Relative Expected Instantaneous Loss Bounds

Jürgen Forster
Computer Science Department
University of California, Santa Cruz
Santa Cruz, CA 95064
forster@lmi.ruhr-uni-bochum.de

Manfred Warmuth
Computer Science Department
University of California, Santa Cruz
Santa Cruz, CA 95064
manfred@cse.ucsc.edu

Abstract

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 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 bound 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 $O(\frac{1}{T})$ bounds. We also prove lower bounds that show that our upper bound on the relative expected instantaneous loss for Gaussian density estimation is optimal. In the case of linear regression we can show that our bounds are tight within a factor of two.

1 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* $\widehat{y}_t$ 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 $\widehat{y}_t$ 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 papers 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.

We obtain such bounds without using the powerful but rough machinery of the fat shattering dimension (See, e.g., Anthony and Bartlett [1]). Instead we want to build 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 [5], Vovk [14], Azoury and Warmuth [2], Forster [3], Gordon [6], Yamanishi [16], [17]). There are standard conversions of on-line algorithms to off-line algorithms (See Helmbold and Warmuth [9] and Kivinen and Warmuth [10]). 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 for Gaussian density estimation by an exact calculation. In the case of Bernoulli density estimation and linear regression we use a generalization of an inequality from Haussler, Littlestone and Warmuth [8] that is based on the leave-one-out loss.

We also give a lower bound that shows that our result for Gaussian density estimation is tight. 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.

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, Gammerman & Vovk [13]). Also Fourier or wavelet transforms can be used to extract frequency-dependent information from the instances, see, e.g., Walker [15] and Graps [7]. 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).

2 NOTATION AND PRELIMINARIES

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

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 $z_1 = (x_1, y_1)$,$\ldots$,$z_{T-1} = (x_{T-1}, y_{T-1})$ $\in \mathcal{X} \times \mathcal{Y}$ in trials 1 through $T-1$ and the instance $x_T \in \mathcal{X}$ in trial T, then its prediction is $Q(z_1, \ldots, z_{T-1}, x_T) \in \widehat{\mathcal{Y}}$. The loss of algorithm Q in trial T is

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

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 expected 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*

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

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

3 GAUSSIAN DENSITY ESTIMATION

In density estimation problems we do not use the instance space $\mathcal{X}$. One of the simplest density estimation problems is the prediction of the mean of a unit variance Gaussian. In this case the prediction space $\widehat{\mathcal{Y}}$ and the outcome space $\mathcal{Y}$ are both $\mathbb{R}^m$ for a fixed dimension $m \in \mathbb{N}$. The loss function is the squared Euclidean norm $L(\widehat{y}, y) = ||\widehat{y} - y||^2$ and the comparison class $\mathcal{C}$ for the best mean vector is again $\mathbb{R}^m$.

For Gaussian density estimation the comparison term in (1) is the variance of the unknown distribution $\mathcal{D}$, i.e.

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

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

$$\mathrm{E}_{y\sim\mathcal{D}}\Big(||c-y||^2\Big) = ||c||^2 - 2c \cdot \mathrm{E}_{y\sim\mathcal{D}}(y) + \mathrm{E}_{y\sim\mathcal{D}}(||y||^2) \qquad (3)$$

is convex in c. Setting the gradient of (3) with respect to c to zero shows that the infimum of (3) is attained for $c = \mathrm{E}_{y\sim\mathcal{D}}(y)$. For this c, (3) is equal to

$$\mathrm{E}_{y\sim\mathcal{D}}\Big(||y||^2\Big) - ||\mathrm{E}_{y\sim\mathcal{D}}(y)||^2 \ .$$

Theorem 3.1 *Consider the following prediction algorithm Q for Gaussian density estimation*

