Competing with Gaussian linear experts

Fedor Zhdanov and Vladimir Vovk
Computer Learning Research Centre,
Department of Computer Science,
Royal Holloway, University of London,
Egham, Surrey, TW20 0EX, UK
{fedor,vovk}@cs.rhul.ac.uk

Abstract

We study the problem of online regression. We do not make any assumptions about input vectors or outcomes. We prove a theoretical bound on the square loss of Ridge Regression. We also show that Bayesian Ridge Regression can be thought of as an online algorithm competing with all the Gaussian linear experts. We then consider the case of infinite-dimensional Hilbert spaces and prove relative loss bounds for the popular non-parametric kernelized Bayesian Ridge Regression and kernelized Ridge Regression. Our main theoretical guarantees have the form of equalities.

1 Introduction

In the online prediction framework we are provided with some input at each step and try to predict an outcome using this input and information from previous steps (Cesa-Bianchi and Lugosi, 2006). In a simple case in statistics, it is assumed that each outcome is the value, corrupted by Gaussian noise, of a linear function of input.

In competitive prediction the learner compares his loss at each step with the loss of any expert from a certain class of experts instead of making statistical assumptions about the data generating process. Experts may follow certain strategies. The learner wishes to predict almost as well as the best expert for *all* sequences.

Our main result is Theorem 1 in the next section, which compares the cumulative weighted square loss of Ridge Regression applied in the on-line mode with the regularized cumulative loss of the best linear predictor. The power of this result can be best appreciated by looking at the range of its implications, both known and new. For example, Corollary 1 answers the question asked by several researchers, see Vovk (2001), whether Ridge Regression has a relative loss bound with the regret term of the order $\ln T$ under the square loss function, where T is the number of steps and the outcomes are assumed bounded; this corollary (as well as all other implications stated in Section 2) is an explicit inequality rather than an asymptotic result. Theorem 1 itself is much stronger, stating an equality rather than inequality and not assuming that the outcomes are bounded. Since it is an equality, it unites upper and lower bounds on the

loss. It appears that all natural bounds on the square loss of Ridge Regression can be easily deduced from our theorem; we give some examples in the next section.

Most of previous research in online prediction considers experts that disregard the presence of noise in observations. We consider experts predicting a distribution on the outcomes. We use Bayesian Ridge Regression and prove that it can predict as well as the best regularized expert; this is our Theorem 2. The loss in this theoretical guarantee is the logarithmic loss. The algorithm that we apply was first used by DeSantis et al. (1988) and similar bounds to ours were obtained by Kakade and Ng (2004); Kakade et al. (2005). Theorem 2 is later used to deduce Theorem 1. Ridge Regression predicts the mean of the Bayesian Ridge Regression predictive distribution, and the logarithmic loss of Bayesian Ridge Regression is close to scaled square loss of Ridge Regression.

We extend our main result to the case of infinite dimensional Hilbert spaces of functions. The algorithm used becomes an analogue of non-parametric Bayesian methods. From Theorem 2 and Theorem 1 we deduce relative loss bounds on the logarithmic loss of kernelized Bayesian Ridge Regression and on the square loss of kernelized Ridge Regression in comparison with the loss of any function from a reproducing kernel Hilbert space. Both bounds have the form of equalities.

There is a lot of research done to prove upper and lower relative loss bounds under different loss functions. If the outcomes are assumed to be bounded, the strongest known theoretical guarantees for square loss are given by Vovk (2001) and Azoury and Warmuth (2001) for the algorithm which we call VAW (Vovk-Azoury-Warmuth) following Cesa-Bianchi and Lugosi (2006). In the case when the inputs and outcomes are not restricted in any way, like for our main guarantees, it is possible to prove certain loss bounds for the Gradient Descent; see Cesa-Bianchi et al. (1996).

In Section 2 of this paper we present the online regression framework and the main theoretical guarantee on the square loss of Ridge Regression. Section 3 describes what we call the Bayesian Algorithm. In Section 4 we show that Bayesian Ridge Regression is competitive with the experts which take into account the presence of noise in observations. In Section 5 we prove the main theorem. Section 6 describes the case of infinite-dimensional Hilbert spaces.

2 The prediction protocol and performance guarantees

In online regression the learner follows this prediction protocol:

Protocol 1 Online regression protocol

for $t = 1, 2, \dots$ **do**
  Reality announces $x_t \in \mathbb{R}^n$
  Learner predicts $\gamma_t \in \mathbb{R}$
  Reality announces $y_t \in \mathbb{R}$
end for

We use the Ridge Regression algorithm for the learner:

Algorithm 1 Online Ridge Regression

Require: $a > 0$
  Initialize $b_0 = 0 \in \mathbb{R}^n$, $A_0 = aI \in \mathbb{R}^{n \times n}$
  for $t = 1, 2, \ldots$ **do**
    Read $x_t \in \mathbb{R}^n$
    Predict $\gamma_t = b'_{t-1} A_{t-1}^{-1} x_t$
    Read y_t
    Update $A_t = A_{t-1} + x_t x'_t$
    Update $b_t = b_{t-1} + y_t x_t$
  end for

Following this algorithm the learner's prediction at step T can be written as

$$\gamma_T = \left(\sum_{t=1}^{T-1} y_t x_t \right)' \left(aI + \sum_{t=1}^{T-1} x_t x'_t \right)^{-1} x_T.$$

The incremental update of the matrix A_t^{-1} can be done effectively by the Sherman-Morrison formula. We prove the following theoretical guarantee for the square loss of the learner following Ridge Regression.

