An Identity for Kernel Ridge Regression

Fedor Zhdanov and Yuri Kalnishkan

Computer Learning Research Centre and Department of Computer Science,
Royal Holloway, University of London, Egham, Surrey, TW20 0EX, United Kingdom
{fedor,yura}@cs.rhul.ac.uk

Abstract. This paper provides a probabilistic derivation of an identity connecting the square loss of ridge regression in on-line mode with the loss of a retrospectively best regressor. Some corollaries of the identity providing upper bounds for the cumulative loss of on-line ridge regression are also discussed.

1 Introduction

Ridge regression is a powerful technique of machine learning. It was introduced in [9]; the kernel version of it is derived in [15].

Ridge regression can be used as a batch or on-line algorithm. This paper proves an identity connecting the square losses of ridge regression used on the same data in batch and on-line fashions. The identity and the approach to the proof are not entirely new. The identity implicitly appears in [2] for the linear case (it can be obtained by summing (4.21) from [2] in an exact rather than estimated form). The proof method based essentially on Bayesian estimation features in [10], which focuses on probabilistic statements and stops one step short of formulating the identity. In this paper we put it all together, explicitly formulate the identity in terms of ridge regression, and give a simple proof for the kernel case. The identity is obtained by calculating the likelihood in a Gaussian processes model by different ways. Another proof of this fact is given in unpublished technical report [18].

We use the identity to derive several inequalities providing upper bounds for the cumulative loss of ridge regression applied in the on-line fashion. Corollaries 2 and 3 deal with 'clipped' ridge regression. The later reproduces Theorem 4.6 from [2] (this result is often confused with Theorem 4 in [17], which, in fact, provides a similar bound for an essentially different algorithm). Corollary 4 (reproduced from [18]) shows that in the linear case the loss of (unclipped) on-line ridge regression is asymptotically close to the loss of a retrospectively best regressor.

In the literature there is a range of specially designed regression-type algorithms with better worst-case bounds or bounds covering wider cases. Aggregating algorithm regression (also known as Vovk-Azoury-Warmuth predictor) is described in [17], [2], and Section 11.8 of [6]. Theorem 1 in [17] provides an upper bound for aggregating algorithm regression, which is better than the bound given by Corollary 3 for clipped ridge regression. The bound from [17] has also been shown to be optimal. The exact relation between the performances of ridge

M. Hutter et al. (Eds.): ALT 2010, LNAI 6331, pp. 405–419, 2010.
© Springer-Verlag Berlin Heidelberg 2010

regression and aggregating algorithm regression is not known. Theorem 3 in [17] describes a case where aggregating algorithm regression performs better, but in the case of unbounded signals. An important class of regression-type algorithms achieving different bounds is based on the gradient descent idea; see [5], [11], and Section 11 in [6]. Algorithms in [8] and [4] provide regression-type algorithms dealing with changing dependencies.

The paper is organised as follows. Section 2 introduces kernels and kernel ridge regression in batch and on-line settings. We take the simplest approach and use an explicit formula to introduce ridge regression. Section 3 contains the statement of the identity and Section 4 discusses corollaries of the identity. The rest of the paper is devoted to the proof of the identity. Section 5 introduces a probabilistic interpretation of ridge regression in the context of Gaussian fields and Section 6 contains the proof. Section 7 contains an outline of an alternative proof based on the aggregating algorithm.

2 Kernel Ridge Regression in On-Line and Batch Settings

2.1 Kernels

A *kernel* on a domain X, which is an arbitrary set with no structure assumed, is a symmetric positive semi-definite function of two arguments, i.e., $\mathcal{K} : X \times X \to \mathbb{R}$ such that

1. for all $x_1, x_2 \in X$ we have $\mathcal{K}(x_1, x_2) = \mathcal{K}(x_2, x_1)$ and
2. for any positive integer T, any $x_1, x_2, \ldots, x_T \in X$ and any real numbers $\alpha_1, \alpha_2, \ldots, \alpha_T \in \mathbb{R}$ we have $\sum_{i,j=1}^{T} \mathcal{K}(x_i, x_j)\alpha_i\alpha_j \geq 0$.

An equivalent definition can be given as follows. There is a Hilbert space $\mathcal{F}$ of functions on X such that

1. for every $x \in X$ the function $\mathcal{K}(x, \cdot)$, i.e., $\mathcal{K}$ considered as a function of the second argument with the first argument fixed, belongs to $\mathcal{F}$ and
2. for every $x \in X$ and every $f \in \mathcal{F}$ the value of f at x equals the scalar product of f by $\mathcal{K}(x, \cdot)$, i.e., $f(x) = \langle f, \mathcal{K}(x, \cdot)\rangle_{\mathcal{F}}$; this property is often called the *reproducing property*.