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

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

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

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

$$\begin{aligned}
&\mathrm{E}_{(y_1,\ldots,y_T)\sim\mathcal{D}^T}\left(||\eta\sum_{t=1}^{T-1} y_t - y_T||^2\right) - \inf_{c\in\mathbb{R}^m} \mathrm{E}_{y\sim\mathcal{D}}\Big(||c-y||^2\Big) \\
&\overset{(2)}{=} \mathrm{E}_{(y_1,\ldots,y_T)\sim\mathcal{D}^T}\left(\eta^2||\sum_{t=1}^{T-1} y_t||^2 - 2\eta\sum_{t=1}^{T-1} y_t \cdot y_T + ||y_T||^2\right) \\
&\quad -\mathrm{E}_{y\sim\mathcal{D}}\Big(||y||^2\Big) + ||\mathrm{E}_{y\sim\mathcal{D}}(y)||^2 \\
&= \eta^2(T-1)\mathrm{E}_{y\sim\mathcal{D}}\Big(||y||^2\Big) + ||\mathrm{E}_{y\sim\mathcal{D}}(y)||^2 \cdot \\
&\quad \cdot \underbrace{(\eta^2(T-1)(T-2) - 2\eta(T-1) + 1)}_{=0} \\
&= \frac{1}{T + 2\sqrt{T-1}} \mathrm{E}_{y\sim\mathcal{D}}\Big(||y||^2\Big) \ .
\end{aligned}$$

For the second equality we used

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

□

The prediction used in Theorem 3.1 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 $c \geq 0$ is the multiplicity of this mean. We chose $y_0 = 0$ and $c = \sqrt{T-1}$. In Azoury and Warmuth [2] worst-case on-line total loss bounds were proven for the algorithm that uses $y_0 = 0$ and $c = 1$. The relative expected instantaneous loss bounds for the latter algorithm are slightly weaker.

In Section 7 we show that the prediction algorithm of Theorem 3.1 is optimal in a very strong sense: For every learning algorithm Q' for Gaussian density estimation there is a simple distribution $\mathcal{D}$ on two points for which the relative expected instantaneous loss of Q' is at least as large as the bound of Theorem 3.1.

4 THE LEAVE-ONE-OUT LOSS

The bound for Gaussian density estimation given in the previous section was proven by an exact calculation. For the relative expected instantaneous loss bounds for Bernoulli density estimation and linear regression we need to use a general inequality given in Theorem 4.1 of this section. This theorem is a generalization of a similar theorem given in Haussler, Littlestone and Warmuth [8]. Theorem 4.1 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 $z_1, \ldots, z_T \in \mathcal{X} \times \mathcal{Y}$ is the average of the losses of the learning algorithm on the last example of certain permutations of the sequence:

$$L_{Q,\mathrm{loo}}(z_1, \ldots, z_T) := \frac{1}{T}\sum_{t=1}^{T} L(Q(z_1, \ldots, z_{t-1}, z_{t+1}, \ldots, z_T, x_t), y_t) \ . \tag{4}$$

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

$$L_{c,\mathrm{se}}(z_1, \ldots, z_T) := \frac{1}{T}\sum_{t=1}^{T} L(c(x_t), y_t) \ . \tag{5}$$

Theorem 4.1 *The relative expected instantaneous loss (1) of any learning algorithm Q is bounded as follows:*

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

Proof. This holds because (1) is the sum of

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

and of

$$\begin{aligned}
&- \inf_{c\in\mathcal{C}} \mathrm{E}_{(x,y)\sim\mathcal{D}}\Big(L(c(x), y)\Big) \\
&\overset{(5)}{=} - \inf_{c\in\mathcal{C}} \mathrm{E}_{(z_1,\ldots,z_T)\sim\mathcal{D}^T}\Big(L_{c,\mathrm{se}}(z_1, \ldots, z_T)\Big) \\
&\leq -\mathrm{E}_{(z_1,\ldots,z_T)\sim\mathcal{D}^T}\Big(\inf_{c\in\mathcal{C}} L_{c,\mathrm{se}}(z_1, \ldots, z_T)\Big) \ .
\end{aligned}$$

□

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, Littlestone and Warmuth [8] the comparison term is zero.

For Gaussian density estimation the infimum of the sample error that appears in the bounds of Theorem 4.1 is the sample variance:

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

This follows because

$$\begin{aligned}
L_{c,\mathrm{se}}(y_1, \ldots, y_T) &\overset{(5)}{=} \frac{1}{T}\sum_{t=1}^{T} ||c - y_t||^2 \\
&= ||c||^2 - 2c \cdot \frac{1}{T}\sum_{t=1}^{T} y_t + \frac{1}{T}\sum_{t=1}^{T} ||y_t||^2
\end{aligned}$$

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.

