

Relative loss bounds for predicting almost as well as any function in a union of Gaussian reproducing kernel spaces with varying widths^{*}

Mark Herbster

Department of Computer Science
University College London
Gower Street, London WC1E 6BT, UK
M.Herbster@cs.ucl.ac.uk

Abstract. We consider a two-layer network algorithm. The first layer consists of an uncountable number of linear units. Each linear unit is an *LMS* algorithm whose inputs are first “kernelized.” Each unit is indexed by the value of a parameter corresponding to a parameterized reproducing kernel, here an isotropic Gaussian Kernel parameterized by its width. The first-layer outputs are then connected to an *Exponential Weights* algorithm which combines them to produce the final output. We give performance guarantees for this algorithm.

As a guarantee of performance, we give a *relative loss bound* for this online algorithm. By online, we refer to the fact that learning proceeds in trials where on each trial the algorithm first receives a pattern, then it makes a prediction, after which it receives the true outcome, and finally incurs a loss on that trial measuring the discrepancy between its prediction and the true outcome. By relative loss bound, we refer to the fact on any trial, we can bound the cumulative loss of the algorithm by the cumulative loss of any predictor in a comparison class of predictors plus an additive term. Hence the goal is that the performance of algorithm be almost as good as any predictor in the class; therefore we desire a small additive term. Often these bounds may be given without any probabilistic assumptions. In this note the comparison class is the set of functions obtained by a union of reproducing kernel spaces formed by isotropic Gaussian kernels of varying widths.

1 Introduction

The key contribution of this note is a lemma (cf. Lemma 1) and its application to bounding the online loss of a particular algorithm (K-LMS-NET) on an adversarially chosen data sequence in terms of the loss any predictor chosen a posteriori from a union of reproducing kernel spaces of Gaussian kernels of varying width. In the lemma we consider a function chosen from a Gaussian reproducing kernel space of particular width then its norm and its squared loss on an arbitrary data sequence are used to bound the norm and squared loss on the same data sequence, of a function with essentially the same representation in a Gaussian kernel space of slightly differing width.

In [3] a full introduction is given to the K-LMS-NET algorithm with an added additional application to the convex combination of two kernels. Here, we consider only a bound for the case of Gaussian kernel spaces. A major consideration in that paper is a demonstration that predictions for specific kernels could be well approximated in polynomial time; in this note we do not discuss the polynomial tractability or lack thereof of for prediction. In the following sections we give a minimal excerpt of the results of [3] to motivate our final section where we present results for Gaussian kernel spaces.

1.1 Preliminaries and notation

The symbol \mathcal{X} denotes an abstract space, for example \mathcal{X} could be a set of strings. A Hilbert space \mathcal{H} denotes a complete inner product space. The inner product between vectors \mathbf{v} and \mathbf{w} in \mathcal{H} is denoted by $\langle \mathbf{v}, \mathbf{w} \rangle$ and the norm by $\|\mathbf{v}\|$. In this note, we will consider Hilbert spaces determined

^{*} This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors’ views.

by a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. The prehilbert space induced by kernel k is the set $H_k = \text{span}(\{k(x, \cdot)\}_{\forall x \in \mathcal{X}})$ and the inner product of $f = \sum_{i=1}^m \beta_i k(x_i, \cdot)$ and $g = \sum_{j=1}^n \beta'_j k(x'_j, \cdot)$ is $\langle f, g \rangle = \sum_{i=1}^m \sum_{j=1}^n \beta_i \beta'_j k(x_i, x'_j)$. The completion of H_k is denoted \mathcal{H}_k . Two kernels $k_0 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and $k_1 : \mathcal{X}' \times \mathcal{X}' \rightarrow \mathbb{R}$ are termed *domain compatible* if $\mathcal{X} = \mathcal{X}'$. Further useful properties of reproducing kernels (including conditions on a function k to be a kernel), and introductory material may be found in [2]. In this note we are particularly interested in the parameterized kernels $k_\alpha(\mathbf{x}, \mathbf{x}') = e^{-s_0 \|\mathbf{x} - \mathbf{x}'\|^2}$ with an associated Hilbert Space \mathcal{H}_α , inner product $\langle \cdot, \cdot \rangle_\alpha$, and norm $\|\cdot\|_\alpha$, for every $\alpha \in [0, 1]$. We denote the Lebesgue measure of a set A by $\mu(A)$.

