On Relative Loss Bounds in Generalized Linear Regression

Jürgen Forster

Universität Bochum, Germany
forster@lmi.ruhr-uni-bochum.de

Abstract. When relative loss bounds are considered, an on-line learning algorithm is compared to the performance of a class of off-line algorithms, called experts. In this paper we reconsider a result by Vovk, namely an upper bound on the on-line relative loss for linear regression with square loss – here the experts are linear functions. We give a shorter and simpler proof of Vovk's result and give a new motivation for the choice of the predictions of Vovk's learning algorithm. This is done by calculating the, in some sense, best prediction for the last trial of a sequence of trials when it is known that the outcome variable is bounded. We try to generalize these ideas to the case of generalized linear regression where the experts are neurons and give a formula for the "best" prediction for the last trial in this case, too. This prediction turns out to be essentially an integral over the "best" expert applied to the last instance. Predictions that are "optimal" in this sense might be good predictions for long sequences of trials as well.

1 Introduction

In the *on-line learning* protocols we consider here, a *learning algorithm* called *Learner* tries to predict real numbers in a sequence of trials. Real-world examples of applications for this protocol are weather or stockmarket predictions, or pattern recognition. This protocol can be seen as a game between Learner and an opponent, *Nature*. After Learner receives an instance x_t in the t–th trial, it makes a prediction $\hat{y}_t$ and Nature responds with the correct outcome y_t. Learner wants to keep the discrepancy between $\hat{y}_t$ and y_t as small as possible. This discrepancy is measured with a *loss function* L, and the total loss of Learner on a sequence of trials is the sum of the losses in each trial. One way to measure the quality of Learner's predictions is to compare the loss of Learner to that of a class of functions from the set of instances to the set of outcomes (such functions are called *experts*), i.e. to give *relative loss bounds*.

Like in [6], we consider the following protocol of interaction between Learner and Nature:

FOR $t = 1, 2, 3, \ldots, T$
  Nature chooses an instance $x_t \in \mathbb{R}^n$
  Learner chooses a prediction $\hat{y}_t \in \mathbb{R}$

G. Ciobanu and G. Păun (Eds.): FCT'99, LNCS 1684, pp. 269–280, 1999.

Nature chooses an outcome $y_t \in \mathbb{R}$
END FOR

Learner does not necessarily know the number of trials T in advance. After t trials, the loss of Learner is

$$L_t(\text{Learner}) := \sum_{s=1}^{t} (y_s - \hat{y}_s)^2 \quad . \tag{1}$$

We use the following notations: The vectors $x \in \mathbb{R}^n$ are column vectors. $x^\top$, the transposed vector of x, is a row vector. For $m, n \in \mathbb{N}$, $\mathbb{R}^{m \times n}$ is the set of real $m \times n$ matrices. The scalar product of $x, y \in \mathbb{R}^n$ is $x \cdot y = x^\top y = \sum_{i=1}^n x_i y_i$ and the 2-norm of x is $\|x\| = (x \cdot x)^{\frac{1}{2}}$. $I \in \mathbb{R}^{n \times n}$ is the $n \times n$ identity matrix. A matrix $A \in \mathbb{R}^{n \times n}$ is called positive semidefinite if $x^\top A x \geq 0$ for all $x \in \mathbb{R}^n$ and it is called positive definite if $x^\top A x > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.

We search for strategies for Learner that ensure that its loss is not much larger than the loss of the, in some sense, best linear expert. A linear expert $w \in \mathbb{R}^n$ makes the prediction $w \cdot x$ on instance $x \in \mathbb{R}^n$ and its loss on the first t trials is

$$L_t(w) := \sum_{s=1}^{t} (y_s - w \cdot x_s)^2 \quad . \tag{2}$$

For a fixed $a > 0$, we try to minimize

$$L_T(\text{Learner}) - \inf_{w \in \mathbb{R}^n} \Big(a\|w\|^2 + L_T(w) \Big) \quad . \tag{3}$$

The term $a\|w\|^2$ gives the learner a start on expert w, according to the expert's "complexity" $\|w\|^2$.

Several prediction strategies for Learner have been considered. In [5], Kivinen and Warmuth gave relative loss bounds for the Exponentiated Gradient Algorithm. The ridge regression strategy was compared to a similar strategy, called AA^{R}, in [6] by Vovk. The best known upper bound on the relative loss holds for AA^{R}.

How can we get good strategies for Learner? It is not clear what the best prediction for Learner at trial t is. This is mainly because, when Learner gives $\hat{y}_t$, it does not know the number of trials T and it does not know anything about the future instances $x_{t+1}, \ldots, x_T$ and the future outcomes $y_t, \ldots, y_T$. But if it is known that there is some $Y \geq 0$ such that $y_T \in [-Y, Y]$, there is a prediction $\hat{y}_T$ for the last trial T that minimizes (3). This prediction is calculated in Theorem 1, it is essentially the prediction of AA^{R}.