5 BERNOULLI DENSITY ESTIMATION

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

$$L(\widehat{y}, y) = -y \ln \widehat{y} - (1-y)\ln(1-\widehat{y}) \qquad (6)$$

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 $L(0,1) = L(1,0) = \infty$. Note that as for Gaussian density estimation we do not use the instance space $\mathcal{X}$.

Theorem 5.1 *Consider the following prediction algorithm Q for Bernoulli density estimation*

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

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

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

Proof. Because of Theorem 4.1 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 $\ln\left(1+\frac{1}{T}\right)$. 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

$$\begin{aligned} L_{c,\mathrm{se}}(y_1,\ldots,y_T) &\stackrel{(5)}{=} \frac{1}{T}\sum_{t=1}^T L(c, y_t) \\ &\stackrel{(6)}{=} \frac{1}{T}\left(-\sum_{t=1}^T y_t \ln c - \sum_{t=1}^T (1-y_t)\ln(1-c)\right) \\ &= -\bar{y}\ln c - (1-\bar{y})\ln(1-c) \ . \end{aligned}$$

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 $y_1, \ldots, y_T$ is

$$\begin{aligned} &L_{Q,\mathrm{loo}}(y_1,\ldots,y_T) \\ &\stackrel{(4)}{=} \frac{1}{T}\sum_{t=1}^T L\Big(\frac{1+\sum_{s\in\{1,\ldots,T\}\setminus\{t\}} y_s}{T+1}, y_t\Big) \\ &= \frac{1}{T}\sum_{\substack{t\in\{1,\ldots,T\}\\ y_t=1}} L\Big(\frac{T\bar{y}}{T+1}, 1\Big) \\ &\qquad + \frac{1}{T}\sum_{\substack{t\in\{1,\ldots,T\}\\ y_t=0}} L\Big(\frac{1+T\bar{y}}{T+1}, 0\Big) \\ &\stackrel{(6)}{=} -\bar{y}\ln\left(\frac{T\bar{y}}{T+1}\right) - (1-\bar{y})\ln\left(\frac{T(1-\bar{y})}{T+1}\right) \ . \end{aligned}$$

Thus

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

□

As in the case of Gaussian density estimation our algorithm again 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 Laplace estimator for which $c = 1$. For the Laplace estimator we could only prove a slightly weaker bound.

6 OVERVIEW OF OLD AND NEW RESULTS FOR LINEAR REGRESSION

6.1 KNOWN RESULTS

In linear regression the instance space is $\mathcal{X} = \mathbb{R}^n$ for a fixed dimension n. The prediction space is $\widehat{\mathcal{Y}} = \mathbb{R}$ and the outcome space is $\mathcal{Y} = [-Y, Y] \subseteq \mathbb{R}$. 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. $L(\widehat{y}, y) = (\widehat{y} - y)^2$. The comparison class $\mathcal{C}$ consists of the linear functions $c : \mathbb{R}^n \to \mathbb{R}$. Thus for off-line linear regression the relative expected instantaneous loss (1) of a learning algorithm Q is

$$\begin{aligned} &\mathrm{E}_{(z_1,\ldots,z_T)\sim\mathcal{D}^T}\left((Q(z_1,\ldots,z_{T-1},x_T) - y_T)^2\right) \\ &\qquad - \inf_{w\in\mathbb{R}^n} \mathrm{E}_{(x,y)\sim\mathcal{D}}\left((w\cdot x - y)^2\right) \ , \qquad (7) \end{aligned}$$

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

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

$$\begin{aligned} &\sum_{t=1}^T (Q(z_1,\ldots,z_{t-1},x_t) - y_t)^2 \\ &\qquad - \inf_{w\in\mathbb{R}^n}\left(\sum_{t=1}^T (w\cdot x_t - y_t)^2 + a||w||^2\right) \ , \qquad (8) \end{aligned}$$

where $z_1 = (x_1, y_1), \ldots, z_T = (x_T, y_T)$ are the T examples of trials 1 through T. Here $a \geq 0$ is a constant and the term $a||w||^2$ is a measure of the complexity of the vector w. The larger a, the smaller the expression (8). Bounds on (8) 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 $a||w||^2$ in (8) to show our relative expected instantaneous loss bounds for off-line linear regression.