We consider the following on-line learning model based on a model introduced by Littlestone [5, 6]. Learning proceeds in trials $t = 1, 2, \dots, \ell$. In each trial t the algorithm receives a *pattern* x_t . It then gives a prediction denoted \hat{y}_t . The algorithm then receives an *outcome* y_t , and incurs a loss $L(y_t, \hat{y}_t)$ measuring the discrepancy between y_t and \hat{y}_t . In this note x_t is usually in the structureless \mathcal{X} ; when \mathcal{X} is a Hilbert space then \mathbf{x}_t will be in bold. Predictions y_t and outcomes \hat{y}_t are always in \mathbb{R} . In this note $L(y_t, \hat{y}_t) = (y_t - \hat{y}_t)^2$.

In the usual methodology of relative loss bounds the total loss of the algorithm is expressed as a function of the total loss of any member $c : \mathcal{X} \rightarrow \mathbb{R}$ of a comparison class \mathcal{C} of predictors [5]. Surprisingly, such bounds are achievable even when there are no probabilistic assumptions made on the sequence of examples. These bounds are of the following form, for all data sequences $S = \langle (x_1, y_1), (x_2, y_2), \dots, (x_\ell, y_\ell) \rangle$,

$$\sum_{t=1}^{\ell} L(y_t, \hat{y}_t) \leq \sum_{t=1}^{\ell} L(y_t, c(x_t)) + O(r(S, \mathcal{C}, c)) \quad \forall c \in \mathcal{C}$$

where $r(S, \mathcal{C}, c)$ is known as the *regret*, since it measures our “regret” at using our algorithm versus the “best” predictor c in the comparison class. In the ideal case the regret is a slowly growing function of the data sequence, the comparison class, and the particular predictor.

2 Predicting well relative to a function in a union of Gaussians kernel spaces of varying width with the K-LMS-NET algorithm

In this section we prove a relative loss bound for the K-LMS-NET algorithm with Gaussian kernels. It is not obvious that it is tractable polynomially or otherwise to calculate (well-approximate) the predictions (cf (2)) of the algorithm. In [3] we show that it possible to give predictions of guaranteed quality for the width parameterized Gaussian kernel in polynomial time when the patterns are drawn from a boolean domain; that result is easily extended to an arbitrarily discretized pattern domain, where though polynomial, the algorithm must “pay” computationally in the degree of discretization. Here we are not concerned with that computational issue hence the data may be arbitrary real vectors. The following general bound for the K-LMS-NET algorithm is given in [3]. The bound is a straightforward chaining of the well known loss bounds [1] of the LMS (GD) algorithm and a variant of the *exponential weights* algorithm [7, 6, 4] that implements direct clipping of the inputs and an amortized clipping of the cumulative loss.

Theorem 1. *The K-LMS-NET algorithm with parameterized kernel function k_α ($\alpha \in [0, 1]$) with any data sequence $\langle (x_1, y_1), (x_2, y_2), \dots, (x_\ell, y_\ell) \rangle \in (\mathcal{X}, [r_1, r_2])^\ell$ when the algorithm is tuned with constants r_1, r_2 , and η , the total square loss of the algorithm will satisfy*

$$\sum_{t=1}^{\ell} L(y_t, \hat{y}_t) \leq \sup_{\alpha \in \mathcal{A}} \left(\sum_{t=1}^{\ell} L(y_t, h_\alpha(x_t)) \right) + 2\sqrt{\hat{L}_{\mathcal{A}} \hat{H}_{\mathcal{A}} \hat{X}_{\mathcal{A}}} + \hat{H}_{\mathcal{A}}^2 \hat{X}_{\mathcal{A}}^2 + 2(r_2 - r_1)^2 \ln \frac{1}{\mu(\mathcal{A})} \quad (5)$$