Theorem 1. *The Ridge Regression algorithm for the learner with $a > 0$ satisfies, at any step T,*

$$\sum_{t=1}^{T} \frac{(y_t - \gamma_t)^2}{1 + x'_t A_{t-1}^{-1} x_t} = \min_{\theta \in \mathbb{R}^n} \left(\sum_{t=1}^{T} (y_t - \theta' x_t)^2 + a\|\theta\|^2 \right). \tag{1}$$

Note that the part $x'_t A_{t-1}^{-1} x_t$ in the denominator is usually close to zero for large t. An equivalent equality is also obtained (but well hidden) in the proof of Theorem 4.6 in Azoury and Warmuth (2001). Our proof is more elegant. We describe it from the point of view of online prediction, but we note the connection with Bayesian learning in derivations. We obtain an upper bound in the form which is more familiar from online prediction literature.

Corollary 1. *Assume $|y_t| \le Y$ for all t, clip the predictions of Ridge Regression to $[-Y, Y]$, and denote them by γ_t^Y. Then*

$$\sum_{t=1}^{T} (y_t - \gamma_t^Y)^2 \le \min_{\theta} \left(\sum_{t=1}^{T} (y_t - \theta' x_t)^2 + a\|\theta\|^2 \right) + 4Y^2 \ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x'_t \right). \tag{2}$$

Proof. We first clip the predictions of Ridge Regression to $[-Y, Y]$ in Theorem 1. In this case the loss at each step can only become smaller, and so the equality transforms to an inequality. Since all the outcomes also lie in $[-Y, Y]$, the maximum square loss at each step is $4Y^2$. We have the following relations:

$$\frac{1}{1 + x'_t A_{t-1}^{-1} x_t} = 1 - \left(\frac{x'_t A_{t-1}^{-1} x_t}{1 + x'_t A_{t-1}^{-1} x_t} \right) \text{ and } \frac{x'_t A_{t-1}^{-1} x_t}{1 + x'_t A_{t-1}^{-1} x_t} \le \ln(1 + x'_t A_{t-1}^{-1} x_t).$$

The last inequality holds because $x_t' A_{t-1}^{-1} x_t$ is non-negative due to the positive definiteness of the matrix A_{t-1}. Thus we can use $\frac{b}{1+b} \le \ln(1+b), b \ge 0$ (it holds at $b = 0$, then take the derivatives of both sides). For the equality $\sum_{t=1}^{T} \ln(1 + x_t' A_{t-1}^{-1} x_t) = \ln\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t x_t'\right)$ see (16). □

The bound (2) is exactly the bound obtained in Theorem 4 in Vovk (2001) for the algorithm merging linear experts with predictions clipped to $[-Y, Y]$, which does not have a closed-form description and so is less interesting than clipped Ridge Regression. The bound for the VAW algorithm obtained in Theorem 1 in Vovk (2001) has Y^2 in place of $4Y^2$ (the VAW algorithm is very similar to Ridge Regression; its predictions are $b_{t-1}' A_t^{-1} x_t$ rather than $b_{t-1}' A_{t-1}^{-1} x_t$). The regret term in (2) has the logarithmic order in T if $\|x_t\|_\infty \le X$ for all t, because

$$\ln\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t x_t'\right) \le n \ln\left(1 + \frac{TX^2}{a}\right) \tag{3}$$

(the determinant of a positive definite matrix is bounded by the product of its diagonal elements; see Chapter 2, Theorem 7 of Beckenbach and Bellman (1961). This bound is also obtained in Theorem 4.6 in Azoury and Warmuth (2001).

From our Theorem 1 we can also deduce Theorem 11.7 of Cesa-Bianchi and Lugosi (2006), which is somewhat similar to our corollary. That theorem implies (2) when Ridge Regression's predictions happen to be in $[-Y, Y]$ without clipping (but this is not what Corollary 1 asserts).

The upper bound (2) does not hold if the coefficient 4 is replaced by any number less than $\frac{3}{2\ln 2} \approx 2.164$, as can be seen from an example given in Theorem 3 in Vovk (2001), where the left-hand side of (2) is $4T + o(T)$, the minimum in the right-hand side is at most T, $Y = 1$, and the logarithm is $2T \ln 2 + O(1)$. It is also known that there is no algorithm achieving (2) with the coefficient less than 1 instead of 4 even in the case where $\|x_t\|_\infty \le X$ for all t; see Theorem 2 in Vovk (2001).

It is also possible to prove an upper bound without the logarithmic part on the cumulative square loss of Ridge Regression without assuming that the outcomes are bounded.

Corollary 2. *If $\|x_t\|_2 \le Z$ for all t then the Ridge Regression algorithm for the learner with $a > 0$ satisfies, at any step T,*

$$\sum_{t=1}^{T} (y_t - \gamma_t)^2 \le \left(1 + \frac{Z^2}{a}\right) \min_{\theta \in \mathbb{R}^n} \left(\sum_{t=1}^{T} (y_t - \theta' x_t)^2 + a\|\theta\|^2\right). \tag{4}$$

Proof. Qazaz et al. (1997) showed that $1 + x_t' A_j^{-1} x_t \le 1 + x_t' A_i^{-1} x_t$ for $j \ge i$. We take $i = 0$ and obtain $1 + x_t' A_{t-1}^{-1} x_t \le 1 + Z^2/a$ for any t. □

This bound is better than the bound in Corollary 3.1 of Kakade and Ng (2004), which has an additional regret term of logarithmic order in time.

Asymptotic properties of the Ridge Regression algorithm can be further studied using Corollary A.1 in Kumon et al. (2009). It states that when $\|x_t\|_2 \le 1$ for all t,

then $x_t' A_{t-1}^{-1} x_t \to 0$ as $t \to \infty$. It is clear that we can replace $\|x_t\|_2 \le 1$ for all t by $\sup_t \|x_t\|_2 < \infty$. The following corollary states that if there exists a very good expert (asymptotically), then Ridge Regression also predicts very well. If there is no such a good expert, Ridge Regression performs asymptotically as well as the best regularized expert.