The second definition is sometimes said to specify a *reproducing kernel*. The space $\mathcal{F}$ is called the *reproducing kernel Hilbert space (RKHS)* for the kernel $\mathcal{K}$ (it can be shown that the RKHS for a kernel $\mathcal{K}$ is unique). The equivalence of the two definitions is proven in [1].

2.2 Regression in Batch and On-Line Settings

Suppose that we are given a sample of pairs

$$S = ((x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)) \ ,$$

where all $x_t \in X$ are called *signals* and $y_t \in \mathbb{R}$ are called *outcomes* for the corresponding signals. A pair (x_t, y_t) is called *an example*.

The task of regression is to fit a function (usually from a particular class) to the data. The method of *kernel ridge regression* with a kernel $\mathcal{K}$ and a real regularisation parameter $a > 0$ suggests the function $f_{\mathrm{RR}}(x) = Y'(K + aI)^{-1}k(x)$, where $Y = (y_1, y_2, \ldots, y_T)'$ is the column vector of outcomes,

$$K = \begin{pmatrix} \mathcal{K}(x_1, x_1) & \mathcal{K}(x_1, x_2) & \ldots & \mathcal{K}(x_1, x_T) \\ \mathcal{K}(x_2, x_1) & \mathcal{K}(x_2, x_2) & \ldots & \mathcal{K}(x_2, x_T) \\ \vdots & \vdots & \ddots & \ldots \\ \mathcal{K}(x_T, x_1) & \mathcal{K}(x_T, x_2) & \ldots & \mathcal{K}(x_T, x_T) \end{pmatrix}$$

is the *kernel matrix* and

$$k(x) = \begin{pmatrix} \mathcal{K}(x_1, x) \\ \mathcal{K}(x_2, x) \\ \vdots \\ \mathcal{K}(x_T, x) \end{pmatrix} .$$

Note that the matrix K is positive-semidefinite by the definition of a kernel, therefore the matrix $K + aI$ is positive-definite and thus non-singular.

It is easy to see that $f_{\mathrm{RR}}(x)$ is a linear combination of functions $\mathcal{K}(x_t, x)$ (note that x does not appear outside of $k(x)$ in the ridge regression formula) and therefore it belongs to the RKHS $\mathcal{F}$ specified by the kernel $\mathcal{K}$. It can be shown that on this f the minimum of the expression $\sum_{t=1}^{T}(f(x) - y_t)^2 + a\|f\|_{\mathcal{F}}^2$ (where $\|\cdot\|_{\mathcal{F}}$ is the norm in $\mathcal{F}$) over all f from the RKHS $\mathcal{F}$ is achieved.

Suppose now that the sample is given to us example by example. For each example we are shown the signal and then asked to produce a prediction for the outcome. One can say that the learner operates according to the following protocol:

Protocol 1

```
for t = 1, 2, ...
  read signal x_t
  output prediction γ_t
  read true outcome y_t
endfor
```

This learning scenario is called *on-line* or *sequential*. The scenario when the whole sample is given to us at once as before is called *batch* to distinguish it from on-line.

One can apply ridge regression in the on-line scenario in the following natural way. On step t we form the sample S_t from the $t - 1$ known examples $(x_1, y_1), (x_2, y_2), \ldots, (x_{t-1}, y_{t-1})$ and output the prediction suggested by the ridge regression function for this sample.

For the on-line scenario we will use the same notations as in the batch mode but with the index t denoting the time. Thus K_t is the kernel matrix on step t

(note that its size is $(t-1)\times(t-1)$), Y_t is the vector of outcomes $y_1, y_2, \ldots, y_{t-1}$, and k_t is $k(x)$ for step t. We will be referring to the prediction output by on-line ridge regression on step t as γ_t^{RR}.

