni-bochum.de, manfred@cse.ucsc.edu

Received October 10, 2000; revised May 29, 2001

In the literature a number of relative loss bounds have been shown for on-line learning algorithms. Here the relative loss is the total loss of the on-line algorithm in all trials minus the total loss of the best comparator that is chosen off-line. However, for many applications instantaneous loss bounds are more interesting where the learner first sees a batch of examples and then uses these examples to make a prediction on a new instance. We show relative expected instantaneous loss bounds for the case when the examples are i.i.d. with an unknown distribution. We bound the expected loss of the algorithm on the last example minus the expected loss of the best comparator on a random example. In particular, we study linear regression and density estimation problems and show how the leave-one-out loss can be used to prove instantaneous loss bounds for these cases. For linear regression we use an algorithm that is similar to a new on-line learning algorithm developed by Vovk. Recently a large number of relative total loss bounds have been shown that have the form $O(\ln T)$, where T is the number of trials/examples. Standard conversions of on-line algorithms to batch algorithms result in relative expected instantaneous loss bounds of the form $O(\frac{\ln T}{T})$. Our methods lead to $O(\frac{1}{T})$ upper bounds. In many cases we give tight lower bounds.

Contents.

0022-0000/02 $35.00

0. INTRODUCTION

Consider a sequence of trials $t = 1, 2, \ldots, T$. In each trial an example is processed. Such an example consists of an *instance vector* x_t and an *outcome* y_t. For some trials t the *learner* has to make a *prediction* $\hat{y}_t$ for the outcome y_t. The learner must do this based on the examples from the previous $t-1$ trials plus the current instance x_t. This means that as the number of trials increases, the learner has more information at its disposal.

A *learning algorithm* is a strategy for choosing predictions. The discrepancy between a prediction $\hat{y}_t$ of the learner and a correct outcome y_t is measured by a loss function. The learner wants to make use of "correlations" between instances and outcomes for the purpose of keeping the loss as small as possible. To model such correlations we compare the loss of a learning algorithm against the loss of the best function from a *comparison class* of predictors.

In this paper we are mainly interested in *off-line learning* where the learner has to make one prediction in the last trial T. *On-line* algorithms must predict in all trials. For off-line algorithms we focus on the *instantaneous* loss in the last trial and for on-line algorithms on the *total* loss of all trials. We always consider the loss of the algorithm minus the loss of the best comparator. Such bounds are called relative loss bounds. They might be shown for worst-case sequences of examples or for the case when the examples are i.i.d. with a fixed but unknown distribution. The main focus of this paper is to prove relative expected instantaneous loss bounds.

Our work builds on the recent successes in proving relative total loss bounds for on-line algorithms. These bounds hold for worst-case sequences and they grow as $O(\ln T)$ (see Foster [7], Vovk [22], Azoury and Warmuth [3], Forster [5], Gordon [9], Yamanishi [24, 25]). There are standard conversions of on-line algorithms to off-line algorithms (see Helmbold and Warmuth [13], Kivinen and Warmuth [14]). These conversions would produce complicated algorithms and their relative expected instantaneous loss bounds would have the form $O(\frac{\ln T}{T})$. Instead we prove bounds of the form $O(\frac{1}{T})$.

We first do this by an exact calculation for the case when the comparators are Gaussian densities (with a fixed variance). In the case when the comparators are Bernoulli distributions and for linear regression we use a generalization of an inequality from Haussler *et al.* [11] that is based on the leave-one-out loss. When the comparators are from a class of densities then we always use the negative log-likelihood as the loss function. So when the comparators are Gaussian densities then the square loss is used. Also for linear regression we naturally use the square loss.

Note that the comparison class does not restrict the distribution by which the examples are generated. In all our expected instantaneous loss bounds the latter distribution on the example domain is arbitrary.

We also give a lower bound that shows the expected instantaneous loss bound is tight for the case when the comparators are Gaussian densities. For linear regression our lower and upper bounds are within a factor of two. We believe that the upper bound for linear regression can be improved so that it meets the lower bound.

The result for the case when the comparators are Gaussian densities is known in the statistics community (see, e.g., Lehmann and Casella [17]). We keep this simple example for the purpose of showing connections to previous techniques. In particular, when the comparators are Gaussian densities then the relative instantaneous loss can be rewritten as a simple expected loss. We later generalize this property to the case when comparators are densities from any member of the exponential family. Such a reformulation of the expected instantaneous loss does not seem to be possible for linear regression. We also keep the Gaussian example, because the lower bound technique used for this case is needed for linear regression.

All algorithms we use in this paper are subtle modifications of previously introduced algorithms. We chose these modifications to optimize our bounds. It remains to be seen whether these modifications result in improved performance on natural data.

An interesting application of our new linear regression algorithm is the case when the instances are expanded to feature vectors and the dot product between two feature vectors is given by a kernel function (see Saunders *et al.* [20]). Also Fourier or wavelet transforms can be used to extract frequency-dependent information from the instances (see, e.g., Walker [23] and Graps [10]). These linear transforms can reduce the dimensionality of the comparison class which leads to smaller relative loss bounds. One of the most salient properties of our bound for linear regression is the fact that it is linear in the *expected* dimension of the instances (or feature vectors).

There is an alternate machinery for proving bounds of the following type: If the number of examples is larger than $m(\epsilon, \delta)$ then the probability that the expected instantaneous loss of the algorithm is by ϵ larger than the expected instantaneous loss of the best comparator is at most $1-\delta$. The accuracy parameters ϵ and δ are small positive numbers and the sample size $m(\epsilon, \delta)$ grows at least linearly in a dimension parameter such as the fat shattering dimension or the pseudo dimension (see, e.g., Anthony and Bartlett [1]). Ignoring the dependence on the dimension, the best bounds produced by this methodology are usually of the form $O(\frac{1}{\epsilon}(\log\frac{1}{\epsilon}+\log\frac{1}{\delta}))$. Converting these bounds to expected instantaneous loss bound produces bounds of the form $O(\frac{\log T}{T})$, whereas our bounds have the form $O(\frac{1}{T})$. What about conversions in the opposite direction? In the case of learning $\{0, 1\}$-valued functions with respect to the discrete loss the following conversion was shown in Haussler *et al.* [11, Sect. 5.]. Assume that the expected instantaneous loss of the best comparator is always zero (i.e., the noise-free case). Then any algorithm that has a relative expected instantaneous loss bound of $O(\frac{1}{T})$ can be converted using standard hypothesis testing to an algorithm with the following bound: After using $O(\frac{1}{\epsilon}\log\frac{1}{\delta})$ examples the expected instantaneous loss is at most ϵ with probability at least $1-\delta$. Similar conversions that hold for other loss functions and the noisy case need to be investigated further.

This paper is outlined as follows. In the next section we begin with some notation used throughout the paper. We then give a simple expected instantaneous loss bound for the case when the comparators are Gaussian densities (with a fixed variance) (Section 2). This is followed by a general theorem based on the leave-one-out loss (Section 3). Most upper bounds given in the rest of the paper use this theorem. We first apply it to the case when the comparators are Bernoulli distributions (Section 4) and then consider a general class of loss functions called Bregman divergences (Section 5). Our upper bound for linear regression is given in Section 6 and the lower bounds for the case when the comparators are Gaussian densities and for linear regression appear in Section 7. We end the paper with some concluding remarks (Section 8).

1. NOTATION AND PRELIMINARIES

In our setting, instances x, predictions $\hat{y}$, and outcomes y are elements of an *instance space* $\mathscr{X}$, a *prediction space* $\hat{\mathscr{Y}}$, and an *outcome space* $\mathscr{Y}$, respectively. An *example* z is a pair (x, y) of an instance $x \in \mathscr{X}$ and an outcome $y \in \mathscr{Y}$. Loss functions map tuples of predictions and outcomes to the non-negative reals. For such a loss function L, the learner incurs loss $L(y, \hat{y})$ if it makes the prediction $\hat{y} \in \hat{\mathscr{Y}}$ and the correct outcome is $y \in \mathscr{Y}$. The comparison class $\mathscr{C}$ consists of functions c that map instances to predictions, i.e., $c\colon \mathscr{X} \to \hat{\mathscr{Y}}$. Such a function is called a *comparator*. The loss of a comparator c on example (x, y) is $L(y, c(x))$.

A learning algorithm Q is a function that maps a batch of examples and an instance to a prediction. If the learner is given the examples $z_1 = (x_1, y_1), \ldots, z_{T-1} = (x_{T-1}, y_{T-1}) \in \mathscr{X} \times \mathscr{Y}$ in trials 1 through $T-1$ and the instance $x_T \in \mathscr{X}$ in trial T, then its prediction is $Q(z_1, \ldots, z_{T-1}, x_T) \in \hat{\mathscr{Y}}$. The loss of algorithm Q in trial T is

$$L(y_T, Q(z_1, \ldots, z_{T-1}, x_T)).$$

Under the assumption that the examples are i.i.d. with unknown distribution we want to find learning algorithms for which we can prove that the expectation of the loss of the learner in trial T on a random example is not much larger than the expected loss of the best function from the comparison class. Formally we want to bound the *relative expected instantaneous loss*

$$\mathrm{E}_{(z_1, \ldots, z_T) \sim \mathscr{D}^T}(L(y_T, Q(z_1, \ldots, z_{T-1}, x_T))) - \inf_{c \in \mathscr{C}} \mathrm{E}_{(x, y) \sim \mathscr{D}}(L(y, c(x))) \qquad (1)$$

for any distribution $\mathscr{D}$ on the set of examples $\mathscr{X} \times \mathscr{Y}$. Here $\mathrm{E}_{(z_1, \ldots, z_T) \sim \mathscr{D}^T}$ denotes the expectation over the random variables $z_1 = (x_1, y_1), \ldots, z_T = (x_T, y_T)$ which are i.i.d. with distribution $\mathscr{D}$, and $\mathrm{E}_{(x, y) \sim \mathscr{D}}$ denotes the expectation over the random variable (x, y) with distribution $\mathscr{D}$. The infimum in (1) is called the *comparison term*.

In particular we are interested in density estimation problems and in regression problems. For density estimation problems the instance space $\mathscr{X}$ is not used.

For $n \in \mathbb{N}$, $\mathbb{R}^n$ denotes the set of n-dimensional real vectors. For $m, n \in \mathbb{N}$, $\mathbb{R}^{m \times n}$ is the set of real matrices with m rows and n columns. Vectors $x \in \mathbb{R}^n$ are column vectors and x' denotes the transpose of x. The scalar product of two vectors $w, x \in \mathbb{R}^n$ is $w \cdot x = w'x$, and the Euclidean norm of a vector $x \in \mathbb{R}^n$ is $\|x\| = (x \cdot x)^{1/2}$.