Corollary 3. *Let $a > 0$ and γ_t be the predictions output by the Ridge Regression algorithm with parameter a. Suppose $\sup_t \|x_t\|_2 < \infty$.*

1. *If*

$$\exists \theta \in \mathbb{R}^n : \sum_{t=1}^{\infty} (y_t - \theta' x_t)^2 < \infty, \tag{5}$$

then

$$\sum_{t=1}^{\infty} (y_t - \gamma_t)^2 < \infty.$$

2. *If*

$$\forall \theta \in \mathbb{R}^n : \sum_{t=1}^{\infty} (y_t - \theta' x_t)^2 = \infty, \tag{6}$$

then

$$\lim_{T \to \infty} \frac{\sum_{t=1}^{T} (y_t - \gamma_t)^2}{\min_{\theta \in \mathbb{R}^n} \left(\sum_{t=1}^{T} (y_t - \theta' x_t)^2 + a \|\theta\|^2 \right)} = 1. \tag{7}$$

Proof. **Part 1.** Suppose that the condition (5) holds. Then the right-hand side of (1) is bounded by a constant (independent of T). By Corollary A.1 in Kumon et al. (2009), the denominators in the left-hand side converge to 1 as $t \to \infty$ and so are bounded. Therefore, the sequence $\sum_{t=1}^{T} (y_t - \gamma_t)^2$ remains bounded as $T \to \infty$.

Part 2. Suppose that the condition (6) holds and the right-hand side of (1) is bounded above by a constant C. Then for each T there exists θ_T such that

$$\sum_{t=1}^{T} (y_t - \theta_T' x_t)^2 + a \|\theta_T\|^2 \le C.$$

It follows that each θ_T belongs to the closed ball with centre 0 and of radius $\sqrt{C/a}$. This ball is a compact set, and thus the sequence θ_T has a subsequence that converges to some $\tilde{\theta}$. For each T_0 we have $\sum_{t=1}^{T_0} (y_t - \tilde{\theta}' x_t)^2 \le C$, because otherwise we would have $\sum_{t=1}^{\hat{T}} (y_t - \theta_{\hat{T}}' x_t)^2 > C$ for a large enough $\hat{T}$ in the subsequence. Therefore, we have arrived at a contradiction: $\sum_{t=1}^{\infty} (y_t - \tilde{\theta}' x_t)^2 \le C < \infty$.

Once we know that the right-hand side of (1) tends to ∞ as $T \to \infty$ and the denominators on the left-hand side tend to 1 (this is true by Corollary A.1 in Kumon et al., 2009), (7) becomes intuitively plausible since, as far as the conclusion (7) is concerned, we can ignore the finite number of ts for which the denominator $1 + x_t' A_{t-1}^{-1} x_t$ is significantly different from 1. We will, however, give a formal argument.

The inequality ≥ 1 in (7) is clear from (1) and $1 + x_t' A_{t-1}^{-1} x_t \geq 1$. We shall prove the inequality ≤ 1 now. Choose a small $\epsilon > 0$. Then starting from some $t = T_0$ we have that the denominators $1 + x_t' A_{t-1}^{-1} x_t$ are less than $1 + \epsilon$. Thus, for $T > T_0$,

$$\begin{aligned}\sum_{t=1}^{T}(y_t - \gamma_t)^2 &= \sum_{t=1}^{T_0}(y_t - \gamma_t)^2 + \sum_{t=T_0+1}^{T}(y_t - \gamma_t)^2 \\ &\leq \sum_{t=1}^{T_0}(y_t - \gamma_t)^2 + (1+\epsilon)\sum_{t=1}^{T}\frac{(y_t - \gamma_t)^2}{1 + x_t' A_{t-1}^{-1} x_t} \\ &= \sum_{t=1}^{T_0}(y_t - \gamma_t)^2 + (1+\epsilon)\min_{\theta \in \mathbb{R}^n}\left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2\right).\end{aligned}$$

This implies that the left-hand side of (7) with $\lim$ replaced by $\limsup$ does not exceed $1 + \epsilon$, and it remains to remember that ϵ can be taken arbitrarily small. □

3 Bayesian algorithm

In this section we describe the main algorithm used to prove our theoretical bounds. Let us denote the set of possible outcomes by Ω, the index set for the experts by Θ, and the set of allowed predictions by Γ. The quality of predictions is measured by a loss function $\lambda : \Gamma \times \Omega \to \mathbb{R}$. We have $\Omega = \mathbb{R}$, $\Theta = \mathbb{R}^n$, and Γ is the set of all measurable functions on the real line integrable to one. The loss function λ is the logarithmic loss $\lambda(\gamma, y) = -\ln \gamma(y)$, where $\gamma \in \Gamma$ and $y \in \Omega$. The learner follows the prediction with expert advice protocol.

Protocol 2 Prediction with expert advice protocol

Initialize $L_0 := 0$ and $L_0(\theta) = 0, \forall \theta \in \Theta$
for $t = 1, 2, \ldots$ **do**
  Experts $\theta \in \Theta$ announce their predictions $\xi_t^\theta \in \Gamma$
  Learner predicts $\gamma_t \in \Gamma$
  Reality announces $y_t \in \Omega$
  Losses are updated: $L_T = L_{T-1} + \lambda(\gamma_t, y_t)$, $L_T(\theta) = L_{T-1}(\theta) + \lambda(\xi_t^\theta, y_t), \forall \theta \in \Theta$
end for

Here by L_T we denote the cumulative loss of the learner at step T, and by $L_T(\theta)$ we denote the cumulative loss of the expert θ at this step.