Vovk [14] proposed a learning algorithm for on-line linear regression for which he showed that the relative loss (8) is at most $nY^2\ln(1+\frac{TX^2}{an})$ 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 $\mathcal{D}$ on the examples for which the expectation of (8),

$$\mathrm{E}_{(z_1,\dots,z_T)\sim\mathcal{D}^T}\Big(\sum_{t=1}^{T}(Q(z_1,\dots,z_{t-1},x_t)-y_t)^2 - \inf_{w\in\mathbb{R}^n}\sum_{t=1}^{T}(w\cdot x_t-y_t)^2\Big) \qquad (9)$$

is at least $(n-o(1))Y^2\ln T$ as $T\to\infty$. Forster and Warmuth [4] give a corresponding upper bound of $nY^2(1+\ln T)$ on (9) that holds for any distribution $\mathcal{D}$ on the examples.

Vovk's algorithm for on-line linear regression and the bound $nY^2(1+\ln T)$ on (9) can be converted to a learning algorithm for off-line linear regression with a relative expected instantaneous loss bound of $nY^2\frac{(1+\ln T)}{T}$.

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

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 Figure 1.

6.2 NEW ALGORITHM AND PROOFS

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

$$b_T:=\sum_{t=1}^{T}y_tx_t\in\mathbb{R}^n\ , \qquad A_T:=\sum_{t=1}^{T}x_tx_t'\in\mathbb{R}^{n\times n}\ . \qquad (10)$$

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 [12]. There it is also shown how the pseudoinverse of a matrix can be computed from the singular value decomposition. The pseudoinverse A_T^+ is positive semi-definite and $A_T^+A_Tx=x=A_TA_T^+x$ holds for all $x\in\operatorname{span}\{x_1,\dots,x_T\}$. For all $t\in\{1,\dots,T\}$,

$$x_t'A_T^+x_t\in[0,1]\ , \qquad (11)$$

and if x_t is linearly independent of the remaining $T-1$ vectors then $x_t'A_T^+x_t=1$. (Details are given in Appendix A.) We also need the following inequality.

Theorem 6.1 (Forster and Warmuth [4]) *For any T vectors $x_1,\dots,x_T\in\mathbb{R}^n$ and for A_T given by (10):*

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

Vovk [14] 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

$$1-x_T'A_T^+x_T\overset{(11)}{\in}[0,1]\ .$$

This new prediction was chosen so that in the the proof of 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*

$$Q(z_1,\dots,z_{T-1},x_T):=(1-x_T'A_T^+x_T)b_{T-1}'A_T^+x_T\ .$$

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

$$\begin{aligned}&\mathrm{E}_{(z_1,\dots,z_T)\sim\mathcal{D}^T}\Big((Q(z_1,\dots,z_{T-1},x_T)-y_T)^2\Big)\\&\quad-\inf_{w\in\mathbb{R}^n}\mathrm{E}_{(x,y)\sim\mathcal{D}}\big((w\cdot x-y)^2\big)\\&\le 2\frac{1}{T}\mathrm{E}_{(z_1,\dots,z_T)\sim\mathcal{D}^T}\Big(\sum_{t=1}^{T}y_t^2x_t'A_T^+x_t\Big)\\&\le 2Y^2\frac{1}{T}\mathrm{E}_{(z_1,\dots,z_T)\sim\mathcal{D}^T}\Big(\dim(\operatorname{span}\{x_1,\dots,x_T\})\Big)\\&\le 2nY^2\frac{1}{T}\ .\end{aligned}$$

Proof. Because of Theorem 4.1 it suffices to bound the term

$$L_{Q,\mathrm{loo}}(z_1,\dots,z_T)-\inf_{c\in\mathcal{C}}L_{c,\mathrm{se}}(z_1,\dots,z_T) \qquad (12)$$

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

$$\xi_t:=x_t'A_T^+x_t\ , \qquad \zeta_t:=\Bigg(\sum_{s\in\{1,\dots,T\}\setminus\{t\}}y_sx_s\Bigg)'A_T^+x_t$$

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

The linear function with minimal sample error on the examples $z_1,\dots,z_T$ is $\mathbb{R}^n\ni x\mapsto b_T'A_T^+x\in\mathbb{R}$. Because of $b_T'A_T^+x_t=y_t\xi_t+\zeta_t$ we can write the infimum in (12) as

$$\inf_{c\in\mathcal{C}}L_{c,\mathrm{se}}(z_1,\dots,z_T)\overset{(5)}{=}\frac{1}{T}\sum_{t=1}^{T}((y_t\xi_t+\zeta_t)-y_t)^2\ . \qquad (13)$$