A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called positive semi-definite if $x'Ax \geqslant 0$ for all vectors $x \in \mathbb{R}^n$. It is called positive definite if $x'Ax > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$. Every positive definite matrix is invertible. For example, the unit matrix $I \in \mathbb{R}^{n\times n}$ is positive definite, and for every vector $x \in \mathbb{R}^n$ the matrix $xx' \in \mathbb{R}^{n\times n}$ is positive semidefinite.

2. GAUSSIAN COMPARATORS

One of the simplest models of the examples is a Gaussian density (with a fixed variance). In this case the prediction space $\hat{\mathscr{Y}}$ and the outcome space $\mathscr{Y}$ are both $\mathbb{R}^m$ for a fixed dimension $m \in \mathbb{N}$. Also the comparison class $\mathscr{C}$ of all possible choices for the mean is again $\mathbb{R}^m$. The form of the negative log-likelihood of a Gaussian density motivates the use the following loss function for this case: The loss function is the squared Euclidean norm $L(y, \hat{y}) = \|\hat{y} - y\|^2$. Thus the relative expected instantaneous loss (1) is

$$\mathrm{E}_{(y_1, ..., y_T) \sim \mathscr{D}^T}(\|Q(y_1, ..., y_{T-1}) - y_T\|^2) - \inf_{c \in \mathbb{R}^m} \mathrm{E}_{y \sim \mathscr{D}}(\|c - y\|^2). \tag{2}$$

For the above choice of loss function the comparison term in the relative expected instantaneous loss is the variance of the unknown distribution $\mathscr{D}$ of the outcomes, i.e.,

$$\inf_{c \in \mathbb{R}^m} \mathrm{E}_{y \sim \mathscr{D}}(\|c - y\|^2) = \mathrm{E}_{y \sim \mathscr{D}}(\|y\|^2) - \|\mathrm{E}_{y \sim \mathscr{D}}(y)\|^2, \tag{3}$$

and the infimum is attained when c is chosen as the expectation $\mathrm{E}_{y \sim \mathscr{D}}(y)$ of $\mathscr{D}$. To see this note that

$$\mathrm{E}_{y \sim \mathscr{D}}(\|c - y\|^2) = \|c\|^2 - 2c \cdot \mathrm{E}_{y \sim \mathscr{D}}(y) + \mathrm{E}_{y \sim \mathscr{D}}(\|y\|^2) \tag{4}$$

is convex in c. Setting the gradient of (4) with respect to c to zero shows that the infimum of (4) is attained for $c = \mathrm{E}_{y \sim \mathscr{D}}(y)$. For this c, the right hand side of (4) is the variance of y.

When the comparators are Gaussian densities then the relative expected instantaneous loss can also be written as

$$\mathrm{E}_{(y_1, ..., y_{T-1}) \sim \mathscr{D}^{T-1}}(\|Q(y_1, ..., y_{T-1}) - \mathrm{E}_{y \sim \mathscr{D}}(y)\|^2). \tag{5}$$

In Section 5.2 we given a generalization of this reformulization of the expected instantaneous loss to other comparison classes. For the case when the comparators are Gaussian densities, the reformulization follows from (2), (3), and

$$\begin{aligned}
&\mathrm{E}_{(y_1, ..., y_T) \sim \mathscr{D}^T}(\|Q(y_1, ..., y_{T-1}) - y_T\|^2) - \mathrm{E}_{y_T \sim \mathscr{D}}(\|y_T\|^2) + \|\mathrm{E}_{y_T \sim \mathscr{D}}(y_T)\|^2 \\
&\quad = \mathrm{E}_{(y_1, ..., y_{T-1}) \sim \mathscr{D}^{T-1}}(\|Q(y_1, ..., y_{T-1})\|^2) \\
&\qquad - 2\mathrm{E}_{(y_1, ..., y_{t-1}) \sim \mathscr{D}^{T-1}}(Q(y_1, ..., y_{T-1})) \cdot \mathrm{E}_{y_T \sim \mathscr{D}}(y_T) \\
&\qquad + \|\mathrm{E}_{y_T \sim \mathscr{D}}(y_T)\|^2 \\
&\quad = \mathrm{E}_{(y_1, ..., y_{T-1}) \sim \mathscr{D}^{T-1}}(\|Q(y_1, ..., y_{T-1}) - \mathrm{E}_{y_T \sim \mathscr{D}}(y_T)\|^2).
\end{aligned}$$

For the first equality we used that $Q(y_1, ..., y_{T-1})$ and y_T are independent random variables.

The above reformulation (5) of the expected instantaneous loss shows that minimizing this loss is the same as minimizing the expected square loss of the estimator Q for the mean. The following bound for a particular estimator Q and the corresponding lower bound are known in the statistics community (Lehmann and Casella [17, Chap. 5.1, Example 1.16]) in the context of minimax estimation of the mean of a Gaussian density.

THEOREM 2.1. *Consider the following prediction algorithm Q for the case when the comparators are Gaussian densities*

$$Q(y_1, ..., y_{T-1}) := \frac{\sum_{t=1}^{T-1} y_t}{T-1+\sqrt{T-1}}.$$

Then for any distribution $\mathscr{D}$ on $\mathbb{R}^m$ the relative expected instantaneous loss of Q is

$$\mathrm{E}_{(y_1, ..., y_T) \sim \mathscr{D}^T}(\|Q(y_1, ..., y_{T-1}) - y_T\|^2) - \inf_{c \in \mathbb{R}^m} \mathrm{E}_{y \sim \mathscr{D}}(\|c - y\|^2)$$
$$= \frac{1}{T+2\sqrt{T-1}} \mathrm{E}_{y \sim \mathscr{D}}(\|y\|^2).$$

Proof. Let $\eta := (T-1+\sqrt{T-1})^{-1}$. The relative expected instantaneous loss (5) is

$$\mathrm{E}_{(y_1, ..., y_{T-1}) \sim \mathscr{D}^{T-1}} \left(\left\| \eta \sum_{t=1}^{T-1} y_t - \mathrm{E}_{y \sim \mathscr{D}}(y) \right\|^2 \right)$$
$$= \mathrm{E}_{(y_1, ..., y_{T-1}) \sim \mathscr{D}^{T-1}} \left(\eta^2 \left\| \sum_{t=1}^{T-1} y_t \right\|^2 - 2\eta \sum_{t=1}^{T-1} y_t \cdot \mathrm{E}_{y \sim \mathscr{D}}(y) + \|\mathrm{E}_{y \sim \mathscr{D}}(y)\|^2 \right)$$
$$= \eta^2 (T-1)\, \mathrm{E}_{y \sim \mathscr{D}}(\|y\|^2) + \|\mathrm{E}_{y \sim \mathscr{D}}(y)\|^2 \underbrace{(\eta^2 (T-1)(T-2) - 2\eta(T-1) + 1)}_{=0}$$
$$= \frac{1}{T+2\sqrt{T-1}} \mathrm{E}_{y \sim \mathscr{D}}(\|y\|^2).$$

For the second equality we used that

$$\left\| \sum_{t=1}^{T-1} y_t \right\|^2 = \sum_{t=1}^{T-1} \|y_t\|^2 + \sum_{\substack{s, t=1 \\ s \neq t}}^{T-1} y_s \cdot y_t. \quad \blacksquare$$

The prediction used in Theorem 2.1 is a special case of the following prediction for trial T:

$$\frac{c y_0 + \sum_{t=1}^{T-1} y_t}{c + T - 1}.$$

Here y_0 is an initial mean and $c \geqslant 0$ is the multiplicity of this mean. We chose $y_0 = 0$ and $c = \sqrt{T-1}$. Azoury and Warmuth [3] proved worst-case on-line total loss bounds for the algorithm that uses $y_0 = 0$ and $c = 1$. The relative expected instantaneous loss bounds for the latter algorithm are slightly weaker.

It is very natural to apply Theorem 2.1 if we know that the norms of the outcomes $y_1, ..., y_T$ are bounded by some constant Y. Then Theorem 2.1 shows that the relative expected instantaneous loss is at most $Y^2/(T+2\sqrt{T-1})$. In Section 7 we show that this bound for the case when the comparators are Gaussian densities is optimal in a very strong sense: For every learning algorithm Q' there is a simple distribution $\mathscr{D}$ on two points for which the relative expected instantaneous loss of Q' is at least as large as our upper bound.

If there is no restriction on the outcomes, then a straightforward modification of the proof of Theorem 2.1 shows that if we predict with the mean of the past outcomes, i.e., $Q(y_1, ..., y_{T-1}) = (y_1 + \cdots + y_{T-1})/(T-1)$, then the relative expected instantaneous loss is the variance (3) divided by $T-1$.

3. THE LEAVE-ONE-OUT LOSS

The bound proven in the previous section for the case when the comparators are Gaussian densities was obtained by an exact calculation. Some more examples of density estimation problems for which an exact calculation is possible are given in Lehmann and Casella [17]. However these examples are essentially based on the square loss. In this paper we give relative expected instantaneous loss bounds for other loss functions and for linear regression (i.e., square loss in a more involved setting). Our central technique for proving such bounds is a general inequality given in Theorem 3.1 of this section. This theorem is a generalization of a similar theorem given in Haussler *et al.* [11]. Theorem 3.1 gives a bound on the relative expected instantaneous loss (1) in terms of the *leave-one-out loss* of a learning algorithm. The leave-one-out loss of a learning algorithm Q on a sequence of T examples $z_1, ..., z_T \in \mathscr{X} \times \mathscr{Y}$ is the average of the losses of the learning algorithm on the last example of certain permutations of the sequence:

$$L_{Q,\,\mathrm{loo}}(z_1, ..., z_T) := \frac{1}{T} \sum_{t=1}^{T} L(y_t, Q(z_1, ..., z_{t-1}, z_{t+1}, ..., z_T, x_t)). \tag{6}$$

The *sample error* of a function c: $\mathscr{X} \to \hat{Y}$ on T examples $z_1, ..., z_T \in \mathscr{X} \times \mathscr{Y}$ is

$$L_{c,\,\mathrm{se}}(z_1, ..., z_T) := \frac{1}{T} \sum_{t=1}^{T} L(y_t, c(x_t)). \tag{7}$$