We use a standard algorithm in prediction with expert advice (a special case of the Aggregating Algorithm for the logarithmic loss function and learning rate 1, going back to DeSantis et al. (1988) in the case of countable Θ and Ω) to derive the main theoretical bound and give predictions. We call it the Bayesian Algorithm (BA) as it is virtually identical to the Bayes rule used in Bayesian learning (the main difference being that the experts are not required to follow any prediction strategies). Instead of

looking for the best expert, the algorithm considers all the experts and takes a weighted average of their predictions as its own prediction. In detail, it works as follows.

Algorithm 2 Bayesian Algorithm

Require: A probability measure $P_0(d\theta) = P_0^*(d\theta)$ on Θ (the prior distribution, or weights)
  for $t = 1, 2, \dots$ **do**
    Read experts' predictions $\xi_t^\theta \in \Gamma, \forall \theta \in \Theta$
    Predict $g_t = \int_\Theta \xi_t^\theta P_{t-1}^*(d\theta)$
    Read y_t
    Update the weights $P_t(d\theta) = \xi_t^\theta(y_t) P_{t-1}(d\theta)$
    Normalize the weights $P_t^*(d\theta) = P_t(d\theta) / \int_\Theta P_t(d\theta)$
  end for

The experts' weights are updated according to their losses at each step: $\xi_t^\theta(y_t) = e^{-\lambda(\xi_t^\theta, y_t)}$; larger losses lead to smaller weights. After t steps the weights become

$$P_t(d\theta) = e^{-L_t(\theta)} P_0(d\theta). \tag{8}$$

The normalized weights $P_T^*(d\theta)$ correspond to the posterior distribution over θ after the step T. As we said, the prediction of the BA at step T is given by the average

$$g_T = \int_\Theta \xi_T^\theta P_{T-1}^*(d\theta) \tag{9}$$

of the experts' predictions.

The next lemma is a special case of Lemma 1 in Vovk (2001). It shows that the cumulative loss of the BA is an average of the experts' cumulative losses in a generalized sense, as in, e.g., Chapter 3 of Hardy et al. (1952).

Lemma 1. *For any prior P_0 and any $T = 1, 2, \dots$, the cumulative loss of the BA can be expressed as*

$$L_T = -\ln \int_\Theta e^{-L_T(\theta)} P_0(d\theta). \tag{10}$$

Proof. We proceed by induction in T: for $T = 0$ the equality is obvious, and for $T > 0$ we have:

$$\begin{aligned} L_T &= L_{T-1} - \ln g_T(y_T) \\ &= -\ln \int_\Theta e^{-L_{T-1}(\theta)} P_0(d\theta) - \ln \int_\Theta \xi_T^\theta \frac{e^{-L_{T-1}(\theta)}}{\int_\Theta e^{-L_{T-1}(\theta)} P_0(d\theta)} P_0(d\theta) \\ &= -\ln \int_\Theta e^{-L_T(\theta)} P_0(d\theta) \end{aligned}$$

(the second equality follows from the inductive assumption, the definition of g_T, and (8)). □

4 Bayesian Ridge Regression as a competitive algorithm

Let us consider experts whose predictions at step t are the densities of the normal distributions $N(\theta' x_t, \sigma^2)$ on the set of outcomes for some fixed variance $\sigma^2 > 0$ (so each expert θ follows a fixed strategy). From the statistical point of view, they predict according to the model $y_t = \theta' x_t + \epsilon_t$ with Gaussian noise $\epsilon_t \sim N(0, \sigma^2)$. In other words, the prediction of each expert $\theta \in \Theta$ is

$$\xi_t^\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta' x_t)^2}{2\sigma^2}}. \tag{11}$$

Let us take the initial distribution $N(0, \frac{\sigma^2}{a} I)$ on the experts with some $a > 0$:

$$P_0(d\theta) = \left(\frac{a}{2\sigma^2\pi}\right)^{n/2} \exp\left(-\frac{a}{2\sigma^2}\|\theta\|^2\right) d\theta.$$

We will prove that in this setting the prediction of the Bayesian Algorithm is equal to the prediction of Bayesian Ridge Regression. But first we need to introduce some notation. For $t \in \{1, 2, \ldots\}$, let X_t be the $t \times n$ matrix of row vectors $x_1', \ldots, x_t'$ and Y_t be the column vector of outcomes $y_1, \ldots, y_t$. Let $A_t = X_t' X_t + aI$, as before. Bayesian Ridge Regression is the algorithm predicting at each step T the normal distribution $N(\gamma_T, \sigma_T^2)$ with the mean and variance given by

$$\gamma_T = Y_{T-1}' X_{T-1} A_{T-1}^{-1} x_T, \quad \sigma_T^2 = \sigma^2 x_T' A_{T-1}^{-1} x_T + \sigma^2 \tag{12}$$

for some $a > 0$ and the known noise variance σ^2.