Thus

$$\begin{aligned}&L_{Q,\mathrm{loo}}(z_1,\dots,z_T)-\inf_{c\in\mathcal{C}}L_{c,\mathrm{se}}(z_1,\dots,z_T)\\&\overset{(4),(13)}{=}\frac{1}{T}\sum_{t=1}^{T}\big(((1-\xi_t)\zeta_t-y_t)^2\end{aligned}$$

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	$nY^2(1+\ln T)$	$(n-o(1))Y^2\ln T$

Figure 1: Upper and lower relative loss bounds for linear regression

$$
\begin{aligned}
&-((y_t\xi_t+\zeta_t)-y_t)^2) \\
&= \frac{1}{T}\sum_{t=1}^{T}\Big(\zeta_t^2-2\xi_t\zeta_t^2+\xi_t^2\zeta_t^2-2y_t\zeta_t+2y_t\xi_t\zeta_t \\
&\qquad -y_t^2\xi_t^2-2y_t\xi_t\zeta_t-\zeta_t^2+2y_t^2\xi_t+2y_t\zeta_t\Big) \\
&= \frac{1}{T}\sum_{t=1}^{T}\Big(\underbrace{-2\xi_t\zeta_t^2+\underbrace{\xi_t^2\zeta_t^2}_{\overset{(11)}{\le}\xi_t\zeta_t^2}}_{\le-\xi_t\zeta_t^2\overset{(11)}{\le}0}\underbrace{-y_t^2\xi_t^2}_{\le 0}+2y_t^2\xi_t\Big) \\
&\le 2\frac{1}{T}\sum_{t=1}^{T}y_t^2x_t'A_T^+x_t \overset{(11)}{\le} 2Y^2\frac{1}{T}\sum_{t=1}^{T}x_t'A_T^+x_t \\
&\overset{\text{Theorem 6.1}}{\le} 2nY^2\frac{1}{T}\ .
\end{aligned}
$$

□

Note that the prediction of the learning algorithm in Theorem 6.2 is zero if $x_T \notin \operatorname{span}\{x_1,\ldots,x_{T-1}\}$. If $x_T \in \operatorname{span}\{x_1,\ldots,x_{T-1}\}$, then the Sherman-Morrison formula (see Press et al. [11]) shows that the prediction of Theorem 6.2 is equal to

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

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 enumerator 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 costs only $O(n^2)$ time.

7 LOWER BOUNDS

In this section we show lower bounds on the relative expected instantaneous loss (1) for Gaussian density estimation and for linear regression. We do this by adapting a lower bound for on-line linear regression of Vovk [14] 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 $T \geq T_0$ 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 one-dimensional Gaussian density estimation 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 $\alpha \geq 1$ 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.i.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 $\frac{\theta_i}{n}$ and the example $(e_i, 0)$ has probability $\frac{1-\theta_i}{n}$. Here $e_1,\ldots,e_n$ 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 Gaussian density estimation and 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. □

ACKNOWLEDGMENTS

Many thanks to Eiji Takimoto for helpful discussions. Jürgen Forster was supported by a "DAAD Doktorandenstipendium im Rahmen des gemeinsamen Hochschulsonderprogramms III von Bund und Ländern". Manfred Warmuth was supported by the NSF grant CCR-9821087.

References

[1] Anthony, M., & Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge: Cambridge University Press.

[2] Azoury, K., & Warmuth, M. K. (1999). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 31–40). San Francisco: Morgan Kaufmann.

[3] Forster, J. (1999). On relative loss bounds in generalized linear regression. *Proceedings of the Twelfth International Symposium on Fundamentals of Computation Theory* (pp. 269–280). Berlin: Springer.

[4] Forster, J., & Warmuth, M. K. (in press). Relative loss bounds for temporal-difference learning. *Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann.

[5] Foster, D. P. (1991). Prediction in the worst case. *The Annals of Statistics*, *19*, 1084–1090.

[6] Gordon, G. J. (1999). *Approximate Solutions to Markov Decision Processes*. Ph. D. thesis, Department of Computer Science, Carnegie Mellon University, Pittsburgh. Technical report CMU-CS-99-143.

[7] Graps, A. L. (1995). An introduction to wavelets. *IEEE Computational Sciences and Engineering*, *2*, 50–61.