THEOREM 3.1. *The relative expected instantaneous loss* (1) *of any learning algorithm Q and any distribution $\mathscr{D}$ on the example space $\mathscr{X} \times \mathscr{Y}$ is bounded as*

$$\begin{aligned} &\mathrm{E}_{(z_1, ..., z_T) \sim \mathscr{D}^T}(L(y_T, Q(z_1, ..., z_{T-1}, x_T))) - \inf_{c \in \mathscr{C}} \mathrm{E}_{(x,y) \sim \mathscr{D}}(L(y, c(x))) \\ &\quad \leqslant \mathrm{E}_{(z_1, ..., z_T) \sim \mathscr{D}^T}(L_{Q,\,\mathrm{loo}}(z_1, ..., z_T) - \inf_{c \in \mathscr{C}} L_{c,\,\mathrm{se}}(z_1, ..., z_T)) \\ &\quad \leqslant \sup_{z_1, ..., z_T \in \mathscr{X} \times \mathscr{Y}} (L_{Q,\,\mathrm{loo}}(z_1, ..., z_T) - \inf_{c \in \mathscr{C}} L_{c,\,\mathrm{se}}(z_1, ..., z_T)). \end{aligned}$$

Proof. Expectations do not change if we permute the i.i.d. examples $z_1, \ldots, z_T$. That is, for any permutation σ of $(1, 2, \ldots, t)$,

$$\begin{aligned} &\mathrm{E}_{(z_1, \ldots, z_T) \sim \mathscr{D}^T}(L(y_T, Q(z_1, \ldots, z_{T-1}, x_T))) \\ &\qquad = \mathrm{E}_{(z_1, \ldots, z_T) \sim \mathscr{D}^T}(L(y_{\sigma(T)}, Q(z_{\sigma(1)}, \ldots, z_{\sigma(t-1)}, x_{\sigma(T)})). \end{aligned}$$

We use this to rewrite the first term of (1):

$$\begin{aligned} &\mathrm{E}_{(z_1, \ldots, z_T) \sim \mathscr{D}^T}(L(y_T, Q(z_1, \ldots, z_{T-1}, x_T))) \\ &\quad = \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}_{(z_1, \ldots, z_T) \sim \mathscr{D}^T}(L(y_t, Q(z_1, \ldots, z_{t-1}, z_{t+1}, \ldots, z_T, x_t))) \\ &\quad = \mathrm{E}_{(z_1, \ldots, z_T) \sim \mathscr{D}^T}(L_{Q, \mathrm{loo}}(z_1, \ldots, z_T)). \end{aligned}$$

The second term of (1) can be bounded as

$$\begin{aligned} -\inf_{c \in \mathscr{C}} \mathrm{E}_{(x, y) \sim \mathscr{D}}(L(y, c(x))) &\overset{(7)}{=} -\inf_{c \in \mathscr{C}} \mathrm{E}_{(z_1, \ldots, z_T) \sim \mathscr{D}^T}(L_{c, \mathrm{se}}(z_1, \ldots, z_T)) \\ &\leqslant -\mathrm{E}_{(z_1, \ldots, z_T) \sim \mathscr{D}^T}(\inf_{c \in \mathscr{C}} L_{c, \mathrm{se}}(z_1, \ldots, z_T)). \end{aligned}$$

Plugging the above into (1) gives the theorem. ▮

The bound in terms of the supremum has the advantage that it does not contain an expectation over an unknown distribution. In the original theorem of Haussler *et al.* [11] the comparison term is zero.

For the case when the comparators are Gaussian densities, the infimum of the sample error that appears in the bounds of Theorem 3.1 is the sample variance:

$$\inf_{c \in \mathbb{R}^m} L_{c, \mathrm{se}}(y_1, \ldots, y_T) = \frac{1}{T} \sum_{t=1}^{T} \left\| y_t - \frac{1}{T} \sum_{s=1}^{T} y_s \right\|^2.$$

This follows because

$$L_{c, \mathrm{se}}(y_1, \ldots, y_T) \overset{(7)}{=} \frac{1}{T} \sum_{t=1}^{T} \|c - y_t\|^2 = \|c\|^2 - 2c \cdot \frac{1}{T} \sum_{t=1}^{T} y_t + \frac{1}{t} \sum_{t=1}^{T} \|y_t\|^2$$

is convex in c and its gradient with respect to c vanishes for the sample mean $c = \frac{1}{T} \sum_{t=1}^{T} y_T$. Plugging this value of c into the above expression gives the sample variance.

4. BERNOULLI COMPARATORS

For the case when the comparators are Bernoulli, the outcomes are coin flips (i.e., $\mathscr{Y} = \{0, 1\}$). The probability of the underlying coin is hidden and the comparison class consists of all possible choices for the hidden coin (i.e., $\mathscr{C} = [0, 1]$). The predictions are estimates of the probability of the hidden coin (i.e., $\hat{\mathscr{Y}} = [0, 1]$). If the learner makes the prediction $\hat{y}$ this means that it assigns probability $\hat{y}$ to the outcome 1 and probability $1-\hat{y}$ to the outcome 0. The loss function

$$L(y, \hat{y}) = -y \ln \hat{y} - (1-y) \ln(1-\hat{y}) \tag{8}$$

is the negated logarithm of the probability that the learner assigned to the correct binary outcome y. We use the notation $0 \cdot \ln \xi := 0$ for all $\xi \in \mathbb{R}$, i.e., $L(0, 0) = L(1, 1) = 0$ and $L(1, 0) = L(0, 1) = \infty$. Note that as in the case when Gaussian densities were the comparators we do not use the instance space $\mathscr{X}$.

Shtarkov [21] showed that the best total relative loss for T trials that can be achieved for the case of Bernoulli comparators by any on-line algorithm is $\frac{1}{2} \ln(T+1) + \frac{1}{2} \ln \frac{\pi}{2}$. Standard conversions would lead to an $O(\frac{\ln T}{T})$ bound on the expected instantaneous loss for the case when the comparators are Bernoulli. In the following theorem we show how the leave-one-out bound from Theorem 3.1 can be used to bound the relative expected instantaneous loss by $\frac{1}{T}$ for case when the comparators are Bernoulli.

THEOREM 4.1. *Consider the following prediction algorithm Q for the case when the comparators are Bernoulli*

$$Q(y_1, \ldots, y_{T-1}) := \frac{1 + \sum_{t=1}^{T-1} y_t}{T+1}.$$

Then for any distribution $\mathscr{D}$ on the outcome space the relative expected instantaneous loss of Q is bounded as

$$\begin{aligned} &\mathrm{E}_{(y_1, \ldots, y_T) \sim \mathscr{D}^T}(L(y_T, Q(y_1, \ldots, y_{T-1}))) - \inf_{c \in \mathscr{C}} \mathrm{E}_{y \sim \mathscr{D}}(L(y, c)) \\ &\quad \leqslant \ln\left(1 + \frac{1}{T}\right) \leqslant \frac{1}{T}. \end{aligned}$$

Proof. Because of Theorem 3.1 it suffices to show that for any T outcomes $y_1, \ldots, y_T$ the term

$$L_{Q,\mathrm{loo}}(y_1, \ldots, y_T) - \inf_{c \in [0,1]} L_{c,\mathrm{se}}(y_1, \ldots, y_T)$$

is equal to $\ln(1 + \frac{1}{T})$. The infimum can be written as

$$\inf_{c \in [0,1]} L_{c,\mathrm{se}}(y_1, \ldots, y_T) = -\bar{y} \ln \bar{y} - (1-\bar{y}) \ln(1-\bar{y}),$$

where $\bar{y} := \frac{1}{T} \sum_{t=1}^{T} y_t$ is the average of the examples. This holds because

$$\begin{aligned} L_{c,\,\mathrm{se}}(y_1, ..., y_T) &\overset{(7)}{=} \frac{1}{T} \sum_{t=1}^{T} L(y_t, c) \\ &\overset{(8)}{=} \frac{1}{T}\left(-\sum_{t=1}^{T} y_t \ln c - \sum_{t=1}^{T} (1-y_t) \ln(1-c) \right) \\ &= -\bar{y} \ln c - (1-\bar{y}) \ln(1-c) \end{aligned}$$

is convex in $c \in [0, 1]$ and its gradient with respect to c vanishes for $c = \bar{y}$.

The leave-one-out loss of the learning algorithm Q on the outcomes $y_1, ..., y_T$ is

$$\begin{aligned} L_{Q,\,\mathrm{loo}}(y_1, ..., y_T) &\overset{(6)}{=} \frac{1}{T} \sum_{t=1}^{T} L\left(y_t, \frac{1 + \sum_{s \in \{1, ..., T\} \setminus \{t\}} y_s}{T+1} \right) \\ &= \frac{1}{T} \sum_{\substack{t \in \{1, ..., T\} \\ y_t = 1}} L\left(1, \frac{T\bar{y}}{T+1}\right) + \frac{1}{T} \sum_{\substack{t \in \{1, ..., T\} \\ y_t = 0}} L\left(0, \frac{1+T\bar{y}}{T+1}\right) \\ &\overset{(8)}{=} -\bar{y} \ln\left(\frac{T\bar{y}}{T+1}\right) - (1-\bar{y}) \ln\left(\frac{T(1-\bar{y})}{T+1}\right). \end{aligned}$$

Thus

$$\begin{aligned} &L_{Q,\,\mathrm{loo}}(y_1, ..., y_T) - \inf_{c \in [0,1]} L_{c,\,\mathrm{se}}(y_1, ..., y_T) \\ &= -\bar{y} \ln\left(\frac{T}{T+1}\right) - (1-\bar{y}) \ln\left(\frac{T}{T+1}\right) \\ &= \ln\left(1 + \frac{1}{T}\right) \leqslant \frac{1}{T}. \quad \blacksquare \end{aligned}$$

As in the case when the comparators are Gaussian densities, our algorithm uses a prediction of the form

$$\frac{c y_0 + \sum_{t=1}^{T-1} y_t}{c + T - 1}.$$

The initial mean is chosen to be unbiased, i.e., $y_0 = \frac{1}{2}$. In our prediction we chose $c = 2$ for the multiplicity of the mean. This is different from the standard Krichevsky–Trofimov estimator for which $c = 1$. For the Krichevsky–Trofimov estimator we could only prove a slightly weaker bound. However, for the latter estimator one can prove a total relative loss bound of $\frac{1}{2} \ln(T+1) + \frac{1}{2} \ln \pi$ [3, 8, 26] and this bound differs only by a constant from the best obtainable bound of $\frac{1}{2} \ln(T+1) + \frac{1}{2} \ln \frac{\pi}{2}$ proven by Shtarkov [21].

5. BREGMAN DIVERGENCES AS LOSS FUNCTIONS

In this section we prove relative expected instantaneous loss bounds for a very general class of loss functions called Bregman divergences. Since the negative log-likelihood of any member of the exponential family of distributions is essentially a Bregman divergence, this covers a wide range of loss functions. In previous sections we proved bounds for the case when the comparators were Gaussian or Bernoulli distributions which are both members of the exponential family. (Note that we always use the negative log-likelihood as the associated loss function whenever the comparison class is a class of densities.) In this section we prove a general bound for arbitrary Bregman divergences as loss functions and discuss what happens when the general bound is applied to the Gaussian and Bernoulli case. This section also makes a connection to the work of Azoury and Warmuth [3] where total loss bounds are proven for the case when the comparison class is a class of densities from the exponential family and more background is given on the subject of Bregman divergences and the exponential family.