If Learner makes in each trial t the prediction that would be optimal if $y_t \in [-Y, Y]$ is known and if trial t was known to be the last one of the sequence, there is an upper bound on (3) of the form $O(\ln T)$. This is proven in Theorem 3, which is a reproof of Theorem 1, [6]. Our new proof is shorter and simpler than Vovk's, which uses the Aggregating Algorithm (AA), "perfectly mixable" games

and contains a lot of difficult calculations. An independent proof of Theorem 3 is given by Azoury and Warmuth in [1]. They discuss relative loss bounds in the context of density estimation using the exponential family of distributions.

It is possible to find other new on-line prediction strategies with the method of Theorem 1. As an example, in Sect. 3 we look at the case of generalized linear regression. Here the linear experts are replaced by neurons, i.e., we compare to functions $\mathbb{R}^n \ni x \mapsto \varphi(w \cdot x)$, $w \in \mathbb{R}^n$, where $\varphi : \mathbb{R} \to \mathbb{R}$ is a fixed differentiable and strictly increasing function. The matching loss is used to measure the loss of Learner and the losses of the experts (there is a motivation for using the matching loss in [4]). Under the assumption that the outcome variable is bounded, we can calculate the optimal prediction of Learner for the last trial of a sequence in this more general case, too.

2 Linear Regression with Square Loss

In this section we calculate the best prediction for Learner in the last trial when the outcome variable is bounded and we prove an upper bound on the relative loss for linear regression with square loss.

For a sequence $(x_1, y_1), \ldots, (x_t, y_t)$ of instances and outcomes it is easy to give a formula for the loss of the "best" linear expert. We use the following notation:

$$b_t := \sum_{s=1}^{t} y_s x_s \in \mathbb{R}^n \ , \tag{4}$$

$$A_t := aI + \sum_{s=1}^{t} x_s x_s^\top \in \mathbb{R}^{n \times n} \ . \tag{5}$$

Note that for $t = 0, 1, 2, \ldots$, A_t is a positive definite symmetric $n \times n$–matrix because aI is positive definite and $xx^\top$ is a positive semidefinite matrix for all $x \in \mathbb{R}^n$. (For each $y \in \mathbb{R}^n$: $y^\top (xx^\top) y = (x \cdot y)^2 \geq 0$.)

Lemma 1. *For all $t \geq 0$, function $f(w) := a\|w\|^2 + L_t(w)$ is minimal at a unique point, say w_t. Furthermore, w_t and $f(w_t)$ are given by*

$$w_t = A_t^{-1} b_t \quad \textit{and} \quad f(w_t) = \sum_{s=1}^{t} y_s^2 - b_t^\top A_t^{-1} b_t \ .$$

Proof. From

$$f(w) = a\|w\|^2 + L_t(w) \overset{(2)}{=} a\|w\|^2 + \sum_{s=1}^{t} (y_s - w \cdot x_s)^2$$

$$= aw^\top w + \sum_{s=1}^{t} \left(y_s^2 - 2y_s (w \cdot x_s) + (w^\top x_s)(x_s^\top w) \right)$$

$$= \sum_{s=1}^{t} y_s^2 - 2\sum_{s=1}^{t} w \cdot (y_s x_s) + w^\top \left(aI + \sum_{s=1}^{t} x_s x_s^\top \right) w$$
$$\overset{(4),(5)}{=} \sum_{s=1}^{t} y_s^2 - 2w \cdot b_t + w^\top A_t w \ .$$

it follows that $\nabla f(w) = 2A_t w - 2b_t$, $Hf(w) = 2A_t$. Thus f is convex and it is minimal if $\nabla f(w) = 0$, i.e. for $w = A_t^{-1} b_t$. This shows that $w_t = A_t^{-1} b_t$ and we obtain

$$f(w_t) = f(A_t^{-1} b_t) = \sum_{s=1}^{t} y_s^2 - 2b_t^\top A_t^{-1} b_t + b_t^\top A_t^{-1} A_t A_t^{-1} b_t = \sum_{s=1}^{t} y_s^2 - b_t^\top A_t^{-1} b_t \ .$$

□

In [6], Vovk proposes the AA^{R} learning algorithm which makes the predictions

$$\hat{y}_t = b_{t-1}^\top A_t^{-1} x_t$$

and shows how these predictions can be computed in time $O(n^2)$ per trial. They are very similar to $w_t \cdot x_t = b_t^\top A_t^{-1} x_t$, the prediction of the expert w_t on the instance x_t. We will now show that $b_{T-1}^\top A_T^{-1} x_T$ is essentially the best prediction for the last trial T. For this, let

$$\mathrm{clip}(z, Z) := \begin{cases} -Z, & z \in (-\infty, -Z] \ , \\ z, & z \in [-Z, Z] \ , \\ Z, & z \in [Z, \infty) \ , \end{cases}$$

for $z, Z \in \mathbb{R}$, $Z \geq 0$. $\mathrm{clip}(z, Z)$ is the number in $[-Z, Z]$ that is closest to z.