for all measurable sets $\mathcal{A} \subseteq [0, 1]$ for which there exists a tuple of functions $(h_\alpha)_{\alpha \in \mathcal{A}} \in \prod_{\alpha \in \mathcal{A}} \mathcal{H}_\alpha$ and constants $\hat{L}_{\mathcal{A}}, \hat{H}_{\mathcal{A}}$, and, $\hat{X}_{\mathcal{A}}$, where for all $\alpha \in \mathcal{A}$ the following conditions must hold:

$$\sum_{t=1}^{\ell} L(y_t, h_\alpha(x_t)) \leq \hat{L}_{\mathcal{A}}, \quad \|h_\alpha\|_\alpha \leq \hat{H}_{\mathcal{A}}, \quad \text{and, } \forall t : k_\alpha(x_t, x_t) \leq \hat{X}_{\mathcal{A}}, \quad (6)$$

Parameters: \mathcal{X} : a pattern space;
 $k_\alpha : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$: a parameterized kernel function ($\alpha \in [0, 1]$);
 $\{\mathcal{H}_\alpha\}$: a set of Hilbert spaces induced by k_α ;
 η : a learning rate; $[r_1, r_2]$: an outcome range.

Data: An online sequence $\langle (x_1, y_1), (x_2, y_2), \dots, (x_\ell, y_\ell) \rangle \in (\mathcal{X}, [r_1, r_2])^\ell$.

Initialization: $r = (r_2 - r_1)$, $c = 2r^2$, $\mathbf{w}_{\alpha,1}^i(x) = \mathbf{0}$, $\mathbf{w}_1^{ii}(\alpha) = 1$,
 $\Phi^i(x) = \max(r_1, \min(r_2, x))$; $\Phi_t^{ii}(w) = \max(\exp(-\frac{t}{2}), w)$.

for $t = 1, \dots, \ell$ **do**

Predict: receive x_t ,

$$\begin{aligned} \hat{\mathbf{y}}_t^i(\alpha) &= \mathbf{w}_{\alpha,t}^i(x_t) = \langle \mathbf{w}_{\alpha,t}^i(\cdot), k_\alpha(x_t, \cdot) \rangle_{\mathcal{H}_\alpha} \\ &= \eta \sum_{j=1}^{t-1} (y_j - \hat{\mathbf{y}}_j^i(\alpha)) k_\alpha(x_j, x_t) \end{aligned} \quad (1)$$

$$\hat{y}_t = \frac{\int_0^1 \mathbf{w}_t^{ii}(\alpha) \Phi^i(\hat{\mathbf{y}}_t^i(\alpha)) d\alpha}{\int_0^1 \mathbf{w}_t^{ii}(\alpha) d\alpha} \quad (2)$$

Update: receive y_t ,

$$\mathbf{w}_{\alpha,t+1}^i(x) = \mathbf{w}_{\alpha,t}^i(x) + \eta(y_t - \hat{\mathbf{y}}_t^i(\alpha)) k_\alpha(x_t, x) \quad (3)$$

$$L_{[1,t]}(\alpha) = L_{[1,t-1]}(\alpha) + (y_t - \hat{\mathbf{y}}_t^i(\alpha))^2$$

$$\mathbf{w}_{t+1}^{ii}(\alpha) = \Phi_t^{ii}(\exp(-\frac{1}{c} L_{[1,t]}(\alpha))) \quad (4)$$

end

Algorithm 1: K-LMS-NET algorithm

and

$$\eta = \frac{1}{(1 + \frac{\sqrt{\hat{L}_A}}{\hat{H}_A \hat{X}_A}) \hat{X}_A^2} . \quad (7)$$

In the following Lemma, the loss and the norm of a particular predictor in $\mathcal{H}_{\alpha'}$ is used to bound the loss and norm of “near” comparable predictors in $\mathcal{H}_{\alpha''}$ when $|\alpha' - \alpha''|$ is small. First we define the surfeit of a function which we use instead of the norm induced by Hilbert space to bound the size of a function in H_α . In this note we avoid the technicalities of defining the surfeit for the complete Hilbert space \mathcal{H}_k ; we consider the definition only on the prehilbert space H_k .