Bregman divergences were introduced by Bregman [4]. The Bregman divergence between two points is a non-negative real number which can be interpreted as a measure of distance. However, a Bregman divergence is usually not symmetric and the triangular inequality may not hold. Let F be any differentiable convex function. For any μ, $\tilde{\mu}$ from the domain $\mathrm{dom}(F) \subseteq \mathbb{R}^m$ of F, the Bregman divergence Δ_F is the difference between $F(\mu)$ and its first-order Taylor approximation at $\tilde{\mu}$:

$$\Delta_F(\mu, \tilde{\mu}) = F(\mu) - F(\tilde{\mu}) - (\mu - \tilde{\mu}) \cdot \nabla F(\tilde{\mu}).$$

Since F is convex, $\Delta_F(\mu, \tilde{\mu}) \geqslant 0$. If F is strictly convex then equality holds iff $\mu = \tilde{\mu}$. See Azoury and Warmuth [3] for a list of additional properties. We denote the gradient of F by $f = \nabla F$ and the Hessian of F by $\nabla^2 F$. Note that by Taylor's Theorem the Bregman divergence $\Delta_F(\mu, \tilde{\mu})$ is bounded by

$$\tfrac{1}{2} \|\mu - \tilde{\mu}\|^2 \sup_{t \in [0,1]} \|\nabla^2 F((1-t)\, \mu + t\tilde{\mu})\|.$$

Now the prediction space $\hat{\mathscr{Y}}$ and the comparison class $\mathscr{C}$ are both equal $\mathrm{dom}(F)$. The outcomes y are assumed to come from a convex subset $\mathscr{Y}$ of $\mathrm{dom}(F)$. For a prediction $\hat{y} \in \mathrm{dom}(F)$ and an outcome $y \in \mathscr{Y}$ we define the loss to be the Bregman divergence

$$L(y, \hat{y}) := \Delta_F(y, \hat{y}) = F(y) - F(\hat{y}) - (y - \hat{y}) \cdot f(\hat{y}).$$

The function F is understood from the context. For example, if $F(\mu) = \|\mu\|^2$ then $L(y, \hat{y}) = \|y - \hat{y}\|^2$.

5.1. Relationship Between Bregman Divergences and Negative log-Likelihood of Any Member of the Exponential Family of Distributions

A class p_θ, $\theta \in \Theta \subseteq \mathbb{R}^m$, of distributions is said to be an *exponential family*, if each p_θ has density function

$$\mathbb{R}^m \to \mathbb{R}, \qquad y \mapsto \exp(\theta \cdot y - G(\theta)) \tag{9}$$

with respect to some common measure on $\mathbb{R}^m$. The parameter θ is called the *natural parameter* and the function G is called the *cumulant function*. Θ is called the *natural parameter space*. The cumulant G must have a number of properties. In particular, G is strictly convex and must have a positive definite Hessian. We denote the gradient of G by g. Since G is a strictly convex function, g is an invertible function. We write $f := g^{-1}$.

For density estimation it is common to use the negative log-likelihood as a loss function. For exponential families, the negative log-likelihood of (9) is

$$G(\theta) - y \cdot \theta \tag{10}$$

for prediction θ and outcome y.

Sometimes it is more convenient to use the *expectation parameter*

$$\mu := \int y \, \mathrm{d}p_\theta(y)$$

instead of the natural parameter θ for exponential families. This expectation μ is equal to $g(\theta) = \nabla G(\theta)$. The function $F(\mu) := \mu \cdot f(\mu) - G(f(\mu))$, $\mu \in g(\Theta)$, is called the *dual* of G. F is also a strictly convex function and it is easy to check that $f = g^{-1}$ is the gradient of F.

Using the dual convex function F of G we can rewrite the negative log-likelihood (10) as

$$-F(\mu) - (y - \mu) \cdot f(\mu).$$

Except for the term $F(y)$ (that does not affect relative losses), this is the Bregman divergence $\Delta_F(y, \mu)$. So Bregman divergences are a generalization of the negative log-likelihood of any member of the exponential family.

5.2. A General Relative Expected Instantaneous Loss Bound for Bregman Divergences

As in the case of Gaussian comparators, there is another way of writing the relative expected instantaneous loss (1) if the loss function is a Bregman divergence:[1]

[1] For the case of Bernoulli density estimation this was pointed out to us by Hans Ulrich Simon.

THEOREM 5.1. *Assume that F is a convex and twice differentiable function. The prediction space $\hat{\mathcal{Y}}$ and the comparison class $\mathcal{C}$ are the domain of F; the outcomes y are assumed to come from a convex subset $\mathcal{Y}$ of the domain of F. For a prediction $\hat{y}$ and an outcome $y \in \mathcal{Y}$ the loss $L(y, \hat{y})$ is the Bregman divergence $\Delta_F(y, \hat{y})$. Then the relative expected instantaneous loss* (1) *of any learning algorithm Q can be written as*

$$\begin{aligned} &\mathrm{E}_{(y_1, \ldots, y_T) \sim \mathcal{D}^T}(L(y_T, Q(y_1, \ldots, y_{T-1}))) - \inf_{c \in \mathcal{C}} \mathrm{E}_{y \sim \mathcal{D}}(L(y, c)) \\ &\quad = \mathrm{E}_{(y_1, \ldots, y_{T-1}) \sim \mathcal{D}^{T-1}}(L(\mathrm{E}_{y \sim \mathcal{D}}(y), Q(y_1, \ldots, y_{T-1}))). \end{aligned}$$

Proof. Let $\mu := \mathrm{E}_{y \sim \mathcal{D}}(y)$. First note that the term

$$\mathrm{E}_{y \sim \mathcal{D}}(L(y, c)) = \mathrm{E}_{y \sim \mathcal{D}}(F(y)) + \Delta_F(\mu, c) - F(\mu)$$

in (1) is minimal if $c = \mu$. We can now rewrite the relative expected instantaneous loss (1) as follows (for brevity we write $\hat{y}_T := Q(y_1, \ldots, y_{T-1})$):

$$\begin{aligned} &\mathrm{E}_{(y_1, \ldots, y_T) \sim \mathcal{D}^T}(L(y_T, Q(y_1, \ldots, y_{T-1}))) - \inf_{c \in \mathcal{C}} \mathrm{E}_{y \sim \mathcal{D}}(L(y, c)) \\ &\quad = \mathrm{E}_{(y_1, \ldots, y_T) \sim \mathcal{D}^T}(\Delta_F(y_T, \hat{y}_T)) - \mathrm{E}_{y \sim \mathcal{D}}(\Delta_F(y, \mu)) \\ &\quad = \mathrm{E}_{(y_1, \ldots, y_T) \sim \mathcal{D}^T}(\Delta_F(y_T, \hat{y}_T) - \Delta_F(y_T, \mu)) \\ &\quad = \mathrm{E}_{(y_1, \ldots, y_T) \sim \mathcal{D}^T}(-F(\hat{y}_T) - (y_T - \hat{y}_T) \cdot f(\hat{y}_T) + F(\mu) + (y_T - \mu) \cdot f(\mu)) \\ &\quad = \mathrm{E}_{(y_1, \ldots, y_{T-1}) \sim \mathcal{D}^{T-1}}(F(\mu) - F(\hat{y}_T) - (\mu - \hat{y}_T) \cdot f(\hat{y}_T)) \\ &\quad = \mathrm{E}_{(y_1, \ldots, y_{T-1}) \sim \mathcal{D}^{T-1}}(\Delta_F(\mu, \hat{y}_T)) \\ &\quad = \mathrm{E}_{(y_1, \ldots, y_{T-1}) \sim \mathcal{D}^{T-1}}(L(\mathrm{E}_{y \sim \mathcal{D}}(y), Q(y_1, \ldots, y_{T-1}))). \end{aligned}$$

For the fourth equality we used that $\hat{y}_T$ and y_T are independent. ▌

From Theorem 5.1 we can easily get a bound on the relative instantaneous loss that is in terms of the variance of the outcomes and of the curvature of F:

COROLLARY 5.1. *Assume that the loss function is a Bregman divergence Δ_F and that the prediction space $\hat{\mathcal{Y}}$, the comparison class $\mathcal{C}$, and the outcome space $\mathcal{Y}$ are as in Theorem* 5.1. *Consider the following prediction algorithm Q:*

$$Q(y_1, \ldots, y_{T-1}) := \frac{1}{T-1} \sum_{t=1}^{T-1} y_t.$$

Then for any distribution $\mathcal{D}$ on the outcome space the relative expected instantaneous loss of Q is bounded as

$$\begin{aligned} &\mathrm{E}_{(y_1, \ldots, y_T) \sim \mathcal{D}^T}(L(y_T, Q(y_1, \ldots, y_{T-1}))) - \inf_{c \in \mathcal{C}} \mathrm{E}_{y \sim \mathcal{D}}(L(y, c)) \\ &\quad \leqslant \frac{1}{2(T-1)} \mathrm{E}_{y \sim \mathcal{D}}(\|\mu - y\|^2) \cdot \sup_{y \in \mathcal{Y}} \|\nabla^2 F(y)\|, \end{aligned}$$

where we set $\mu := \mathrm{E}_{y \sim \mathcal{D}}(y)$.

	Gaussian	Bernoulli
$F(\mu)$	$\lVert\mu\rVert^2$	$\mu\ln\mu+(1-\mu)\ln(1-\mu)$
$f(\mu)$	2μ	$\ln\frac{\mu}{1-\mu}$
$\Delta_F(y,\hat{y})$	$\lVert y-\hat{y}\rVert^2$	$y\ln\frac{y}{\hat{y}}+(1-y)\ln\frac{1-y}{1-\hat{y}}$

FIG. 1. Gaussian and Bernoulli density estimation.