Lemma 2. *In our setting the prediction* (9) *of the Bayesian Algorithm is the prediction density of Bayesian Ridge Regression in the notation of* (12)*:*

$$g_T(y) = \frac{1}{\sqrt{2\pi\sigma_T^2}} e^{-\frac{(y-\gamma_T)^2}{2\sigma_T^2}}. \tag{13}$$

Proof. The prediction

$$g_T(y) = \int_\Theta \xi_T^\theta(y) P_{T-1}^*(d\theta) = \frac{\int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta' x_T)^2}{2\sigma^2}} \prod_{t=1}^{T-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t-\theta' x_t)^2}{2\sigma^2}} P_0(d\theta)}{\int_{\mathbb{R}^n} \prod_{t=1}^{T-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t-\theta' x_t)^2}{2\sigma^2}} P_0(d\theta)}$$

formally coincides with the density of the predictive distribution of the Bayesian Gaussian linear model, and so equality (13) is true: see Section 3.3.2 of Bishop (2006). □

Remark 1. From the probabilistic point of view Lemma 2 is usually explained in the following way (Hoerl and Kennard, 2000). The posterior distribution $P_{T-1}^*(\theta)$ is $N(A_{T-1}^{-1} X_{T-1}' Y_{T-1}, \sigma^2 A_{T-1}^{-1})$. The conditional distribution of $\theta' x_T$ given the training examples is then $N(Y_{T-1}' X_{T-1} A_{T-1}^{-1} x_T, \sigma^2 x_T' A_{T-1}^{-1} x_T)$, and so the predictive distribution is $N(Y_{T-1}' X_{T-1} A_{T-1}^{-1} x_T, \sigma^2 x_T' A_{T-1}^{-1} x_T + \sigma^2)$.

For the subsequent derivations, we will need the following well-known lemma, whose proof can be found in Lemma 8 of Busuttil (2008) or extracted from Chapter 2, Theorem 3 of Beckenbach and Bellman (1961).

Lemma 3. *Let $W(\theta) = \theta' A \theta + b'\theta + c$ for $\theta, b \in \mathbb{R}^n$, c be a scalar, and A be a symmetric positive definite $n \times n$ matrix. Then*

$$\int_{\mathbb{R}^n} e^{-W(\theta)} d\theta = e^{-W_0} \frac{\pi^{n/2}}{\sqrt{\det A}},$$

where $W_0 = \min_\theta W(\theta)$.

The right-hand side of (10) can be transformed to the regularized cumulative loss of the best expert θ and a regret term:

Theorem 2. *For any sequence $x_1, y_1, x_2, y_2, \ldots$, the cumulative logarithmic loss of the Bayesian Ridge Regression algorithm* (13) *at any step T can be expressed as*

$$L_T = \min_\theta \left(L_T(\theta) + \frac{a}{2\sigma^2} \|\theta\|^2 \right) + \frac{1}{2} \ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right). \tag{14}$$

If $\|x_t\|_\infty \le X$ for any $t = 1, 2, \ldots$, then

$$L_T \le \min_\theta \left(L_T(\theta) + \frac{a}{2\sigma^2} \|\theta\|^2 \right) + \frac{n}{2} \ln \left(1 + \frac{TX^2}{a} \right). \tag{15}$$

Proof. We have to calculate the right-hand side of (10). The integral is expressed as

$$\int_\Theta \frac{1}{(2\pi\sigma^2)^{T/2}} \left(\frac{a}{2\sigma^2 \pi} \right)^{n/2} e^{-\frac{1}{2\sigma^2}\left(\sum_{t=1}^{T} (y_t - \theta' x_t)^2 + a\|\theta\|^2 \right)} d\theta.$$

By Lemma 3 it is equal to

$$\frac{1}{(2\pi\sigma^2)^{T/2}} \left(\frac{a}{2\sigma^2 \pi} \right)^{n/2} e^{-\frac{1}{2\sigma^2}\left(\sum_{t=1}^{T} (y_t - \theta_0' x_t)^2 + a\|\theta_0\|^2 \right)} \frac{\pi^{n/2}}{\sqrt{\det A_T}},$$

where A_T is the coefficient matrix in the quadratic part: $A_T = \frac{1}{2\sigma^2}(aI + \sum_{t=1}^{T} x_t x_t')$ and θ_0 is the best predictor: $\theta_0 = \arg\min_\theta \left(\sum_{t=1}^{T} (y_t - \theta' x_t)^2 + a\|\theta\|^2 \right)$. Taking the minus logarithm of this expression we get

$$-\sum_{t=1}^{T} \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_t - \theta_0' x_t)^2} \right) + \frac{a}{2\sigma^2} \|\theta_0\|^2 + \frac{1}{2} \ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right).$$

To obtain the upper bound (15) it suffices to apply (3). □

This theorem shows that the Bayesian Ridge Regression algorithm can be thought of as an online algorithm successfully competing with all the Gaussian linear models under the logarithmic loss function. Similar bounds on the logarithmic loss of Bayesian Ridge Regression are proven by Kakade and Ng (2004).

5 Proof of Theorem 1

Let us rewrite L_T and $L_T(\theta)$ using (13), the expression for σ_t^2 given by (12), and (11):

$$\begin{aligned}
L_T &= -\sum_{t=1}^{T} \ln\left(\frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{(y-\gamma_t)^2}{2\sigma_t^2}}\right) \\
&= \frac{1}{2}\ln\left((2\pi\sigma^2)^T \prod_{t=1}^{T}(1+x_t' A_{t-1}^{-1} x_t)\right) + \frac{1}{2\sigma^2}\sum_{t=1}^{T}\frac{(y_t-\gamma_t)^2}{1+x_t' A_{t-1}^{-1} x_t}, \\
L_T(\theta) &= \sum_{t=1}^{T}\lambda(\xi_t^\theta, y_t) = -\ln\left(\frac{1}{(2\pi\sigma^2)^{T/2}} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t-\theta' x_t)^2}\right) \\
&= \frac{T}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t-\theta' x_t)^2.
\end{aligned}$$

Substituting these expression into (14) we have:

$$\begin{aligned}
&\frac{1}{2}\ln\prod_{t=1}^{T}(1+x_t' A_{t-1}^{-1} x_t) + \frac{1}{2\sigma^2}\sum_{t=1}^{T}\frac{(y_t-\gamma_t)^2}{1+x_t' A_{t-1}^{-1} x_t} \\
&\quad = \frac{1}{2\sigma^2}\min_{\theta}\left(\sum_{t=1}^{T}(y_t-\theta' x_t)^2 + a\|\theta\|^2\right) + \frac{1}{2}\ln\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t x_t'\right).
\end{aligned}$$