3 The Identity

Theorem 1. *Take a kernel $\mathcal{K}$ on a domain X and a parameter $a > 0$. Let $\mathcal{F}$ be the RKHS for the kernel $\mathcal{K}$. For a sample $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ let $\gamma_1^{\mathrm{RR}}, \gamma_2^{\mathrm{RR}}, \ldots, \gamma_T^{\mathrm{RR}}$ be the predictions output by ridge regression with the kernel $\mathcal{K}$ and the parameter a in the on-line mode. Then*

$$\sum_{t=1}^{T} \frac{(\gamma_t^{\mathrm{RR}} - y_t)^2}{1 + d_t/a} = \min_{f\in\mathcal{F}} \left(\sum_{t=1}^{T} (f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2 \right) = aY'(K_{T+1} + aI)^{-1}Y \ ,$$

where $d_t = \mathcal{K}(x_t, x_t) - k_t'(x_t)(K_t + aI)^{-1}k_t(x_t) > 0$ and all other notation is as above.

The left-hand side term in this equality is *close* to the cumulative squared loss of ridge regression in the on-line mode. The difference is in the denominators $1 + d_t/a$. The values d_t have the meaning of variances of ridge regression predictions according to the probabilistic view discussed below.

Note that the minimum in the middle term is attained on f specified by batch ridge regression knowing the whole sample. It is thus *nearly* the squared loss of the *retrospectively* best fit $f \in \mathcal{F}$.

The right-hand side term is a simple closed-form expression.

4 Corollaries

In this section we use the identity to obtain upper bounds on cumulative losses of on-line algorithms.

It is easy to obtain a basic multiplicative bound on the loss of on-line ridge regression. The matrix $(K_t + aI)^{-1}$ is positive-definite as the inverse of a positive-definite, therefore $k_t'(x_t)(K_t + aI)^{-1}k_t(x_t) \geq 0$ and $d_t \leq \mathcal{K}(x_t, x_t)$. Assuming that there is $c_{\mathcal{F}} > 0$ such that $\mathcal{K}(x, x) \leq c_{\mathcal{F}}^2$ on X (i.e., the evaluation functional on $\mathcal{F}$ is uniformly bounded by $c_{\mathcal{F}}$), we get

$$\sum_{t=1}^{T} (\gamma_t^{\mathrm{RR}} - y_t)^2 \leq \left(1 + \frac{c_{\mathcal{F}}^2}{a}\right) \min_{f\in\mathcal{F}} \left(\sum_{t=1}^{T} (f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2 \right) = a\left(1 + \frac{c_{\mathcal{F}}^2}{a}\right) Y'(K_{T+1} + aI)^{-1}Y \ . \quad (1)$$

More interesting bounds can be obtained on the following assumption. Suppose that we know in advance that outcomes y come from an interval $[-Y, Y]$, and Y is known to us. It does not make sense then to make predictions outside of the

interval. One may consider *clipped ridge regression*, which operates as follows. For a given signal the ridge regression prediction γ^{RR} is calculated; if it falls inside the interval, it is kept; if it is outside of the interval, it is replaced by the closest point from the interval. Denote the prediction of clipped ridge regression by $\gamma^{\mathrm{RR},Y}$. If $y \in [-Y, Y]$ indeed holds, then $(\gamma^{\mathrm{RR},Y} - y)^2 \le (\gamma^{\mathrm{RR}} - y)^2$ and $(\gamma^{\mathrm{RR},Y} - y)^2 \le 4Y^2$.

Corollary 2. *Take a kernel $\mathcal{K}$ on a domain X and a parameter $a > 0$. Let $\mathcal{F}$ be the RKHS for the kernel $\mathcal{K}$. For a sample $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ such that $y_t \in [-Y, Y]$ for all $t = 1, 2, \ldots, T$, let $\gamma_1^{\mathrm{RR},Y}, \gamma_2^{\mathrm{RR},Y}, \ldots, \gamma_T^{\mathrm{RR},Y}$ be the predictions output by clipped ridge regression with the kernel $\mathcal{K}$ and the parameter a in the on-line mode. Then*

$$\sum_{t=1}^{T} (\gamma_t^{\mathrm{RR},Y} - y_t)^2 \le \min_{f \in \mathcal{F}} \left(\sum_{t=1}^{T} (f(x_t) - y_t)^2 + a \|f\|_{\mathcal{F}}^2 \right) + 4Y^2 \ln \det \left(I + \frac{1}{a} K_{T+1} \right) \ ,$$

where K_{T+1} is as above.