Theorem 1. *If Learner knows that $y_T \in [-Y, Y]$, then the optimal prediction for the last trial T is*

$$\hat{y}_T = \mathrm{clip}(b_{T-1}^\top A_T^{-1} x_T, Y) \ .$$

Proof. Any $y_T \in [-Y, Y]$ can be chosen by Nature. Thus Learner should choose a $\hat{y}_T \in \mathbb{R}$ such that

$$\sup_{y_T \in [-Y,Y]} \left(L_T(\text{Learner}) - \inf_{w \in \mathbb{R}^n} \left(a\|w\|^2 + L_T(w) \right) \right)$$
$$\overset{(1),\ \text{Lemma 1}}{=} \sup_{y_T \in [-Y,Y]} \left(\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 - \sum_{t=1}^{T} y_t^2 + b_T^\top A_T^{-1} b_T \right).$$

is minimal. Because of

$$b_T^\top A_T^{-1} b_T \overset{(4)}{=} (b_{T-1} + y_T x_T)^\top A_T^{-1} (b_{T-1} + y_T x_T)$$
$$= b_{T-1}^\top A_T^{-1} b_{T-1} + 2y_T b_{T-1}^\top A_T^{-1} x_T + y_T^2 x_T^\top A_T^{-1} x_T$$

this expression is minimal if and only if (only terms that depend on y_T or on $\hat{y}_T$ are important)

$$\begin{aligned}&\sup_{y_T\in[-Y,Y]}\left(-2y_T\hat{y}_T+\hat{y}_T^2+2y_Tb_{T-1}^\top A_T^{-1}x_T+y_T^2x_T^\top A_T^{-1}x_T\right)\\&=\sup_{y_T\in[-Y,Y]}\left(y_T^2(x_T^\top A_T^{-1}x_T)+2y_T(b_{T-1}^\top A_T^{-1}x_T-\hat{y}_T)+\hat{y}_T^2\right)\end{aligned}$$

is minimal. Since $x_T^\top A_T^{-1}x_T \geq 0$ (A_T^{-1} is positive definite), we only have to consider $y_T=-Y$ and $y_T=Y$. Thus we have to find a $\hat{y}_T$ for which

$$2Y|b_{T-1}^\top A_T^{-1}x_T-\hat{y}_T|+\hat{y}_T^2 \tag{6}$$

is minimal. For $\hat{y}_T \leq b_{T-1}^\top A_T^{-1}x_T$, (6) equals $2Y(b_{T-1}^\top A_T^{-1}x_T-\hat{y}_T)+\hat{y}_T^2$ and this is minimal at $\min(Y, b_{T-1}^\top A_T^{-1}x_T)$ on this domain. For $\hat{y}_T \geq b_{T-1}^\top A_T^{-1}x_T$, (6) equals $2Y(\hat{y}_T-b_{T-1}^\top A_T^{-1}x_T)+\hat{y}_T^2$ and this is minimal at $\max(-Y, b_{T-1}^\top A_T^{-1}x_T)$ on this domain. If $b_{T-1}^\top A_T^{-1}x_T \in [-Y,Y]$ this obviously implies the assertion. If $b_{T-1}^\top A_T^{-1}x_T \geq Y$, (6) is decreasing for $\hat{y}_T \leq Y$ and increasing for $\hat{y}_T \geq Y$. Thus (6) is minimal at Y. The case $b_{T-1}^\top A_T^{-1}x_T \leq -Y$ is similar. □

Theorem 1 does not show that $\hat{y}_t=\mathrm{clip}(b_{t-1}^\top A_t^{-1}x_t, Y)$, $t=1,\ldots,T$, are the optimal predictions over a sequence of T trials when it is known that $y_1,\ldots,y_T$ are bounded by Y, because the "best" expert on the first t trials might differ from the "best" expert on all T trials. But there are very good relative loss bounds for these predictions as we will see in Theorem 3.

A standard application of Theorem 1 results, when the learner knows in advance a global upper bound $Y \geq 0$ on all potential outcomes y_t (and thus in particular on the last outcome y_T). The next result shows that a learner without prior knowledge of such a global upper bound Y can use a sort of "empirical upper bound" without suffering much extra loss.