[8] Haussler, D., Littlestone, N. & Warmuth, M. K. (1994). Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, *115*, 284–293.

[9] Helmbold, H., & Warmuth, M. K. (1995). On weak learning, *Journal of Computer and System Sciences*, *50*, 551–573.

[10] Kivinen, J., & Warmuth, M. K. (1997). Additive versus exponentiated gradient updates for linear prediction, *Journal of Information and Computation*, *132*, 1–64.

[11] Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical Recipes in Pascal*. Cambridge: Cambridge University Press.

[12] Rektorys, K. (1994). *Survey of Applicable Mathematics*, 2nd rev. ed. Kluwer Academic Publishers.

[13] Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 515–521). San Francisco: Morgan Kaufmann.

[14] Vovk, V. (1997). *Competitive on-line linear regression*. Technical Report CSD-TR-97-13, Department of Computer Science, Royal Holloway, University of London.

[15] Walker, J. S. (1996). *Fast Fourier Transforms*, 2nd ed. New York: CRC Press.

[16] Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transaction on Information Theory*, *44*, 1424-1439.

[17] Yamanishi, K. (1999). Extended stochastic complexity and minimax relative loss analysis. *Algorithmic Learning Theory: Lecture Notes in Artificial Intelligence 1720*, 26-38.

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 (10)) is always invertible on the subspace

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

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*

$$A_T : X_T \to X_T$$

is invertible and

$$A_T : X_T^\perp \to X_T^\perp$$

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. $x \in \{x_1, \ldots, x_T\}^\perp = X_T^\perp$. 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 $x \in X_T^\perp$:

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

□

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*

$$\begin{array}{ll} \forall x \in X_T : & A_T^+ A_T x = x = A_T A_T^+ x \ , \\ \forall x \in X_T^\perp : & A_T^+ x = 0 \end{array}$$

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 on $X_T^\perp$ 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 $t \in \{1, \ldots, T\}$ let $\xi_t := x_t' A_T^+ x_t$. $\xi_t \geq 0$ holds because A_T^+ is positive semi-definite. The following calculation shows that $\xi_t \leq 1$:

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

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

$$T L_{c,\mathrm{se}}(z_1, \ldots, z_T) \overset{(5)}{=} \sum_{t=1}^{T} (c(x_t) - y_t)^2$$

$$= w' A_T w - 2 b_T \cdot w + \sum_{t=1}^{T} y_t^2 \quad .$$

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

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

Thus

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

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 $x_T \in \mathrm{span}\{x_1, \ldots, x_{T-1}\}$, then the Sherman-Morrison formula (see Press et al. [11]) applied to the linear function $A_T = A_{T-1} + x_T x_T'$ on $X_{T-1} = X_T$ shows that

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

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

APPENDIX B: PROOF OF THEOREM 7.1

THE BETA DISTRIBUTION

For parameters $\alpha, \beta > 0$, the distribution $\mathrm{Beta}(\alpha, \beta)$ has the density function

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

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

$$B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,\mathrm{d}\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad (14)$$

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

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

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

For every $\alpha > 0$:

$$\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha,\alpha)}(\theta) \overset{(14)}{=} \frac{B(\alpha+1,\alpha)}{B(\alpha,\alpha)} \overset{(14)}{=} \frac{\alpha}{2\alpha} = \frac{1}{2} \quad ,$$

$$\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha,\alpha)}(\theta^2) \overset{(14)}{=} \frac{B(\alpha+2,\alpha)}{B(\alpha,\alpha)}$$

$$\overset{(14)}{=} \frac{(\alpha+1)\alpha}{(2\alpha+1)(2\alpha)} = \frac{\alpha+1}{4\alpha+2} \quad ,$$

$$\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha,\alpha)}(\theta - \theta^2) = \frac{1}{2} - \frac{\alpha+1}{4\alpha+2}$$

$$= \frac{2\alpha + 1 - \alpha - 1}{4\alpha+2} = \frac{\alpha}{4\alpha+2} \quad .$$

THE BAYES OPTIMAL LEARNING ALGORITHM

We show that the learning algorithm

$$Q(z_1, \ldots, z_{T-1}, x_T) = \frac{k_T + \alpha}{K_T + 2\alpha} \quad , \qquad (15)$$

where

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

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 $x_1, \ldots, x_{T-1}$ and k_T counts how often the example $(x_T, 1)$ occurs among $(x_1, y_1), \ldots, (x_{T-1}, y_{T-1})$.