Definition 1. Given a positive kernel ($\forall x, y \in \mathcal{X}^2 : k(x, y) \geq 0$), let $f \in H_k$; then define the surfeit by

$$\mathcal{S}^2(f) = \inf \left[\|f^+\|^2 + \|f^-\|^2 \right] . \quad (8)$$

The infimum is taken over all decompositions $f^+ + f^- = f$, where $f^+ = \sum_{i:\beta_i>0} \beta_i k(x_i, \cdot)$ and $f^- = \sum_{i:\beta_i<0} \beta_i k(x_i, \cdot)$ are a positive linear and negative linear combination of kernel functions, respectively, such that $f = f^+ + f^- = \sum_{i=1}^m \beta_i k(x_i, \cdot)$.

The infimum exists since $\|f\|^2 \leq \mathcal{S}^2(f)$.

Lemma 1. Let $k_\alpha(\mathbf{v}_1, \mathbf{v}_2) = \exp(-s_0 \alpha \|\mathbf{v}_1 - \mathbf{v}_2\|^2)$ denote a parameterized ($\alpha \in [0, 1]$) Gaussian kernel with fixed scale constant $s_0 \geq 1$ over the domain $[x_1, x_2]^n \times [x_1, x_2]^n$ with associated prehilbert spaces H_α . Given a function $h_{\alpha'} \in H_{\alpha'}$ such that $\|h_{\alpha'}\|_{\alpha'} \geq 1$ with representation

$$h_{\alpha'}(\cdot) = \sum_{i=1}^m \beta_i k_{\alpha'}(\mathbf{v}_i, \cdot) \quad (9)$$

then let

$$h_{\alpha'+\delta}(\cdot) = \sum_{i=1}^m \beta_i k_{\alpha'+\delta}(\mathbf{v}_i, \cdot). \quad (10)$$

Then the square loss and squared norm of $h_{\alpha'+\delta}$ may be bounded by those of $h_{\alpha'}$ plus any constant $0 < c < 1$ for all sequences $\langle (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_\ell, y_\ell) \rangle \in ([x_1, x_2]^n, [r_1, r_2])^\ell$ where $\max(|r_1|, |r_2|) \geq 1$. Hence

$$\sum_{t=1}^{\ell} (y_t - h_{\alpha'+\delta}(\mathbf{x}_t))^2 \leq \sum_{t=1}^{\ell} (y_t - h_{\alpha'}(\mathbf{x}_t))^2 + c \quad (11)$$

$$\|h_{\alpha'+\delta}\|_{\alpha'+\delta}^2 \leq \|h_{\alpha'}\|_{\alpha'}^2 + c \quad (12)$$

for all $\delta \in [0, \frac{c}{5s_0\ell \max(|r_1|, |r_2|)n(x_2-x_1)^2 \mathcal{S}^2(h_{\alpha'})}]$.

Proof. We will need the inequality,

$$(1+x) \leq e^x, \quad x \in (-\infty, \infty). \quad (13)$$

Let $\epsilon = s_0\delta n(x_1 - x_2)^2$ and given $h_{\alpha'}(\cdot) = \sum_{i=1}^m \beta_i k_{\alpha'}(\mathbf{v}_i, \cdot)$ define

$$\begin{aligned} h_{\alpha'}^+(\cdot) &= \sum_{i:\beta_i>0} \beta_i k_{\alpha'}(\mathbf{v}_i, \cdot); \quad h_{\alpha'}^-(\cdot) = \sum_{i:\beta_i<0} \beta_i k_{\alpha'}(\mathbf{v}_i, \cdot); \\ h_{\alpha'+\delta}^+(\cdot) &= \sum_{i:\beta_i>0} \beta_i k_{\alpha'+\delta}(\mathbf{v}_i, \cdot); \quad h_{\alpha'+\delta}^-(\cdot) = \sum_{i:\beta_i<0} \beta_i k_{\alpha'+\delta}(\mathbf{v}_i, \cdot), \end{aligned}$$

and assume without loss of generality that

$$\mathcal{S}^2(h_{\alpha'}) = \|h_{\alpha'}^+\|_{\alpha'}^2 + \|h_{\alpha'}^-\|_{\alpha'}^2 \quad (14)$$