Equation (1) follows from the fact that

$$\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t x_t'\right) = \prod_{t=1}^{T}(1+x_t' A_{t-1}^{-1} x_t) \tag{16}$$

for $A_t = aI + \sum_{i=1}^{t} x_i x_i'$. This fact can be proven by induction in T: for $T = 0$ it is obvious ($1 = 1$) and for $T \geq 1$ we have

$$\begin{aligned}
&\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t x_t'\right) = a^{-n}\det A_T = a^{-n}\det\left(A_{T-1} + x_T x_T'\right) \\
&= a^{-n}(1+x_T' A_{T-1}^{-1} x_T)\det A_{T-1} = \det\left(I + \frac{1}{a}\sum_{t=1}^{T-1} x_t x_t'\right)(1+x_T' A_{T-1}^{-1} x_T) \\
&\qquad\qquad = \prod_{t=1}^{T}(1+x_t' A_{t-1}^{-1} x_t).
\end{aligned}$$

The third equality follows from the Matrix Determinant Lemma: see, e.g., Theorem 18.1.1 of Harville (1997). The last equality follows from the inductive assumption. Note that σ^2 canceled out; this is natural as Ridge Regression (unlike Bayesian Ridge Regression) does not depend on σ.

6 Kernelized Ridge Regression

In this section we prove bounds on the square loss of kernelized Ridge Regression. We also prove bounds on the logarithmic loss for a commonly used non-parametric Gaussian algorithm: kernelized Bayesian Ridge Regression. These bounds explicitly handle infinite dimensional classes of experts.

Let $\mathbf{X}$ be an arbitrary set of inputs. We define a *reproducing kernel Hilbert space (RKHS)* $\mathcal{F}$ of functions $\mathbf{X} \to \mathbb{R}$ as a functional Hilbert space with continuous evaluation functional $f \in \mathcal{F} \mapsto f(x)$ for each $x \in \mathbf{X}$. By the Riesz-Fischer theorem for any $x \in \mathbf{X}$ there is a unique $k_x \in \mathcal{F}$ such that $\langle k_x, f \rangle_{\mathcal{F}} = f(x)$ for any $f \in \mathcal{F}$. The *kernel* $\mathcal{K} : \mathbf{X}^2 \to \mathbb{R}$ of the RKHS $\mathcal{F}$ is defined as $\mathcal{K}(x_1, x_2) = \langle k_{x_1}, k_{x_2} \rangle$ for any $x_1, x_2 \in \mathbf{X}$. For more information about kernels please refer to Schölkopf and Smola (2002).

Let us introduce some notation. Let $\mathbf{K}_t$ be the kernel matrix $\mathcal{K}(x_i, x_j)$ at step t, where $i, j = 1, \dots, t$. Let $\mathbf{k}_t$ be the column vector $\mathcal{K}(x_i, x_t)$ for $i = 1, \dots, t-1$. As before, Y_t is the column vector of outcomes $y_1, \dots, y_t$. The kernelized Ridge Regression is defined as the learner's strategy in Protocol 1 that predicts $\gamma_T = Y'_{T-1}(aI + \mathbf{K}_{T-1})^{-1}\mathbf{k}_T$ at each step T; see, e.g., Saunders et al. (1998). The following theorem is an analogue of Theorem 1 for kernelized Ridge Regression; in its proof we will see how kernelized Ridge Regression is connected with Ridge Regression.

Theorem 3. *The kernelized Ridge Regression algorithm for the learner with $a > 0$ satisfies, at any step T,*

$$\sum_{t=1}^{T} \frac{(y_t - \gamma_t)^2}{1 + (\mathcal{K}(x_t, x_t) - \mathbf{k}'_t(aI + \mathbf{K}_{t-1})^{-1}\mathbf{k}_t)/a} = \min_{f \in \mathcal{F}} \left(\sum_{t=1}^{T} (y_t - f(x_t))^2 + a\|f\|^2_{\mathcal{F}} \right). \tag{17}$$

Proof. It suffices to prove that for each $T \in \{1, 2, \dots\}$ and every sequence of input vectors and outcomes $(x_1, y_1, \dots, x_T, y_T) \in (\mathbf{X} \times \mathbb{R})^T$ the equality (17) is satisfied. Fix such T and $(x_1, y_1, \dots, x_T, y_T)$; our goal is to prove (17). Fix an isomorphism between the linear span of $k_{x_1}, \dots, k_{x_T}$ and $\mathbb{R}^{\tilde{T}}$, where $\tilde{T} \le T$ is the dimension of the linear span of $k_{x_1}, \dots, k_{x_T}$. Let $\tilde{x}_1, \dots, \tilde{x}_T \in \mathbb{R}^{\tilde{T}}$ be the images of $k_{x_1}, \dots, k_{x_T}$, respectively, under this isomorphism. Notice that, for all t, $\mathbf{K}_t$ is the matrix $\langle \tilde{x}_i, \tilde{x}_j \rangle$, $i, j = 1, \dots, t$, and $\mathbf{k}_t$ is the column vector $\langle \tilde{x}_i, \tilde{x}_t \rangle$ for $i = 1, \dots, t-1$. We know that (1) with $\tilde{x}_t$ in place of x_t and $\tilde{\gamma}_t$ in place of γ_t holds for Ridge Regression, whose predictions are now denoted $\tilde{\gamma}_t$ (in order not to confuse them with kernelized Ridge Regression's predictions γ_t). The predictions output by Ridge Regression on $\tilde{x}_1, y_1, \dots, \tilde{x}_T, y_T$ and by kernelized Ridge Regression on $x_1, y_1, \dots, x_T, y_T$ are the same:

$$\begin{aligned} \gamma_t = Y'_{t-1}(aI + \mathbf{K}_{t-1})^{-1}\mathbf{k}_t &= Y'_{t-1}(aI + \tilde{X}_{t-1}\tilde{X}'_{t-1})^{-1}\tilde{X}_{t-1}\tilde{x}_t \\ &= Y'_{t-1}\tilde{X}_{t-1}(aI + \tilde{X}'_{t-1}\tilde{X}_{t-1})^{-1}\tilde{x}_t = \tilde{\gamma}_t \end{aligned}$$