Theorem 2. *Let*

$$Y_0 := 0 \ , \qquad Y_t := \max_{s=1}^{t}|y_s| \ , \qquad t=1,2,3,\ldots \ .$$

For all $p_t=p_t(x_1,\ldots,x_t,y_1,\ldots,y_{t-1}) \in \mathbb{R}$, $t=1,2,\ldots,T$, *with the predictions*

$$\hat{y}_t=\mathrm{clip}(p_t, Y_{t-1})$$

the loss of Learner on a sequence of T *trials is at most by* Y_T^2 *larger than with the predictions* $\hat{y}_t=\mathrm{clip}(p_t,Y)$ *for each* $Y \geq Y_{T-1}$.

Proof. We show that for all $t \in \{1,\ldots,T\}$:

$$(y_t-\mathrm{clip}(p_t,Y_{t-1}))^2 \leq (y_t-\mathrm{clip}(p_t,Y))^2+(Y_t-Y_{t-1})^2 \ . \tag{7}$$

From this it would follow that the additional loss is at most

$$\sum_{t=1}^{T}(Y_t-Y_{t-1})^2 \leq \left(\sum_{t=1}^{T}(Y_t-Y_{t-1})\right)^2 = Y_T^2 \ .$$

Let $t \in \{1, \ldots, T\}$. If $|y_t| \leq Y_{t-1}$, then p_t clipped by Y_{t-1} is closer to y_t than p_t clipped by $Y \geq Y_{t-1}$. Thus

$$(y_t - \mathrm{clip}(p_t, Y_{t-1}))^2 \leq (y_t - \mathrm{clip}(p_t, Y))^2 \ .$$

So we can assume that $|y_t| > Y_{t-1}$. Then $|y_t| = Y_t$. If $\mathrm{clip}(p_t, Y_{t-1}) = \mathrm{clip}(p_t, Y)$, (7) is obvious. Assume that $\mathrm{clip}(p_t, Y_{t-1}) \neq \mathrm{clip}(p_t, Y)$. Thus $|\mathrm{clip}(p_t, Y_{t-1})| = Y_{t-1}$. If y_t and p_t have the same sign then

$$(y_t - \mathrm{clip}(p_t, Y_{t-1}))^2 = (Y_t - Y_{t-1})^2 \ .$$

Otherwise

$$\begin{aligned}(y_t - \mathrm{clip}(p_t, Y_{t-1}))^2 &= (Y_t + |\mathrm{clip}(p_t, Y_{t-1})|)^2 \\ &\leq (Y_t + |\mathrm{clip}(p_t, Y)|)^2 = (y_t - \mathrm{clip}(p_t, Y))^2 \ .\end{aligned}$$

□

Even without clipping the $b_{t-1}^\top A_t^{-1} x_t$ (not clipping them means that the loss of Learner $L_T(\text{Learner})$ will only increase, but has the advantage that Learner does not need to know Y), there is a very good upper bound on the relative loss. To show this, we need the following lemma:

Lemma 2. *For all $t \geq 1$:*

$$A_{t-1}^{-1} - A_t^{-1} - A_t^{-1} x_t x_t^\top A_t^{-1} = (x_t^\top A_{t-1}^{-1} x_t) A_t^{-1} x_t x_t^\top A_t^{-1} \ .$$

Proof. From the equality $A_t - A_{t-1} \overset{(5)}{=} x_t x_t^\top$ we get

$$A_{t-1}^{-1} - A_t^{-1} = A_{t-1}^{-1} x_t x_t^\top A_t^{-1} \ , \tag{8}$$

$$A_{t-1}^{-1} - A_t^{-1} = A_t^{-1} x_t x_t^\top A_{t-1}^{-1} \ . \tag{9}$$

Thus

$$\begin{aligned}&A_{t-1}^{-1} - A_t^{-1} - A_t^{-1} x_t x_t^\top A_t^{-1} \\ &\quad \overset{(8)}{=} (A_{t-1}^{-1} - A_t^{-1}) x_t x_t^\top A_t^{-1} \\ &\quad \overset{(9)}{=} A_t^{-1} x_t x_t^\top A_{t-1}^{-1} x_t x_t^\top A_t^{-1} \ .\end{aligned}$$

□

Theorem 3 ([6], Theorem 1). *If Learner predicts with $\hat{y}_t = b_{t-1}^\top A_t^{-1} x_t$ for $1 \leq t \leq T$ and if the outcome variables $y_1, \ldots, y_T$ are bounded by $Y \geq 0$, then*

$$L_T(\text{Learner}) \leq \inf_{w \in \mathbb{R}^n} \left(a\|w\|^2 + L_T(w) \right) + Y^2 \ln \left| \frac{1}{a} A_T \right| \ .$$