Proof. We apply Theorem 5.1 and observe that

$$\begin{aligned}
&\mathrm{E}_{(y_1,\dots,y_{T-1})\sim\mathscr{D}^{T-1}}(L(\mu, Q(y_1,\dots,y_{T-1})))\\
&\leqslant \mathrm{E}_{(y_1,\dots,y_{T-1})\sim\mathscr{D}^{T-1}}\left(\frac{1}{2}\left\|\mu-\frac{1}{T-1}\sum_{t=1}^{T-1}y_t\right\|^2\cdot\sup_{y\in\mathscr{Y}}\lVert\nabla^2F(y)\rVert\right)\\
&=\frac{1}{2(T-1)}\,\mathrm{E}_{(y_1,\dots,y_{T-1})\sim\mathscr{D}^{T-1}}\left(\left\|(T-1)\,\mu-\sum_{t=1}^{T-1}y_t\right\|^2\right)\cdot\sup_{y\in\mathscr{Y}}\lVert\nabla^2F(y)\rVert\\
&=\frac{1}{2(T-1)^2}\,\mathrm{E}_{(y_1,\dots,y_{T-1})\sim\mathscr{D}^{T-1}}\left(\sum_{t=1}^{T-1}\lVert\mu-y_t\rVert^2\right)\cdot\sup_{y\in\mathscr{Y}}\lVert\nabla^2F(y)\rVert. \quad\blacksquare
\end{aligned}$$

Note that Theorem 3.1 can be used to obtain a bound of the same form as the bound given in the above corollary. However, the bound obtained this way is weaker.

The cases of Gaussian and Bernoulli distributions as comparators are summarized in Fig. 1. If the norms of the outcomes are bounded by a constant Y then Theorem 5.1 gives a bound of $Y^2/(T-1)$ on the relative expected instantaneous loss for Gaussian comparators. This bound is not optimal: We have seen in Section 2 that a bound of $Y^2/(T+2\sqrt{T-1})$ is possible in this case.

For the case when the comparators are Bernoulli first observe that the loss in Fig. 1 and the loss in (8) differ by a constant and thus lead to the same relative losses. Binary outcomes lie at the boundary of the domain (0, 1) of $F(\mu)$. Thus we can only apply Theorem 5.1 if the outcomes y are bounded away from 0 and 1. Otherwise Theorem 5.1 does not give a meaningful bound, because the second derivative of F, $\nabla^2F(\mu)=f'(\mu)=\frac{1}{\mu(1-\mu)}$, is unbounded.

6. LINEAR REGRESSION

6.1. Known Results

In linear regression the instance space is $\mathscr{X}=\mathbb{R}^n$ for a fixed dimension n. The prediction space is $\hat{\mathscr{Y}}=\mathbb{R}$ and the outcome space is $\mathscr{Y}=[-Y,Y]\subseteq\mathbb{R}$. This means that we make the assumption that the outcomes are bounded by some constant Y. However, the learner does not need to know Y. The loss function L is the square loss, i.e., $L(y,\hat{y})=(\hat{y}-y)^2$. The comparison class $\mathscr{C}$ consists of the linear functions

$c\colon \mathbb{R}^n \to \mathbb{R}$. Thus for off-line linear regression the relative expected instantaneous loss (1) of a learning algorithm Q is

$$\mathrm{E}_{(z_1, ..., z_T) \sim \mathscr{D}^T}((Q(z_1, ..., z_{T-1}, x_T) - y_T)^2) - \inf_{w \in \mathbb{R}^n} \mathrm{E}_{(x, y) \sim \mathscr{D}}((w \cdot x - y)^2),$$

where $\mathscr{D}$ is an unknown distribution on the examples. For the comparison term we represent linear functions $c\colon \mathbb{R}^n \to \mathbb{R}$ as n-dimensional vectors w.

For on-line linear regression the regularized relative total loss of a learning algorithm Q is

$$\sum_{t=1}^{T} (Q(z_1, ..., z_{t-1}, x_t) - y_t)^2 - \inf_{w \in \mathbb{R}^n} \left(\sum_{t=1}^{T} (w \cdot x_t - y_t)^2 + a \, \|w\|^2 \right), \tag{11}$$

where $z_1 = (x_1, y_1), ..., z_T = (x_T, y_T)$ are the T examples of trials 1 through T. Here the regularization term $\|w\|^2$ is a measure of the complexity of the vector w and $a \geqslant 0$ is a constant trade-off parameter. The larger a, the smaller the expression (11). Bounds on (11) that hold for almost arbitrary sequences of examples have only been shown for the case $a > 0$. Note that we will not need a term like the $a \, \|w\|^2$ in (11) to show our relative expected instantaneous loss bounds for off-line linear regression.

Vovk [22] proposed a learning algorithm for on-line linear regression for which he showed that the relative loss (11) is at most $nY^2 \ln(1 + TX^2/an)$ for all example sequences of length T for which the Euclidean norm of the instances is bounded by a constant X. For the case $a = 0$, he also showed that for any learning algorithm there is a distribution $\mathscr{D}$ on the examples for which the expectation of (11),

$$\mathrm{E}_{(z_1, ..., z_T) \sim \mathscr{D}^T} \left(\sum_{t=1}^{T} (Q(z_1, ..., z_{t-1}, x_t) - y_t)^2 - \inf_{w \in \mathbb{R}^n} \sum_{t=1}^{T} (w \cdot x_t - y_t)^2 \right) \tag{12}$$

is at least $(n - o(1)) \, Y^2 \ln T$ as $T \to \infty$. Forster and Warmuth [6] give a corresponding upper bound of $nY^2(1 + \ln T)$ on (12) that holds for any distribution $\mathscr{D}$ on the examples.

Vovk's algorithm for on-line linear regression and the bound $nY^2(1 + \ln T)$ on (12) can be converted to a learning algorithm for off-line linear regression with a relative expected instantaneous loss bound of $nY^2 \, \frac{(1 + \ln T)}{T}$. A bound of $O(nY^2 \, \frac{\ln T}{T})$ on (12) also follows from the fact that the class of linear functions on $\mathbb{R}^n$ has pseudodimension n (see Lee *et al.* (15]).

In Theorem 6.2 we improve on this and propose a new learning algorithm for linear regression for which we can prove that its relative expected instantaneous loss is at most $2nY^2 \, \frac{1}{T}$. (Note that for any particular distribution $\mathscr{D}$ the dimension n in this upper bound can be replaced by the expected dimension of T random instances.)

Relative loss bounds for linear regression	Upper bound	Lower bound
Expected instantaneous	$2nY^2\frac{1}{T}$	$(n-o(1))Y^2\frac{1}{T}$
Worst case on-line total	$nY^2\ln\left(1+\frac{TX^2}{an}\right)$	
Expected case on-line total	$nY^2(1+\ln T)$	$(n-o(1))Y^2\ln T$

FIG. 2. Upper and lower relative loss bounds for linear regression.

In Section 7 we also show a lower bound on the relative expected instantaneous loss of any learning algorithm. This lower bound is $(n-o(1))\,Y^2\frac{1}{T}$ as $T\to\infty$. This shows that our upper bound cannot be improved by more than a factor of 2.

An overview of upper and lower relative loss bounds for linear regression is given in Fig. 2.

6.2. New Algorithm and Proofs

We need some notation: For any T examples $z_1=(x_1, y_1), ..., z_T=(x_T, y_T)\in \mathcal{X}\times\mathcal{Y}=\mathbb{R}^n\times\mathbb{R}$ let

$$b_T := \sum_{t=1}^{T} y_t x_t \in \mathbb{R}^n, \qquad A_T := \sum_{t=1}^{T} x_t x_t' \in \mathbb{R}^{n\times n}. \tag{13}$$

A_T is a positive semi-definite matrix that might not be invertible, but the pseudoinverse $A_T^+\in\mathbb{R}^{n\times n}$ of A_T is always defined. For the definition of the pseudoinverse of a matrix see, e.g., Rektorys [19]. There it is also shown how the pseudoinverse of a matrix can be computed from the singular value decomposition. In Appendix A we show the following facts: The pseudoinverse A_T^+ is positive semi-definite and $A_T^+A_T x = x = A_T A_T^+ x$ holds for all $x\in\operatorname{span}\{x_1, ..., x_T\}$. Furthermore, for all $t\in\{1, ..., T\}$,

$$x_t' A_T^+ x_t \in [0, 1]. \tag{14}$$

(Details are given in Appendix A.) We also need the following inequality.

THEOREM 6.1 (Forster and Warmuth [6]). *For any T vectors $x_1, ..., x_T\in\mathbb{R}^n$ and for A_T given by* (13),

$$\sum_{t=1}^{T} x_t' A_T^+ x_t = \dim(\operatorname{span}\{x_1, ..., x_T\}) \leqslant n.$$

Vovk [22] proposed the prediction $b'_{T-1}A_T^+x_T$ for trial T in the on-line linear regression setting. Our new prediction algorithm multiplies Vovk's prediction by the factor

$$1-x'_TA_T^+x_T \overset{(14)}{\in} [0,1].$$

This new prediction was chosen so that in the proof of the following theorem all terms that are linear in y_t cancel out. It remains to be seen whether the factor improves the performance on some natural data.

The relative expected instantaneous loss bound we prove for the new algorithm is of the form $O(\frac{1}{T})$. We were not able to prove a bound of the same form for Vovk's algorithm nor for the standard least squares algorithm.

THEOREM 6.2. *Consider the following prediction algorithm Q for linear regression*

$$Q(z_1, ..., z_{T-1}, x_T) := (1-x'_TA_T^+x_T)\, b'_{T-1}A_T^+x_T.$$

Then for any distribution $\mathscr{D}$ on the examples with outcomes in $[-Y, Y]$ the relative expected instantaneous loss of Q is bounded as

$$\begin{aligned}
&\mathrm{E}_{(z_1, ..., z_T)\sim\mathscr{D}^T}((Q(z_1, ..., z_{T-1}, x_T)-y_T)^2)-\inf_{w\in\mathbb{R}^n}\mathrm{E}_{(x,y)\sim\mathscr{D}}((w\cdot x-y)^2)\\
&\leqslant 2\frac{1}{T}\mathrm{E}_{(z_1, ..., z_T)\sim\mathscr{D}^T}\left(\sum_{t=1}^{T} y_t^2x'_tA_T^+x_t\right)\\
&\leqslant 2Y^2\frac{1}{T}\mathrm{E}_{(z_1, ..., z_T)\sim\mathscr{D}^T}(\dim(\mathrm{span}\{x_1, ..., x_T\}))\\
&\leqslant 2nY^2\frac{1}{T}.
\end{aligned}$$

Proof. Because of Theorem 3.1 it suffices to bound the term

$$L_{Q,\mathrm{loo}}(z_1, ..., z_T)-\inf_{c\in\mathscr{C}} L_{c,\mathrm{se}}(z_1, ..., z_T) \tag{15}$$

for any T examples $z_1=(x_1, y_1), ..., z_T=(x_T, y_T)\in\mathscr{X}\times\mathscr{Y}$. Let

$$\xi_t := x'_tA_T^+x_t, \qquad \zeta_t := \left(\sum_{s\in\{1, ..., T\}\setminus\{t\}} y_sx_s\right)' A_T^+x_t$$

for $t\in\{1, ..., T\}$. The learner's prediction can be written as $(1-\xi_T)\,\zeta_T$.