(for the notation see (12), with tildes added). The denominators in (17) and (1) are also

the same:

$$\begin{aligned}
&1 + (\mathcal{K}(x_t, x_t) - \mathbf{k}_t'(aI + \mathbf{K}_{t-1})^{-1}\mathbf{k}_t)/a \\
&\qquad = 1 + \tilde{x}_t'(I - \tilde{X}_{t-1}'(aI + \tilde{X}_{t-1}\tilde{X}_{t-1}')^{-1}\tilde{X}_{t-1})\tilde{x}_t/a \\
&\quad = 1 + \tilde{x}_t'(aI + \tilde{X}_{t-1}'\tilde{X}_{t-1})^{-1}((aI + \tilde{X}_{t-1}'\tilde{X}_{t-1}) - \tilde{X}_{t-1}'\tilde{X}_{t-1})\tilde{x}_t/a \\
&\qquad = 1 + \tilde{x}_t'(aI + \tilde{X}_{t-1}'\tilde{X}_{t-1})^{-1}\tilde{x}_t.
\end{aligned}$$

The right-hand sides are the same by the representer theorem (see, e.g., Theorem 4.2 in Schölkopf and Smola, 2002). Indeed, by this theorem we have

$$\begin{aligned}
&\min_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(y_t - f(x_t))^2 + a\|f\|_{\mathcal{F}}^2\right) \\
&\qquad = \min_{c_1,\dots,c_T\in\mathbb{R}}\left(\sum_{t=1}^{T}\left(y_t - \sum_{i=1}^{T}c_i\mathcal{K}(x_i, x_t)\right)^2 + a\left\|\sum_{i=1}^{T}c_i k_{x_i}\right\|_{\mathcal{F}}^2\right) \\
&\qquad\qquad = \min_{c_1,\dots,c_T\in\mathbb{R}}\left(\sum_{t=1}^{T}\left(y_t - \sum_{i=1}^{T}c_i\langle\tilde{x}_i, \tilde{x}_t\rangle\right)^2 + a\left\|\sum_{i=1}^{T}c_i\tilde{x}_i\right\|_2^2\right)
\end{aligned}$$

(the last equality holds due to the isomorphism). Denoting $\theta = \sum_{i=1}^{T}c_i\tilde{x}_i \in \mathbb{R}^{\tilde{T}}$ we obtain the expression for the minimum in (1): θ ranges over the whole of $\mathbb{R}^{\tilde{T}}$ (as $c_1, \dots, c_T$ range over $\mathbb{R}$) since $\tilde{x}_1, \dots, \tilde{x}_T$ span $\mathbb{R}^{\tilde{T}}$. □

Similarly to the proof of Theorem 3 we can prove an analogue of Theorem 2 for kernelized Bayesian Ridge Regression. At step T kernelized Bayesian Ridge Regression predicts the normal density on outcomes with the mean γ_T and variance $\sigma^2 + \sigma^2(\mathcal{K}(x_T, x_T) - \mathbf{k}_T'(aI + \mathbf{K}_{T-1})^{-1}\mathbf{k}_T)/a$. We denote by L_T the cumulative logarithmic loss, over the first T steps, of the algorithm and by $L_T(f)$ the cumulative logarithmic loss of the expert f predicting normal density with the mean $f(x_t)$ and variance σ^2.

Theorem 4. *For any sequence $x_1, y_1, x_2, y_2, \dots$, the cumulative logarithmic loss of the kernelized Bayesian Ridge Regression algorithm at any step T can be expressed as*

$$L_T = \min_{f\in\mathcal{F}}\left(L_T(f) + \frac{a}{2\sigma^2}\|f\|_{\mathcal{F}}^2\right) + \frac{1}{2}\ln\det\left(I + \frac{1}{a}\mathbf{K}_T\right).$$

This theorem is proven by Kakade et al. (2005) for $a = 1$.

We can see from Theorem 13.3.8 of Harville (1997) that

$$\begin{aligned}
\det\left(I + \frac{1}{a}\mathbf{K}_T\right) &= \det\begin{pmatrix} I + \mathbf{K}_{T-1}/a & \mathbf{k}_T/a \\ \mathbf{k}_T'/a & 1 + \mathcal{K}(x_T, x_T)/a \end{pmatrix} \\
&= \det\left(I + \frac{1}{a}\mathbf{K}_{T-1}\right)(1 + (\mathcal{K}(x_T, x_T) - \mathbf{k}_T'(aI + \mathbf{K}_{T-1})^{-1}\mathbf{k}_T)/a),
\end{aligned}$$

and so by induction we have

$$\det\left(I+\frac{1}{a}\mathbf{K}_T\right)=\prod_{t=1}^{T}(1+(\mathcal{K}(x_t,x_t)-\mathbf{k}_t'(aI+\mathbf{K}_{t-1})^{-1}\mathbf{k}_t)/a),$$

with $\mathbf{k}_1'(aI+\mathbf{K}_0)^{-1}\mathbf{k}_1$ understood to be 0. Using this equality and following the arguments of the proof of Corollary 1 we obtain the following corollary from Theorem 3.