(we discuss this simplifying assumption at the end of this proof). We have $\forall \mathbf{x} \in [x_1, x_2]^n$ that

$$h_{\alpha'}^+(\mathbf{x}) \geq h_{\alpha'+\delta}^+(\mathbf{x}) \geq (1-\epsilon)h_{\alpha'}^+(\mathbf{x}); \quad h_{\alpha'}^-(\mathbf{x}) \leq h_{\alpha'+\delta}^-(\mathbf{x}) \leq (1-\epsilon)h_{\alpha'}^-(\mathbf{x}) \quad (15)$$

by (13). Since $h_{\alpha'}(\mathbf{x}) = h_{\alpha'}^+(\mathbf{x}) + h_{\alpha'}^-(\mathbf{x})$ we have that

$$|h_{\alpha'}(\mathbf{x}) - h_{\alpha'+\delta}(\mathbf{x})| \leq \epsilon \max(h_{\alpha'}^+(\mathbf{x}), |h_{\alpha'}^-(\mathbf{x})|). \quad (16)$$

We now bound the square loss. Let $h_m = \max_{\mathbf{x}}(h_{\alpha'}^+(\mathbf{x}), |h_{\alpha'}^-(\mathbf{x})|)$; then

$$\begin{aligned} \sum_{t=1}^{\ell} (y_t - h_{\alpha'+\delta}(\mathbf{x}_t))^2 &\leq \sum_{t=1}^{\ell} (y_t - h_{\alpha'}(\mathbf{x}_t))^2 + \sum_{t=1}^{\ell} 2\epsilon |y_t - h_{\alpha'}(\mathbf{x}_t)| h_m + \sum_{t=1}^{\ell} \epsilon^2 h_m^2 \\ &\leq \sum_{t=1}^{\ell} (y_t - h_{\alpha'}(\mathbf{x}_t))^2 + \ell \epsilon h_m [2 \max(|r_1|, |r_2|) + 2\|h_{\alpha'}\|_{\alpha'} + \epsilon h_m] \end{aligned}$$

observing that we may bound $|h_{\alpha'}(\mathbf{x})| \leq \|h_{\alpha'}\|_{\alpha'}$ by the Cauchy-Schwarz inequality applied to $h_{\alpha'}(\mathbf{x})$. We now bound the norm squared,

$$\begin{aligned} \|h_{\alpha'+\delta}\|_{\alpha'+\delta}^2 &= \sum_{i=1}^m \sum_{j=1}^m \beta_i \beta_j k_{\alpha'}(\mathbf{v}_i, \mathbf{v}_j) k_{\delta}(\mathbf{v}_i, \mathbf{v}_j) \\ &\leq \|h_{\alpha'}^+\|_{\alpha'}^2 + \|h_{\alpha'}^-\|_{\alpha'}^2 + 2(1-\epsilon) \sum_{\{i:\beta_i>0\}} \sum_{\{j:\beta_j<0\}} \beta_i \beta_j k_{\alpha'}(\mathbf{v}_i, \mathbf{v}_j) \\ &\leq \|h_{\alpha'}\|_{\alpha'}^2 + \epsilon \left(\|h_{\alpha'}^+\|_{\alpha'}^2 + \|h_{\alpha'}^-\|_{\alpha'}^2 \right) \\ &\leq \|h_{\alpha'}\|_{\alpha'}^2 + \epsilon \mathcal{S}^2(h_{\alpha'}). \end{aligned} \quad (17)$$