Proof. For $1 \le t \le T$ we have

$$\begin{aligned}
&(y_t - \hat{y}_t)^2 + \inf_{w \in \mathbb{R}^n} \Big(a\|w\|^2 + L_{t-1}(w) \Big) - \inf_{w \in \mathbb{R}^n} \Big(a\|w\|^2 + L_t(w) \Big) \\
&\quad \overset{\text{Lemma 1}}{=} -2y_t\hat{y}_t + \hat{y}_t^2 - b_{t-1}^\top A_{t-1}^{-1} b_{t-1} + b_t^\top A_t^{-1} b_t \\
&\quad \overset{(4)}{=} -2y_t b_{t-1}^\top A_t^{-1} x_t + b_{t-1}^\top A_t^{-1} x_t x_t^\top A_t^{-1} b_{t-1} - b_{t-1}^\top A_{t-1}^{-1} b_{t-1} \\
&\qquad\qquad + (b_{t-1} + y_t x_t)^\top A_t^{-1} (b_{t-1} + y_t x_t) \\
&\quad = b_{t-1}^\top (A_t^{-1} x_t x_t^\top A_t^{-1} - A_{t-1}^{-1} + A_t^{-1}) b_{t-1} + y_t^2 x_t^\top A_t^{-1} x_t \\
&\quad \overset{\text{Lemma 2}}{=} y_t^2 x_t^\top A_t^{-1} x_t - (x_t^\top A_{t-1}^{-1} x_t) b_{t-1}^\top A_t^{-1} x_t x_t^\top A_t^{-1} b_{t-1} \\
&\quad = y_t^2 x_t^\top A_t^{-1} x_t - (x_t^\top A_{t-1}^{-1} x_t) \hat{y}_t^2 \\
&\quad \le Y^2 x_t A_t^{-1} x_t \ .
\end{aligned}$$

Summing over $t \in \{1, \ldots, T\}$ and using (1) gives

$$L_T(\text{Learner}) - \inf_{w \in \mathbb{R}^n} \Big(a\|w\|^2 + L_T(w) \Big) \le Y^2 \sum_{t=1}^T x_t A_t^{-1} x_t \ .$$

Because of $\ln |\frac{1}{a} A_0| \overset{(5)}{=} 0$, it now suffices to show that for $t \in \{1, \ldots, T\}$:

$$x_t^\top A_t^{-1} x_t \le \ln \frac{|A_t|}{|A_{t-1}|} \ .$$

We first show that $x_t^\top A_t^{-1} x_t < 1$. This is trivial for $x_t = 0$. For $x_t \neq 0$, it is obtained from the following calculation:

$$\begin{aligned}
(x_t^\top A_t^{-1} x_t)^2 &= x_t^\top A_t^{-1} x_t x_t^\top A_t^{-1} x_t \overset{(5)}{=} x_t^\top A_t^{-1} (A_t - A_{t-1}) A_t^{-1} x_t \\
&= x_t^\top A_t^{-1} x_t - \underbrace{x_t^\top A_t^{-1} A_{t-1} A_t^{-1} x_t}_{>0} < x_t^\top A_t^{-1} x_t \ .
\end{aligned}$$

There is a symmetric, positive definite matrix $A \in \mathbb{R}^{n \times n}$ such that $A_t = AA$. Let $\xi := A^{-1} x_t$. Thus $x_t = A\xi$. From $\xi^\top \xi = x_t^\top A_t^{-1} x_t < 1$ we know that $I - \xi\xi^\top$ is positive definite and we get

$$|I - \xi\xi^\top| \le \prod_{i=1}^n (1 - \xi_i^2) \le \prod_{i=1}^n e^{-\xi_i^2} = e^{-\xi^\top \xi} \ , \tag{10}$$

where the first inequality holds because the determinant of a positive semidefinite matrix is bounded by the product of the entries on the diagonal of the matrix (e.g., see [2], Theorem 7 in Chapter 2). It follows that

$$x_t^\top A_t^{-1} x_t = \xi^\top \xi \overset{(10)}{\le} \ln \frac{1}{|I - \xi\xi^\top|} = \ln \frac{|AA|}{|AA - A\xi\xi^\top A|} \overset{(5)}{=} \ln \left| \frac{A_t}{A_{t-1}} \right| \ .$$

□

Vovk gives the following upper bounds on the term $\ln\left|\frac{1}{a}A_T\right|$ in [6]:

$$\ln\left|\frac{1}{a}A_T\right| \overset{(5)}{=} \ln\left|I + \frac{1}{a}\sum_{t=1}^{T} x_t x_t^\top\right|$$
$$\leq \sum_{i=1}^{n} \ln\left(1 + \frac{1}{a}\sum_{t=1}^{T} x_{t,i}^2\right) \leq n \ln\left(1 + \frac{TX^2}{a}\right) \quad ,$$

where, for the first inequality, again [2], Chap. 2, Theorem 7 is used, and where we assume that $|x_{t,i}| \leq X$ for $t \in \{1, \ldots, T\}$, $i \in \{1, \ldots, n\}$.