Corollary 4. *Assume $|y_t|\le Y$ for all t, clip the predictions of kernelized Ridge Regression to $[-Y,Y]$, and denote them by γ_t^Y. Then*

$$\sum_{t=1}^{T}(y_t-\gamma_t^Y)^2\le\min_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(y_t-f(x_t))^2+a\|f\|_{\mathcal{F}}^2\right)+4Y^2\ln\det\left(I+\frac{1}{a}\mathbf{K}_T\right). \tag{18}$$

It is possible to prove this corollary directly from Corollary 1 using the same argument as in the proof of Theorem 3.

The order of the regret term in (18) is not clear on the face of it. We show that it has the order $O(\sqrt{T})$ in many cases. We will use the notation $c_{\mathcal{F}}^2=\sup_{x\in\mathbf{X}}\mathcal{K}(x,x)$. We bounding the logarithm of the determinant and obtain that $\ln\det\left(I+\frac{1}{a}\mathbf{K}_T\right)\le T\ln\left(1+\frac{c_{\mathcal{F}}^2}{a}\right)$ (cf. (3)). If we know the number T of steps in advance, then we can choose a specific value for a; let $a=c_{\mathcal{F}}\sqrt{T}$. Thus we get an upper bound with the regret term of the order $O(\sqrt{T})$ for any $f\in\mathcal{F}$:

$$\sum_{t=1}^{T}(y_t-\gamma_t^Y)^2\le\sum_{t=1}^{T}(y_t-f(x_t))^2+c_{\mathcal{F}}(\|f\|_{\mathcal{F}}^2+4Y^2)\sqrt{T}.$$

If we do not know the number of steps in advance, it is possible to achieve a similar bound using the Bayesian Algorithm with a suitable prior over the parameter a:

$$\sum_{t=1}^{T}(y_t-\gamma_t^Y)^2\le\sum_{t=1}^{T}(y_t-f(x_t))^2+8Y\max\left(c_{\mathcal{F}}\|f\|_{\mathcal{F}},Y\delta T^{-1/2+\delta}\right)\sqrt{T+2}$$
$$+6Y^2\ln T+c_{\mathcal{F}}^2\|f\|_{\mathcal{F}}^2+O(Y^2) \tag{19}$$

for any arbitrarily small $\delta>0$, where the constant implicit in $O(Y^2)$ depends only on δ. (Proof omitted.)

In particular, (19) shows that if $\mathbf{X}$ is a universal kernel (Steinwart, 2001) on a topological space $\mathbf{X}$, Ridge Regression is competitive with all continuous functions on $\mathbf{X}$: for any continuous $f:\mathbf{X}\to\mathbb{R}$,

$$\limsup_{T\to\infty}\frac{1}{T}\left(\sum_{t=1}^{T}(y_t-\gamma_t^Y)^2-\sum_{t=1}^{T}(y_t-f(x_t))^2\right)\le 0 \tag{20}$$

(assuming $|y_t|\le Y$ for all t). For example, (20) holds for $\mathbf{X}$ a compact set in $\mathbb{R}^n$, $\mathcal{K}$ an RBF kernel, and $f:\mathbf{X}\to\mathbb{R}$ any continuous function, see Example 1 of Steinwart (2001).

Acknowledgements

We are very grateful for useful comments to Yuri Kalnishkan and Alexey Chernov. Thanks to the organizers and lecturers of the Cambridge Machine Learning Summer School 2009, whose work helped us to look at the usual problems from a new viewpoint. This work was supported by EPSRC (grant EP/F002998/1).

References

Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43:211–246, 2001.

Edwin F. Beckenbach and Richard Bellman. *Inequalities*. Springer, Berlin, 1961.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

Steven Busuttil. *The Aggregating Algorithm and Regression*. PhD thesis, Department of Computer Science, Royal Holloway University of London, UK, 2008.

Nicolò Cesa-Bianchi, Philip M. Long, and Manfred K. Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Transactions on Neural Networks*, 7:604–619, 1996.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, UK, 2006.

Alfredo DeSantis, George Markowsky, and Mark N. Wegman. Learning probabilistic prediction functions. In *Proceedings of the 29th Annual IEEE Symposium on Foundations of Computer Science*, pages 110–119, Los Alamitos, CA, USA, 1988. IEEE Computer Society.

Godfrey H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, Cambridge, UK, second edition, 1952.

David A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer, New York, 1997.

Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: biased estimation for nonorthogonal problems. *Technometrics*, 42:80–86, 2000.

Sham M. Kakade and Andrew Y. Ng. Online bounds for Bayesian algorithms. In *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, 2004.

Sham M. Kakade, Matthias W. Seeger, and Dean P. Foster. Worst-case bounds for Gaussian process models. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, 2005.

Masayuki Kumon, Akimichi Takemura, and Kei Takeuchi. Sequential optimizing strategy in multi-dimensional bounded forecasting games. *Technical report, arXiv:0911.3933 [math.PR], arXiv.org e-Print archive*, 2009.

Cazhaow S. Qazaz, Christopher K. I. Williams, and Christopher M. Bishop. An upper bound on the Bayesian error bars for generalized linear regression. In *Proceedings of the First International Conference on Mathematics of Neural Networks: Models, Algorithms and Applications*, pages 295–299, 1997.

Craig Saunders, Alex Gammerman, and Vladimir Vovk. Ridge regression learning algorithm in dual variables. In Jude W. Shavlik, editor, *Machine Learning, Proceedings of the Fifteenth International Conference*, pages 515–521, San Francisco, CA, 1998. Morgan Kaufmann.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2002.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69: 213–248, 2001.