The linear function with minimal sample error on the examples $z_1, ..., z_T$ is $\mathbb{R}^n\ni x\mapsto b'_TA_T^+x\in\mathbb{R}$. Because of $b'_TA_T^+x_t=y_t\xi_t+\zeta_t$ we can write the infimum in (15) as

$$\inf_{c\in\mathscr{C}} L_{c,\mathrm{se}}(z_1, ..., z_T) \overset{(7)}{=} \frac{1}{T}\sum_{t=1}^{T}((y_t\xi_t+\zeta_t)-y_t)^2. \tag{16}$$

Thus

$$
\begin{aligned}
&L_{Q,\mathrm{loo}}(z_1, \dots, z_T) - \inf_{c \in \mathscr{C}} L_{c,\mathrm{se}}(z_1, \dots, z_T) \\
&\overset{(6),\,(16)}{=} \frac{1}{T} \sum_{t=1}^{T} (((1-\xi_t)\,\zeta_t - y_t)^2 - ((y_t\xi_t + \zeta_t) - y_t)^2) \\
&= \frac{1}{T} \sum_{t=1}^{T} (\zeta_t^2 - 2\xi_t\zeta_t^2 + \xi_t^2\zeta_t^2 - 2y_t\zeta_t + 2y_t\xi_t\zeta_t \\
&\quad - y_t^2\xi_t^2 - 2y_t\xi_t\zeta_t - \zeta_t^2 + 2y_t^2\xi_t + 2y_t\zeta_t) \\
&= \frac{1}{T} \sum_{t=1}^{T} (\underbrace{-2\xi_t\zeta_t^2 + \underbrace{\xi_t^2\zeta_t^2}_{\overset{(14)}{\leqslant}\ \xi_t\zeta_t^2}}_{\leqslant\ -\xi_t\zeta_t^2\ \overset{(14)}{\leqslant}\ 0} - \underbrace{y_t^2\xi_t^2}_{\leqslant 0} + 2y_t^2\xi_t) \\
&\leqslant 2\frac{1}{T} \sum_{t=1}^{T} y^2 t x_t' A_T^+ x_t \\
&\overset{(14)}{\leqslant} 2Y^2 \frac{1}{T} \sum_{t=1}^{T} x_t' A_T^+ x_t \\
&\overset{\text{Theorem 6.1}}{\leqslant} 2nY^2 \frac{1}{T}. \qquad \blacksquare
\end{aligned}
$$

Note that the prediction of the learning algorithm in Theorem 6.2 is zero if $x_T \notin \operatorname{span}\{x_1, \dots, x_{T-1}\}$. If $x_T \in \operatorname{span}\{x_1, x_{T-1}\}$, then the Sherman–Morrison formula (see Press *et al.* [18]) shows that the prediction of Theorem 6.2 is equal to

$$
\begin{aligned}
&(1 - x_T' A_T^+ x_T)\, b_{T-1}' A_T^+ x_T \\
&= \left(1 - x_T' A_{T-1}^+ x_T + \frac{(x_T' A_{T-1}^+ x_T)^2}{1 + x_T' A_{T-1}^+ x_T}\right) b_{T-1}' A_{T-1}^+ x_T \left(1 - \frac{x_T' A_{T-1}^+ x_T}{1 + x_T' A_{T-1}^+ x_T}\right) \\
&= \frac{b_{T-1}' A_{T-1}^+ x_T}{(1 + x_T' A_{T-1}^+ x_T)^2}.
\end{aligned}
$$

Vovk's original prediction is the same as the above except that the denominator is not squared. The standard least squares algorithm is simply the numerator of the above. For all of these algorithms the prediction as a function of x_T is determined by the pseudoinverse A_{T-1}^+. If this pseudoinverse has been calculated in advance, then predictions for different values of x_T cost only $O(n^2)$ time.

7. LOWER BOUNDS

In this section we show lower bounds on the relative expected instantaneous loss (1) for the case when the comparators are Gaussian densities and for linear regression. We do this by adapting a lower bound for on-line linear regression of Vovk [22] to the off-line setting.

THEOREM 7.1. *For every learning algorithm Q for linear regression and for every $\varepsilon > 0$ there is a $T_0 \in \mathbb{N}$ such that for all $T \geqslant T_0$ there is a distribution $\mathcal{D}$ on the examples (with outcomes bounded by Y) such that the relative expected instantaneous loss is at least $(n-\varepsilon)\, Y^2 \frac{1}{T}$.*

For every learning algorithm Q for the case when the comparators are one-dimensional Gaussian densities there is a distribution $\mathcal{D}$ on $\{-1, 1\}$ such that the relative expected instantaneous loss is at least $(T+2\sqrt{T-1})^{-1}$.

Proof. For a fixed parameter $\alpha \geqslant 1$ we generate a distribution $\mathcal{D}$ on the examples with the following stochastic strategy: A vector $\theta \in [0, 1]^n$ is chosen from the prior distribution $\mathrm{Beta}(\alpha, \alpha)^n$, i.e., the components of θ are i.d.d. with distribution $\mathrm{Beta}(\alpha, \alpha)$. Then $\mathcal{D} = \mathcal{D}_\theta$ is the distribution for which the example $(e_i, 1)$ has probability θ_i / n and the example $(e_i, 0)$ has probability $(1-\theta_i)/n$. Here $e_1, ..., e_n$ are the unit vectors of $\mathbb{R}^n$. In each trial the examples are generated i.i.d. with $\mathcal{D}_\theta$. We can calculate the Bayes optimal learning algorithm for which the expectation (over the prior) of the relative expected instantaneous loss (1) is minimal. We show that for the Bayes optimal algorithm the expectation of (1) is

$$\frac{\alpha}{4\alpha+2} \frac{1}{n^{T-1}} \sum_{t=0}^{T-1} \binom{T-1}{t} \frac{(n-1)^{T-1-t}}{t+2\alpha}.$$

The lower bounds for the case when the comparators are Gaussian densities and for linear regression follow by suitably choosing α and with an affine transformation of the outcomes (using the fact that all instances are unit vectors).

Details of the proof are given in Appendix B. ▮

8. CONCLUSION

We have shown relative expected instantaneous loss bounds for off-line linear regression and various classes of densities as comparators where the associated loss function is not quadratic. We achieved this by using an inequality based on the leave-one-out loss. The upper and lower bounds for Gaussian comparators are identical and for linear regression they essentially differ by a factor of 2. We conjecture that the upper bound for linear regression can be improved by a factor of 2.

We believe that our techniques for proving relative expected instantaneous loss bounds using the leave-one-out loss can be applied to other learning problems. For example, they might be used to prove bounds for linear regression where the square loss as well as the quadratic regularization in (11) is exchanged for other functions. Good choices of loss function are the matching loss functions introduced in Herbster *et al.* [2] and Helmbold *et al.* [12]. Such loss functions can be defined in terms of Bregman divergences (see Kivinen and Warmuth [16]). The typical example is the entropic loss

$$y_t \ln \frac{y_t}{\hat{y}_t} + (1-y_t) \ln \frac{1-y_t}{1-\hat{y}_t}, \qquad \text{where} \quad \hat{y}_t = \frac{e^{w \cdot x_t}}{1+e^{w \cdot x_t}}.$$

Similarly, good candidates for the regularization term are Bregman divergences between w and some initial w_0 such as the relative entropy $\sum_{i=1}^{n} w_{0,i} \ln(w_i/w_{0,i})$. Our techniques should also be useful to prove relative expected instantaneous loss bounds for regularized classification problems where the classifier is a linear threshold function.

APPENDICES

Note that there is a fair amount of overlap between the following two appendices and the appendices of Forster and Warmuth [6].

APPENDIX A: OMITTED CALCULATIONS FOR LINEAR REGRESSION

Properties of the Pseudoinverse A_T^+

We begin showing that $A_T = \sum_{t=1}^{T} x_t x_t'$ (as defined in (13)) is always invertible on the subspace

$$X_T := \operatorname{span}\{x_1, ..., x_T\}$$

of $\mathbb{R}^n$.

Lemma A.1. *Let $X_T^\perp$ be the orthogonal complement of X_T in $\mathbb{R}^n$. Then the linear map*

$$A_T \colon X_T \to X_T$$

is invertible and

$$A_T \colon X_T^\perp \to X_T^\perp$$

is the zero function.

Proof. $A_T : X_T \to X_T$ is one-to-one because if $A_T x = 0$ holds for a vector $x \in X_T$ it follows that

$$0 = x' A_T x = \sum_{t=1}^{T} x' x_t x_t' x = \sum_{t=1}^{T} (x_t \cdot x)^2,$$

i.e., $x \in \{x_1, ..., x_T\}^\perp = X_T^\perp$. Thus $x \in X_T \cap X_T^\perp = \{0\}$. Since $\dim X_T$ is finite this also implies that $A_T : X_T \to X_T$ is onto.

Finally we have that for $x \in X_T^\perp$,

$$A_T x = \sum_{t=1}^{T} \underbrace{(x \cdot x_t)}_{=0} x_t = 0. \quad \blacksquare$$

Now we can calculate the pseudoinverse A_T^+ of the linear map A_T:

Lemma A.2. *The matrix A_T^+ is positive semi-definite and*

$$\begin{aligned} \forall x \in X_T, \qquad & A_T^+ A_T x = x = A_T A_T^+ x, \\ \forall x \in X_T^\perp, \qquad & A_T^+ x = 0. \end{aligned}$$

Proof. Since A_T maps X_T into X_T and $X_T^\perp$ into $X_T^\perp$ (by Lemma A.1) we can calculate the pseudoinverse on X_T and $X_T^\perp$ separately. Now we only have to note that the pseudoinverse of the zero function is the zero function and that the pseudoinverse of an invertible matrix is the inverse matrix. ▮

To show that $x_t' A_T^+ x_t \in [0, 1]$ for $t \in \{1, ..., T\}$ let $\xi_t := x_t' A_T^+ x_t$. $\xi_t \geqslant 0$ holds because A_T^+ is positive semi-definite. The following calculation shows that $\xi_t \leqslant 1$:

$$\begin{aligned} \xi_t - \xi_t^2 &= x_t'(A_T^+ - A_T^+ x_t x_t' A_T^+)\, x_t \\ &= x_t'(A_T^+ A_T A_T^+ - A_T^+ x_t x_t' A_T^+)\, x_t \\ &= x_t' A_T^+ \left(\sum_{s \in \{1, ..., T\} \setminus \{t\}} x_s x_s' \right) A_T^+ x_t \\ &= \sum_{s \in \{1, ..., T\} \setminus \{t\}} (x_s' A_T^+ x_t)^2 \geqslant 0. \end{aligned}$$

Linear Function with Minimal Sample Error

For $w \in \mathbb{R}^n$ let c be the linear function $c(x) = w \cdot x$ for $x \in \mathbb{R}^n$. The sample error is minimal for $w = A_T^+ b_T$ because

$$TL_{c,\,\mathrm{se}}(z_1, ..., z_T) \overset{(7)}{=} \sum_{t=1}^{T} (c(x_t) - y_t)^2 = w' A_T w - 2 b_T \cdot w + \sum_{t=1}^{T} y_t^2$$

is convex in w (because A_T is positive semi-definite) and its gradient with respect to w vanishes for $w = A_T^+ b_T$.