3 Generalized Linear Regression

In generalized linear regression we consider the same protocol of interaction between Nature and Learner as before, but now expert $w \in \mathbb{R}^n$ makes the prediction $\varphi(w \cdot x)$ on an instance $x \in \mathbb{R}^n$, where $\varphi : \mathbb{R} \to \mathbb{R}$ is a strictly increasing, differentiable function. Now the losses are measured with the matching loss L_φ,

$$L_\varphi(y, \hat{y}) := \int_{\varphi^{-1}(y)}^{\varphi^{-1}(\hat{y})} (\varphi(\tau) - y)\, d\tau \tag{11}$$
$$= \int_{\varphi^{-1}(y)}^{\varphi^{-1}(\hat{y})} \varphi(\tau)\, d\tau + y\varphi^{-1}(y) - y\varphi^{-1}(\hat{y}) \tag{12}$$

for $y, \hat{y}$ in the range of φ (see Fig. 1). Examples of matching loss functions are the square loss ($\varphi = \mathrm{id}_{\mathbb{R}}$, $L_\varphi(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$) and the entropic loss ($\varphi(z) = \frac{1}{1+e^{-z}}$, $L_\varphi(y, \hat{y}) = y \ln \frac{y}{\hat{y}} + (1 - y) \ln \frac{1-y}{1-\hat{y}}$). Note that in the case of square loss ($\varphi = \mathrm{id}_{\mathbb{R}}$) all losses in this section differ by a factor of $\frac{1}{2}$ from those in Sect. 2. This was not harmonized because in Sect. 2 we wanted to use the same definitions as in [6], and in Sect. 3 we want to use the common definition of the matching loss.

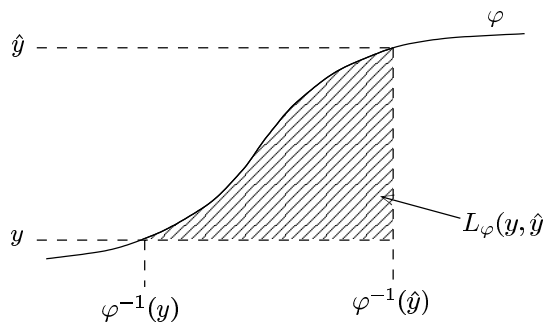

Fig. 1. The matching loss function L_φ

The loss of Learner and the loss of expert $w \in \mathbb{R}^n$ are now defined as

$$L_{\varphi,t}(\text{Learner}) = \sum_{s=1}^{t} L_\varphi(y_s, \hat{y}_s) \ , \qquad L_{\varphi,t}(w) = \sum_{s=1}^{t} L_\varphi(y_s, \varphi(w \cdot x_s)) \ .$$

Lemma 3. *For all $t \geq 0$, function $f(w) := \frac{a}{2}\|w\|^2 + L_{\varphi,t}(w)$ is minimal at a unique point, say w_t. w_t is differentiable in y_t and we have*

$$\frac{\partial w_t}{\partial y_t} \cdot x_t \geq 0 \ . \tag{13}$$

Let

$$\Phi(z) := -\frac{1}{2} z\varphi(z) + \int_0^z \varphi(\tau)\, d\tau \ , \qquad z \in \mathbb{R} \ ,$$

$$\Gamma_t := \sum_{s=1}^{t} \left(\frac{y_s}{2} w_t \cdot x_s - \Phi(w_t \cdot x_s) \right) \ .$$

Then

$$\begin{aligned} L_{\varphi,t}(\text{Learner}) &- \inf_{w \in \mathbb{R}^n} \left(\frac{a}{2}\|w\|^2 + L_{\varphi,t}(w) \right) \\ &= \sum_{s=1}^{t} \left(\int_0^{\varphi^{-1}(\hat{y}_s)} \varphi(\tau)\, d\tau - y_s \varphi^{-1}(\hat{y}_s) \right) + \Gamma_t \end{aligned} \tag{14}$$

and

$$\frac{\partial \Gamma_t}{\partial y_t} = w_t \cdot x_t \ . \tag{15}$$

The proof of Lemma 3 is omitted here. We need one more small Lemma.