The expected loss of any learning algorithm in trial T is

$$\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha,\alpha)^n} \mathrm{E}_{(z_1,\ldots,z_T) \sim \mathcal{D}_\theta^T} \left((\widehat{y}_T - y_T)^2 \right)$$

$$= \int_{[0,1]^n} \left(\prod_{i=1}^{n} \frac{(\theta_i - \theta_i^2)^{\alpha-1}}{B(\alpha,\alpha)} \right) \sum_{\substack{x_1,\ldots,x_T \\ \in \{e_1,\ldots,e_n\}}} \sum_{\substack{y_1,\ldots,y_T \\ \in \{0,1\}}}$$

$$\left(\prod_{i=1}^{n} \left(\frac{1-\theta_i}{n} \right)^{\sum_{t=1}^{T}(1-y_t) x_t \cdot e_i} \left(\frac{\theta_i}{n} \right)^{\sum_{t=1}^{T} y_t x_t \cdot e_i} \right)$$

$$\cdot (\widehat{y}_T - y_T)^2 \,\mathrm{d}\theta$$

$$\overset{(14)}{=} \frac{1}{n^T} \sum_{\substack{x_1,\ldots,x_T \\ \in \{e_1,\ldots,e_n\}}} \sum_{\substack{y_1,\ldots,y_T \\ \in \{0,1\}}}$$

$$\left(\prod_{i=1}^{n} \frac{B(\alpha + \sum_{t=1}^{T} y_t x_t \cdot e_i, \alpha + \sum_{t=1}^{T} (1-y_t) x_t \cdot e_i)}{B(\alpha,\alpha)} \right)$$

$$\cdot (\widehat{y}_T - y_T)^2 \quad .$$

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

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

is a positive constant times

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

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

This is minimal if

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

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

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 (15) is

$$\mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha,\alpha)^n} \mathrm{E}_{(z_1,\ldots,z_T) \sim \mathcal{D}_\theta^T} \left(\left(\frac{k_T + \alpha}{K_T + 2\alpha} - y_T \right)^2 \right)$$

$$= \frac{1}{n^T} \sum_{\substack{x_1,\ldots,x_T \\ \in \{e_1,\ldots,e_n\}}} \mathrm{E}_{\theta \sim \mathrm{Beta}(\alpha,\alpha)^n}$$

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

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

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

it follows that

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

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

$$
\begin{aligned}
&\frac{\alpha}{4\alpha+2}\left(1+\frac{K_T-4\alpha^2+4\alpha^2+2\alpha}{(K_T+2\alpha)^2}\right)\\
&\qquad=\frac{\alpha}{4\alpha+2}\left(1+\frac{1}{K_T+2\alpha}\right) .
\end{aligned}
$$

THE COMPARISON TERM

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

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

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 BOUND 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

$$\frac{\alpha}{4\alpha+2}\frac{1}{n^T}\sum_{\substack{x_1,\ldots,x_T\\ \in\{e_1,\ldots,e_n\}}}\left(\frac{1}{K_T+2\alpha}\right) . \qquad (16)$$

LOWER BOUND FOR LINEAR REGRESSION

The lower bound (16) can be written as $\frac{\alpha}{4\alpha+2}$ times

$$\frac{1}{n^{T-1}}\sum_{t=0}^{T-1}\binom{T-1}{t}\frac{(n-1)^{T-1-t}}{t+2\alpha} . \qquad (17)$$

We show that this is at least

$$\frac{n}{T}\left(1-\left(1-\frac{1}{n}\right)^T\right)-(2\alpha-1)\frac{n^2}{T^2} . \qquad (18)$$

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.)

(18) is a lower bound on (17) because

$$
\begin{aligned}
\frac{1}{t+2\alpha}&=\frac{1}{t+1}-\frac{2\alpha-1}{(t+1)(t+2\alpha)}\\
&\overset{\alpha\geq 1}{\geq}\frac{1}{t+1}-\frac{2\alpha-1}{(t+1)(t+2)}
\end{aligned}
$$

and

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

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

LOWER BOUND FOR GAUSSIAN DENSITY ESTIMATION

For $n = 1$ the lower bound (16) is equal to

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

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

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

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

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

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