Prediction Is Zero If $x_T \notin X_{T-1}$

Because of $X_T \setminus X_{T-1} \neq \emptyset$ there exists a $z \in X_T \cap X_{T-1}^\perp$, $z \neq 0$. For this z,

$$A_T z \overset{z \in X_{T-1}^\perp}{=} x_T x_T' z = (z \cdot x_T)\, x_T.$$

Thus

$$(z \cdot x_T)\, A_T^+ x_T = A_T^+ A_T z \overset{z \in X_T}{=} z \in X_{T-1}^\perp \setminus \{0\}.$$

This shows that $A_T^+ x_T \in X_{T-1}^\perp$. Because of $b_{T-1} \in X_{T-1}$ the prediction is zero in this case.

Sherman–Morrison Formula

If $x_T \in \operatorname{span}\{x_1, ..., x_{T-1}\}$, then the Sherman–Morrison formula (see Press *et al.* [18]) applied to the linear function $A_T = A_{T-1} + x_T x'_T$ on $X_{T-1} = X_T$ shows that

$$c' A_T^+ d = c' A_{T-1}^+ d - \frac{(c' A_{T-1}^+ x_T)(d' A_{T-1}^+ x_T)}{1 + x'_T A_{T-1}^+ x_T}.$$

for all $c, d \in \operatorname{span}\{x_1, ..., x_T\}$.

APPENDIX B: PROOF OF THEOREM 7.1

The Beta Distribution

For parameters $\alpha, \beta > 0$, the distribution Beta(α, β) has the density function

$$\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}, \qquad \theta \in [0, 1],$$

with regard to the Lebesque measure on $[0, 1]$. Here

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} \, \mathrm{d}\theta = \frac{\Gamma(\alpha)\, \Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{17}$$

is the beta function. Γ is the gamma function which satisfies

$$\Gamma(\alpha) = (\alpha-1)\, \Gamma(\alpha-1)$$

for $\alpha \in \mathbb{R} \setminus \{1, 0, -1, -2, -3, ...\}$.

For every $\alpha > 0$,

$$\begin{aligned}
\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha, \alpha)}(\theta) &\overset{(17)}{=} \frac{B(\alpha+1, \alpha)}{B(\alpha, \alpha)} \overset{(17)}{=} \frac{\alpha}{2\alpha} = \frac{1}{2}, \\
\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha, \alpha)}(\theta^2) &\overset{(17)}{=} \frac{B(\alpha+2, \alpha)}{B(\alpha, \alpha)} \overset{(17)}{=} \frac{(\alpha+1)\, \alpha}{(2\alpha+1)(2\alpha)} = \frac{\alpha+1}{4\alpha+2}, \\
\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha, \alpha)}(\theta-\theta^2) &= \frac{1}{2} - \frac{\alpha+1}{4\alpha+2} = \frac{2\alpha+1-\alpha-1}{4\alpha+2} = \frac{\alpha}{4\alpha+2}.
\end{aligned}$$

The Bayes Optimal Learning Algorithm

We show that the learning algorithm

$$Q(z_1, ..., z_{T-1}, x_T) = \frac{k_T + \alpha}{K_T + 2\alpha}, \tag{18}$$

where

$$K_T = \sum_{t=1}^{T-1} x_t \cdot x_T, \qquad k_T = \sum_{t=1}^{T-1} y_t x_t \cdot x_T,$$

has minimal expected relative expected instantaneous loss in the stochastic setting we consider here. K_T counts how often the instance x_T occurs among the previous instances $x_1, ..., x_{T-1}$ and k_T counts how often the example $(x_T, 1)$ occurs among $(x_1, y_1), ..., (x_{T-1}, y_{T-1})$.

The expected loss of any learning algorithm in trial T is

$$\begin{aligned}
&\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha, \alpha)^n}\, \mathrm{E}_{(z_1, ..., z_T) \sim \mathcal{D}_\theta^T}((\hat{y}_T - y_T)^2) \\
&\quad = \int_{[0,1]^n} \left(\prod_{i=1}^n \frac{(\theta_i - \theta_i^2)^{\alpha-1}}{B(\alpha, \alpha)} \right) \\
&\qquad \times \sum_{\substack{x_1, ..., x_T \\ \in \{e_1, ..., e_n\}}} \sum_{\substack{y_1, ..., y_T \\ \in \{0, 1\}}} \left(\prod_{i=1}^n \left(\frac{1-\theta_i}{n} \right)^{\sum_{t=1}^T (1-y_t)\, x_t \cdot e_i} \left(\frac{\theta_i}{n} \right)^{\sum_{t=1}^T y_t x_t \cdot e_i} \right) (\hat{y}_T - y_T)^2 \, \mathrm{d}\theta \\
&\quad \overset{(17)}{=} \frac{1}{n^T} \sum_{\substack{x_1, ..., x_T \\ \in \{e_1, ..., e_n\}}} \sum_{\substack{y_1, ..., y_T \\ \in \{0, 1\}}} \left(\prod_{i=1}^n \frac{B(\alpha + \sum_{t=1}^T y_t x_t \cdot e_i, \alpha + \sum_{t=1}^T (1-y_t)\, x_t \cdot e_i)}{B(\alpha, \alpha)} \right) \\
&\qquad \times (\hat{y}_T - y_T)^2.
\end{aligned}$$

From this formula we see what the best learning algorithm in this setting is as follows: For fixed $x_1, ..., x_T$ and $y_1, ..., y_{T-1}$, the sum of the terms that depend on the prediction

$$\hat{y}_T(x_1, ..., x_T, y_1, ..., y_{T-1})$$

is a positive constant times

$$B(\alpha + k_T, \alpha + K_T - k_T + 1)\, \hat{y}_T^2 + B(\alpha + k_T + 1, \alpha + K_T - k_T)(\hat{y}_T - 1)^2.$$

This is minimal if

$$\hat{y}_T = \left(\frac{B(\alpha + k_T, \alpha + K_T - k_T + 1)}{B(\alpha + k_T + 1, \alpha + K_T - k_T)} + 1 \right)^{-1} = \left(\frac{\alpha + K_T - k_T}{\alpha + k_T} + 1 \right)^{-1} = \frac{k_T + \alpha}{K_T + 2\alpha}.$$

Loss of Bayes Optimal Algorithm

Let $\mathrm{Ber}(p)$ denote the Bernoulli distribution on $\{0, 1\}$ with mean $p \in [0, 1]$. The expected loss of the optimal learning algorithm (18) is

$$\begin{aligned}
&\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha, \alpha)^n}\, \mathrm{E}_{(z_1, ..., z_T) \sim \mathcal{D}_\theta^T} \left(\left(\frac{k_T + \alpha}{K_T + 2\alpha} - y_T \right)^2 \right) \\
&\quad = \frac{1}{n^T} \sum_{\substack{x_1, ..., x_T \\ \in \{e_1, ..., e_n\}}} \mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha, \alpha)^n}\, \mathrm{E}_{y_1 \sim \mathrm{Ber}(\theta \cdot x_1)} \cdots \mathrm{E}_{y_T \sim \mathrm{Ber}(\theta \cdot x_T)} \left(\left(\frac{k_T + \alpha}{K_T + 2\alpha} - y_T \right)^2 \right) \\
&\quad = \frac{\alpha}{4\alpha + 2} \frac{1}{n^T} \sum_{\substack{x_1, ..., x_T \\ \in \{e_1, ..., e_n\}}} \left(1 + \frac{1}{K_T + 2\alpha} \right).
\end{aligned}$$

For the last equality fix θ, x_1, ..., x_T. The instance x_T is some unit vector e_i. Thus $\theta \cdot x_T = \theta_i$. Since

$$\mathrm{E}_{y_T \sim \mathrm{Ber}(\theta \cdot x_T)}(y_T) = \mathrm{E}_{y_T \sim \mathrm{Ber}(\theta \cdot x_T)}(y_T^2) = \theta \cdot x_T = \theta_i,$$

$$\mathrm{E}_{y_1 \sim \mathrm{Ber}(\theta \cdot x_1)} \cdots \mathrm{E}_{y_{T-1} \sim \mathrm{Ber}(\theta \cdot x_{T-1})}(k_T) = K_T \theta_i,$$

$$\mathrm{E}_{y_1 \sim \mathrm{Ber}(\theta \cdot x_1)} \cdots \mathrm{E}_{y_{T-1} \sim \mathrm{Ber}(\theta \cdot x_{T-1})}(k_T^2) = (K_T^2 - K_T)\, \theta_i^2 + K_T \theta_i,$$

it follows that

$$\begin{aligned}
&\mathrm{E}_{y_1 \sim \mathrm{Ber}(\theta \cdot x_1)} \cdots \mathrm{E}_{y_T \sim \mathrm{Ber}(\theta \cdot x_T)} \left(\left(\frac{k_T + \alpha}{K_T + 2\alpha} - y_T \right)^2 \right) \\
&= \frac{(K_T^2 - K_T)\, \theta_i^2 + K_T \theta_i + 2\alpha K_T \theta_i + \alpha^2}{(K_T + 2\alpha)^2} - 2\theta_i \frac{K_T \theta_i + \alpha}{K_T + 2\alpha} + \theta_i \\
&= \theta_i - \theta_i^2 + \frac{1}{(K_T + 2\alpha)^2} (\theta_i^2 ((K_T + 2\alpha)^2 + K_T^2 - K_T - 2K_T (K_T + 2\alpha)) \\
&\quad + \theta_i (K_T + 2\alpha K_T - 2\alpha (K_T + 2\alpha)) + \alpha^2) \\
&= (\theta_i - \theta_i^2) \left(1 + \frac{K_T - 4\alpha^2}{(K_T + 2\alpha)^2} \right) + \frac{\alpha^2}{(K_T + 2\alpha)^2}.
\end{aligned}$$

Integrating the above over $\theta \in \mathrm{Beta}(\alpha, \alpha)^n$ gives

$$\frac{\alpha}{4\alpha + 2} \left(1 + \frac{K_T - 4\alpha^2 + 4\alpha^2 + 2\alpha}{(K_T + 2\alpha)^2} \right) = \frac{\alpha}{4\alpha + 2} \left(1 + \frac{1}{K_T + 2\alpha} \right).$$

The Comparison Term

For a fixed distribution $\mathscr{D}_\theta$, $\theta \in [0, 1]^n$, the comparison term is

$$\begin{aligned}
&\inf_{w \in \mathbb{R}^n} \mathrm{E}_{(x, y) \sim \mathscr{D}_\theta} (w \cdot x - y)^2 \\
&\quad = \inf_{w \in \mathbb{R}^n} \sum_{i=1}^n \left(\frac{1 - \theta_i}{n} w_i^2 + \frac{\theta_i}{n} (w_i - 1)^2 \right) \\
&\quad = \inf_{w \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n (w_i^2 - \theta_i w_i^2 + \theta_i w_i^2 - 2\theta_i w_i + \theta_i) \\
&\quad = \frac{1}{n} \inf_{w \in \mathbb{R}^n} \sum_{i=1}^n ((w_i - \theta_i)^2 + \theta_i (1 - \theta_i)) \\
&\quad = \frac{1}{n} \sum_{i=1}^n \theta_i (1 - \theta_i).
\end{aligned}$$

Integrating this over $\theta \sim \mathrm{Beta}(\alpha, \alpha)^n$ shows that the expectation (over the prior) of the comparison term is $\frac{\alpha}{4\alpha+2}$.