Lemma 4. *Let $f, g : \mathbb{R} \to \mathbb{R}$ be differentiable, convex functions with $f'(z) > g'(z)$ for all $z \in \mathbb{R}$. If there are numbers $Y^- \leq Y^+$, $Z \in \mathbb{R}$ such that f is minimal at Y^-, g is minimal at Y^+ and $f(Z) = g(Z)$, then $h := \max(f, g)$ is minimal at*

$$\text{clip}(Z, Y^-, Y^+) := \begin{cases} Y^- \ , & Z \in (-\infty, Y^-] \ , \\ Z \ , & Z \in [Y^-, Y^+] \ , \\ Y^+ \ , & Z \in [Y^+, \infty) \ . \end{cases}$$

Proof. Because of $f'(z) > g'(z)$ for $z \in \mathbb{R}$ and because of $f(Z) = g(Z)$, we have

$$h(z) = \begin{cases} g(z) \ , & z \leq Z \ , \\ f(z) \ , & z \geq Z \ . \end{cases}$$

If $Z \in [Y^-, Y^+]$, then h is decreasing on $(-\infty, Z]$ (because g is decreasing there) and h is increasing on $[Z, \infty)$ (because f is increasing there). Thus h is minimal at Z. If $Z < Y^-$ then h is decreasing on $(-\infty, Y^-]$ (because both f and g are decreasing there) and h is increasing on $[Y^-, \infty)$ (because f is increasing there). Thus h is minimal at Y^-. The case $Z > Y^+$ is very similar to the case $Z < Y^-$. $\square$

Now we are ready to calculate the best prediction $\hat{y}_T$ for the last trial T when it is known that $y_T \in [Y^-, Y^+]$.

w_T and Γ_T are functions in $x_1, \ldots, x_T, y_1, \ldots, y_T$. If only $x_1, \ldots, x_T$ and $y_1, \ldots, y_{T-1}$ are known (which is the case when Learner makes the last prediction $\hat{y}_T$), we write $w_T = w_T(y_T)$, $\Gamma_T = \Gamma_T(y_T)$.

Theorem 4. *Let $Y^- < Y^+$ be in the range of φ. As a function of $\hat{y}_T$,*

$$\sup_{y_T \in [Y^-, Y^+]} \left(L_{\varphi,T}(\text{Learner}) - \inf_{w \in \mathbb{R}^n} \left(\frac{a}{2} \|w\|^2 + L_{\varphi,T}(w) \right) \right)$$

is minimal for

$$\hat{y}_T = \text{clip}\left(\varphi \left(\frac{1}{Y^+ - Y^-} \int_{Y^-}^{Y^+} w_T(y_T) \cdot x_T \, dy_T \right), Y^-, Y^+ \right) \quad .$$

Proof. From (14) and (15) it follows that

$$\begin{aligned} &\frac{\partial}{\partial y_T} \left(L_{\varphi,T}(\text{Learner}) - \inf_{w \in \mathbb{R}^n} \left(\frac{a}{2} \|w\|^2 + L_{\varphi,T}(w) \right) \right) \\ &\qquad \overset{(15)}{=} w_T \cdot x_T - \varphi^{-1}(\hat{y}_T) \ , \\ &\frac{\partial}{\partial y_T} \frac{\partial}{\partial y_T} \left(L_{\varphi,T}(\text{Learner}) - \inf_{w \in \mathbb{R}^n} \left(\frac{a}{2} \|w\|^2 + L_{\varphi,T}(w) \right) \right) = \frac{\partial w_T}{\partial y_T} \cdot x_T \overset{(13)}{\geq} 0 \ . \end{aligned}$$

Thus, for fixed $\hat{y}_T$, (14) (with $t = T$) is maximal for $y_T = Y^-$ or for $y_T = Y^+$. So we have to find a $\hat{y}_T$ such that

$$\begin{aligned} \int_0^{\varphi^{-1}(\hat{y}_T)} \varphi(\tau) \, d\tau + \max\Big(-Y^- \varphi^{-1}(\hat{y}_T) + \Gamma_T(Y^-), -Y^+ \varphi^{-1}(\hat{y}_T) + \Gamma_T(Y^+) \Big) \\ = \max\Big(f(\varphi^{-1}(\hat{y}_T)), g(\varphi^{-1}(\hat{y}_T)) \Big) \end{aligned}$$

with

$$\begin{aligned} f(z) &= \int_0^z \varphi(\tau) \, d\tau - Y^- z + \Gamma_T(Y^-), \qquad & f'(z) &= \varphi(z) - Y^- \ , \\ g(z) &= \int_0^z \varphi(\tau) \, d\tau - Y^+ z + \Gamma_T(Y^+) \ , \qquad & g'(z) &= \varphi(z) - Y^+ \ , \end{aligned}$$