Here (17) follows from the assumption in (14). Equations (11) and (12) now hold by substitution of the upper bound of δ into the definition ϵ with use of the assumption that $\|h_{\alpha'}\|_{\alpha'} \geq 1$ and the inequalities $\|h_{\alpha'}\|_{\alpha'}^2 \leq \mathcal{S}^2(h_{\alpha'})$ and $h_m^2 \leq \mathcal{S}^2(h_{\alpha'})$. Regarding the simplifying assumption (14) it could be the case that there does not exist a decomposition $\{h_{\alpha'}^+, h_{\alpha'}^-\}$ for which the infimum $\mathcal{S}^2(h_{\alpha'})$ is obtained; in this case $\mathcal{S}^2(h_{\alpha'})$ in (14) is then simply an upper bound for the true surfeit derived from the particular decomposition $\{h_{\alpha'}^+, h_{\alpha'}^-\}$. This upper bound holds (and also (17)), however, for every decomposition, hence there is always a decomposition whose upper bound is as arbitrarily close to the surfeit as desired. \square

Theorem 2. *Given the K-LMS-NET algorithm with learning rate η , an outcome range $[r_1, r_2]$ a parameterized ($\alpha \in [0, 1]$) Gaussian kernel, $k_\alpha(\mathbf{v}_1, \mathbf{v}_2) = \exp(-s_0\alpha\|\mathbf{v}_1 - \mathbf{v}_2\|^2)$ with fixed scale constant $s_0 \geq 1$ over the domain $[x_1, x_2]^n \times [x_1, x_2]^n$ with associated prehilbert spaces H_α a data sequence $\langle (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_\ell, y_\ell) \rangle \in ([x_1, x_2]^n, [r_1, r_2])^\ell$, and the constants $c \in (0, 1]$, $s_0 \geq 1$, $\hat{L} \geq 0$, $\hat{H} \geq 1 + c$ and with*

$$\eta = \frac{1}{(1 + \frac{\sqrt{\hat{L}}}{\hat{H}})} \quad (18)$$

then the total loss of the algorithm satisfies

$$\begin{aligned} \sum_{t=1}^{\ell} L(y_t, \hat{y}_t) &\leq \sum_{t=1}^{\ell} L(y_t, h_\alpha(\mathbf{x}_t)) + 2\sqrt{\hat{L}}\hat{H} + \hat{H}^2 + c \\ &+ 2(r_2 - r_1)^2 \left[\ln \ell + 2 \ln \mathcal{S}(h_\alpha) + \ln s_0 + \ln n + 2 \ln(x_2 - x_1) + \ln \max(|r_1|, |r_2|) + \ln \frac{1}{c} + \ln 5 \right] \end{aligned} \quad (19)$$

for every $h_\alpha \in \bigcup_{\alpha \in [0,1]} H_\alpha$ such that $\|h_\alpha\|_\alpha^2 + c \leq \hat{H}^2$, $\sum_{t=1}^{\ell} L(y_t, h_\alpha(\mathbf{x}_t)) + c \leq \hat{L}$,

$$\text{and } \alpha \in \left[0, 1 - \frac{c}{5s_0\ell \max(|r_1|, |r_2|)n(x_2 - x_1)^2\mathcal{S}^2(h_\alpha)} \right]. \quad (20)$$

Proof. The Theorem follows immediately from Theorem 1 and Lemma 1.

References

1. N. Cesa-Bianchi, P. Long, and M.K. Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Transactions on Neural Networks*, 7(2):604–619, May 1996.
2. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.
3. M. Herbster. Relative loss bounds and polynomial-time predictions for the K-LMS-NET algorithm (submitted for review to ALT '04), 2004.
4. Jyrki Kivinen and Manfred K. Warmuth. Averaging expert predictions. *Lecture Notes in Computer Science*, 1572:153–167, 1999.
5. N. Littlestone. *Mistake Bounds and Logarithmic Linear-threshold Learning Algorithms*. PhD thesis, Technical Report UCSC-CRL-89-11, University of California Santa Cruz, 1989.
6. N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.
7. V. Vovk. Aggregating strategies. In *Proc. 3rd Annu. Workshop on Comput. Learning Theory*, pages 371–383. Morgan Kaufmann, 1990.