Lower Bounds on Relative Loss

Subtracting the expectation of the comparison term from the expected loss of the Bayes optimal algorithm shows that the expectation of the relative expected instantaneous loss (1) of the optimal learning algorithm is exactly

$$\frac{\alpha}{4\alpha+2}\,\frac{1}{n^T}\sum_{\substack{x_1,\ldots,x_T\\ \in\{e_1,\ldots,e_n\}}}\left(\frac{1}{K_T+2\alpha}\right). \tag{19}$$

Lower Bound for Linear Regression

The lower bound (19) can be written as $\frac{\alpha}{4\alpha+2}$ times

$$\frac{1}{n^{T-1}}\sum_{t=0}^{T-1}\binom{T-1}{t}\frac{(n-1)^{T-1-t}}{t+2\alpha}. \tag{20}$$

We show that this is at least

$$\frac{n}{T}\left(1-\left(1-\frac{1}{n}\right)^T\right)-(2\alpha-1)\,\frac{n^2}{T^2}. \tag{21}$$

This proves the lower bound of Theorem 7.1 for linear regression. (We use an affine transformation of the outcomes from $[0, 1]$ to $[-Y, Y]$. This transformation leads to a factor of $4Y^2$ in the lower bound.)

The bound (21) is a lower bound on (20) because

$$\frac{1}{t+2\alpha}=\frac{1}{t+1}-\frac{2\alpha-1}{(t+1)(t+2\alpha)}\overset{\alpha\geqslant 1}{\geqslant}\frac{1}{t+1}-\frac{2\alpha-1}{(t+1)(t+2)}$$

and

$$\begin{aligned}&\frac{1}{n^{T-1}}\sum_{t=0}^{T-1}\binom{T-1}{t}\frac{(n-1)^{T-1-t}}{t+1}\\&\quad=\frac{n}{T}\sum_{t=0}^{T-1}\binom{T}{T+1}\left(\frac{1}{n}\right)^{t+1}\left(1-\frac{1}{n}\right)^{T-(t+1)}=\frac{n}{T}\left(1-\left(1-\frac{1}{n}\right)^T\right),\end{aligned}$$

$$\begin{aligned}&\frac{1}{n^{T-1}}\sum_{t=0}^{T-1}\binom{T-1}{t}\frac{(n-1)^{T-1-t}}{(t+1)(t+2)}\\&\quad=\frac{n^2}{T(T+1)}\sum_{t=0}^{T-1}\binom{T+1}{t+2}\left(\frac{1}{n}\right)^{t+2}\left(1-\frac{1}{n}\right)^{(T+1)-(t+2)}\\&\quad\leqslant\frac{n^2}{T(T+1)}\leqslant\frac{n^2}{T^2}.\end{aligned}$$

Lower Bound for Gaussian Comparators

For $n=1$ the lower bound (19) is equal to

$$f(\alpha)=\frac{\alpha}{4\alpha+2}\frac{1}{T-1+2\alpha}.$$

To find the maximum of f we calculate $f'(\alpha)$:

$$\begin{aligned}f'(\alpha)&=\frac{4\alpha+2-4\alpha}{(4\alpha+2)^2}\frac{1}{T-1+2\alpha}+\frac{\alpha}{4\alpha+2}\frac{-2}{(T-1+2\alpha)^2}\\&=\frac{2}{(4\alpha+2)^2\,(T-1+2\alpha)}-\frac{2\alpha}{(4\alpha+2)(T-1+2\alpha)^2}\\&=\frac{2T-2+4\alpha-8\alpha^2-4\alpha}{(4\alpha+2)^2\,(T-1+2\alpha)^2}=\frac{2T-2-8\alpha^2}{(4\alpha+2)^2\,(T-1+2\alpha)^2}.\end{aligned}$$

Thus $f'(\alpha)=0$ holds if and only if $\alpha=\sqrt{T-1}/2$. For this α,

$$\begin{aligned}f\left(\frac{\sqrt{T-1}}{2}\right)&=\frac{\sqrt{T-1}}{2(2\sqrt{T-1}+2)}\frac{1}{T-1+\sqrt{T-1}}\\&=\frac{1}{4(\sqrt{T-1}+1)^2}=\frac{1}{4}\frac{1}{T+2\sqrt{T-1}}.\end{aligned}$$

This proves the lower bound for the case of Gaussian comparators. (Again we have to use an affine transformation of the outcomes that gives a factor of 4.)

ACKNOWLEDGMENTS

Many thanks to Eiji Takimoto, David McAllester, Hans-Ulrich Simon, Axel Munk, and the anonymous referees for helpful comments. Jürgen Forster was supported by a "DAAD Doktorandenstipendium im Rahmen des gemeinsamen Hochschulsonderprogramms III von Bund und Ländern," by the Deutsche Forschungsgemeinschaft Grant SI 498/4-1, and by a grant from the G.I.F., the German–Israeli Foundation for Scientific Research and Development. Manfred Warmuth was supported by the NSF Grant CCR-9821087.

REFERENCES

1. M. Anthony and P. L. Bartlett, "Neural Network Learning: Theoretical Foundations," Cambridge Univ. Press, Cambridge, UK, 1999.
2. P. Auer, M. Herbster, and M. K. Warmuth, Exponentially many local minima for single neurons, *in* "Proceedings of the 1995 Neural Information Processing Conference," pp. 316–317, MIT Press, Cambridge, MA, 1995.
3. K. Azoury and M. K. Warmuth, Relative loss bounds for on-line density estimation with the exponential family of distributions, *J. Mach. Learning* **43** (2001), 211–246.
4. L. M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, *USSR Comput. Math. Phys.* **7** (1967), 200–217.

5. J. Forster, On relative loss bounds in generalized linear regression, *in* "Proceedings of the Twelfth International Symposium on Fundamentals of Computation Theory," pp. 269–280, Springer-Verlag, Berlin, 1999.
6. J. Forster and M. K. Warmuth, Relative loss bounds for temporal-difference learning, *in* "Proceedings of the Seventeenth International Conference on Machine Learning," pp. 295–302, Morgan Kaufmann, San Francisco, 2000; *J. Mach. Learning*, to appear.
7. D. P. Foster, Prediction in the worst case, *Ann. Statist.* **19** (1991), 1084–1090.
8. Y. Freund, Predicting a binary sequence almost as well as the optimal biased coin, *in* "Proceedings of the Ninth Annual Conference on Computational Learning Theory," pp. 89–98, Assoc. Comput. Mach., New York, 1996.
9. G. J. Gordon, "Approximate Solutions to Markov Decision Processes," Ph.D. thesis, Department of Computer Science, Carnegie Mellon University, Pittsburgh, 1999; Technical Report CMU-CS-99-143.
10. A. L. Graps, An introduction to wavelets, *IEEE Comput. Sci. Engrg.* **2** (1995), 50–61.
11. D. Haussler, N. Littlestone, and M. K. Warmuth, Predicting $\{0, 1\}$-functions on randomly drawn points, *Inform. and Comput.* **115** (1994), 284–293.
12. D. P. Helmbold, J. Kivinen, and M. K. Warmuth, Relative loss bounds for single neurons, *IEEE Trans. Neural Networks* **10** (1999), 1291–1304.
13. H. Helmbold and M. K. Warmuth, On weak learning, *J. Comput. System Sci.* **50** (1995), 551–573.
14. J. Kivinen and M. K. Warmuth, Additive versus exponentiated gradient updates for linear prediction, *J. Inform. Comput.* **132** (1997), 1–64.
15. W. S. Lee, P. L. Bartlett, and R. C. Williamson, The importance of convexity in learning with squared loss, *IEEE Trans. Inform. Theory* **44** (1998), 1977–1998.
16. J. Kivinen and M. K. Warmuth, Relative loss bounds for multidimensional regression problems, *in* "Advances in Neural Information Processing Systems 10," pp. 287–293, MIT Press, Cambridge, MA, 1998.
17. E. L. Lehmann and G. Casella, "Theory of Point Estimation," 2nd ed., Springer-Verlag, New York, 1998.
18. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical Recipes in Pascal," Cambridge Univ. Press, Cambridge, UK, 1989.
19. K. Rektorys, "Survey of Applicable Mathematics," 2nd rev. ed., Kluwer Academic, Dordrecht, 1994.
20. C. Saunders, A. Gammerman, and V. Vovk, Ridge regression learning algorithm in dual variables, *in* "Proceedings of the Fifteenth International Conference on Machine Learning," Morgan Kaufmann, San Francisco, pp. 515–521, 1998.
21. Y. M. Shtarkov, Universal sequential coding of single messages, *Prob. Pered. Inf.* **23** (1987), 175–186.
22. V. Vovk, "Competitive On-line Linear Regression," Technical Report, CSD-TR-97-13, Department of Computer Science, Royal Holloway, University of London, 1997.
23. J. S. Walker, "Fast Fourier Transforms," 2nd ed., CRC Press, New York, 1996.
24. K. Yamanishi, A decision-theoretic extension of stochastic complexity and its applications to learning, *IEEE Trans. Inform. Theory* **44** (1998), 1424–1439.
25. K. Yamanishi, Extended stochastic complexity and minimax relative loss analysis, *in* "Algorithmic Learning Theory," Lecture Notes in Artificial Intelligence, Vol. 1720, pp. 26–38, Springer-Verlag, New York/Berlin, 1999.
26. Q. Xie and A. R. Barron, Minimax redundancy for the class of memoryless sources, *IEEE Trans. Inform. Theory* **43** (1997), 646–657.