is minimal. f and g are equal if $-Y^- \varphi^{-1}(\hat{y}_T) + \Gamma_T(Y^-) = -Y^+ \varphi^{-1}(\hat{y}_T) + \Gamma_T(Y^+)$, i.e., if

$$\begin{aligned} \varphi^{-1}(\hat{y}_T) &= \frac{\Gamma_T(Y^+) - \Gamma_T(Y^-)}{Y^+ - Y^-} = \frac{1}{Y^+ - Y^-} \int_{Y^-}^{Y^+} \frac{\partial \Gamma_T}{\partial y_T} \, dy_T \\ &\overset{(15)}{=} \frac{1}{Y^+ - Y^-} \int_{Y^-}^{Y^+} w_T \cdot x_T \, dy_T \ . \end{aligned}$$

Because of Lemma 4 (f is minimal at $\varphi^{-1}(Y^-)$, g at $\varphi^{-1}(Y^+)$) this implies the assertion. □

Like in Theorem 2 it might be a good idea to use the predictions

$$\hat{y}_t = \mathrm{clip}\left(\varphi\left(\frac{1}{Y_{t-1}^+ - Y_{t-1}^-}\int_{Y_{t-1}^-}^{Y_{t-1}^+} w_t \cdot x_t \, dy_t\right), Y_{t-1}^-, Y_{t-1}^+\right)$$

where $Y_t^+ = \max_{s=1}^t y_s$, $Y_t^- = \min_{s=1}^t y_s$ when no bounds Y^-, Y^+ on the outcomes y_t are known.

4 Conclusion

We have shown that Vovk's prediction rule is essentially characterized by the property that it minimizes the maximal extra loss (compared to the best off-line expert) that might be suffered in the last trial (Theorem 1). For the sake of simplicity, let's call this the minmax-property. Note that instead of first inventing the learning rule and then proving Theorem 1, one could have gone the other direction. The calculus applied in the proof of Theorem 1 will then lead automatically to Vovk's rule. This is precisely the line of attack that we pursued in Section 3. There, we considered the (much more involved) generalized regression problem. In order to find a good *explicit* candidate prediction rule, we tried to find the rule *implicitly* given by the minmax-property. This finally has lead to the rule given in Theorem 4.

It is straightforward to ask whether this rule has (provably) good relative loss bounds. In other words, we need the analogue of Theorem 3 for the generalized regression problem. As yet, we can show that

$$\begin{aligned} L_\varphi(y_t, \hat{y}_t) + \inf_{w \in \mathbb{R}^n}\left(\frac{a}{2}\|w\|^2 + L_{\varphi,t-1}(w)\right) - \inf_{w \in \mathbb{R}^n}\left(\frac{a}{2}\|w\|^2 + L_{\varphi,t}(w)\right) \\ \leq \int_0^{\frac{\Gamma_t(Y^+)-\Gamma_t(Y^-)}{Y^+-Y^-}} \varphi(\tau)\, d\tau + \frac{Y^+\Gamma_t(Y^-) - Y^-\Gamma_t(Y^+)}{Y^+ - Y^-} - \Gamma_{t-1} \end{aligned} \tag{16}$$

for all t. By summing over $t \in \{1, \ldots, T\}$ we get a bound on the relative loss over the sequence of T trials. Thus upper bounds on the terms on the right hand side of (16) would be very interesting.

5 Acknowledgements

The author is grateful to Hans Ulrich Simon and Manfred Warmuth for a lot of helpful suggestions.

References

1. Azoury, K., Warmuth, M.: Relative Loss Bounds for On-line Density Estimation with the Exponential Family of Distributions, to appear at the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99.
2. Beckenbach, E. F., Bellman, R.: Inequalities, Berlin: Springer, 1965.
3. Foster, D. P.: Prediction in the worst case, Annals of Statistics 19, 1084–1090.
4. Kivinen, J., Warmuth, M.: Relative Loss Bounds for Multidimensional Regression Problems. In Jordan, M., Kearns, M., Solla, S., editors, Advances in Neural Information Processing Systems 10 (NIPS 97), 287–293, MIT Press, Cambridge, MA, 1998.
5. Kivinen, J., Warmuth, M.: Additive versus exponentiated gradient updates for linear prediction, Information and Computation 132:1–64, 1997.
6. Vovk, V.: Competitive On-Line Linear Regression. Technical Report CSD-TR-97-13, Department of Computer Science, Royal Holloway, University of London, 1997.