Proof. We have

$$\frac{1}{1 + d_t/a} = 1 - \frac{d_t/a}{1 + d_t/a}$$

and

$$\frac{d_t/a}{1 + d_t/a} \le \ln(1 + d_t/a) \ ;$$

indeed, for $b \ge 0$ the inequality $b/(1+b) \le \ln(1+b)$ holds and can be checked by differentiation. Therefore

$$\begin{aligned} \sum_{t=1}^{T} (\gamma_t^{\mathrm{RR},Y} - y_t)^2 &= \sum_{t=1}^{T} (\gamma_t^{\mathrm{RR},Y} - y_t)^2 \frac{1}{1 + d_t/a} + \sum_{t=1}^{T} (\gamma_t^{\mathrm{RR},Y} - y_t)^2 \frac{d_t/a}{1 + d_t/a} \\ &\le \sum_{t=1}^{T} (\gamma_t^{\mathrm{RR}} - y_t)^2 \frac{1}{1 + d_t/a} + 4Y^2 \sum_{t=1}^{T} \ln(1 + d_t/a) \ . \end{aligned}$$

Lemma 7 proved below yields

$$\prod_{t=1}^{T} (1 + d_t/a) = \frac{1}{a^T} \det(K_{T+1} + aI) = \det \left(I + \frac{1}{a} K_{T+1} \right) \ .$$

□

There is no sublinear upper bound on the regret term $4Y^2 \ln \det \left(I + \frac{1}{a} K_{T+1} \right)$ in the general case; indeed, consider the kernel

$$\delta(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 = x_2; \\ 0 & \text{otherwise.} \end{cases}$$

However we can get good bounds in special cases.

It is shown in [16] that for the radial-basis kernel $\mathcal{K}(x_1, x_2) = e^{-b\|x_1 - x_2\|^2}$, where $x_1, x_2 \in \mathbb{R}^d$, we can get an upper bound on average. Suppose that all xs are independently identically distributed according to the Gaussian distribution with the mean of 0 and variance of cI. Then for the expectation we have $E \ln \det \left(I + \frac{1}{a} K_{T+1}\right) = O((\ln T)^{d+1})$ (see Section IV.B in [16]). This yields a bound on the expected loss of clipped ridge regression.

Consider the linear kernel $\mathcal{K}(x_1, x_2) = x_1' x_2$ defined on column vectors from $\mathbb{R}^n$. We have $\mathcal{K}(x, x) = \|x\|^2$, where $\|\cdot\|$ is the quadratic norm in $\mathbb{R}^n$. The reproducing kernel Hilbert space is the set of all linear functions on $\mathbb{R}^n$. We have $K_t = X_t' X_t$, where X_{T+1} is the *design matrix* made up of column vectors $x_1, x_2, \ldots, x_T$. The Sylvester determinant identity (see, e.g., [7]) implies that

$$\det\left(I + \frac{1}{a} X_{T+1}' X_{T+1}\right) = \det\left(I + \frac{1}{a} X_{T+1} X_{T+1}'\right) = \det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t x_t'\right) \quad .$$

Estimating the determinant by the product of its diagonal elements (see, e.g., Theorem 7 in Chapter 2 of [3]) and assuming that all coordinates of x_t are bounded by B, we get

$$\det\left(I + \frac{1}{a}\sum_{t=1}^{T} x_t x_t'\right) \le \left(1 + \frac{TB^2}{a}\right)^n \quad .$$

We get the following corollary.

Corollary 3. *For a sample $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$, where $x_t \in [-B, B]^n$ and $y_t \in [-Y, Y]$ for all $t = 1, 2, \ldots, T$, let $\gamma_1^{\mathrm{RR},Y}, \gamma_2^{\mathrm{RR},Y}, \ldots, \gamma_T^{\mathrm{RR},Y}$ be the predictions output by clipped linear ridge regression with a parameter $a > 0$ in the on-line mode. Then*

$$\sum_{t=1}^{T} (\gamma_t^{\mathrm{RR},Y} - y_t)^2 \le \min_{\theta \in \mathbb{R}^n} \left(\sum_{t=1}^{T} (\theta' x_t - y_t)^2 + a\|\theta\|^2\right) + 4Y^2 n \ln\left(1 + \frac{TB^2}{a}\right) \quad .$$

It is an interesting problem if the bound is optimal. As far as we know, there is a gap in existing bounds. Theorem 2 in [17] shows that $Y^2 n \ln T$ is a lower bound for *any* learner and in the constructed example $\|x_t\|_\infty = 1$. Theorem 3 in [17] provides a stronger lower bound, but at a cost of allowing unbounded xs.

For the linear kernel the expression d_t/a in the denominator of the identity can be rewritten as follows:

$$\begin{aligned} \frac{d_t}{a} &= \frac{1}{a}\left[\mathcal{K}(x_t, x_t) - k_t'(x_t)(K_t + aI)^{-1} k_t(x_t)\right] \\ &= \frac{1}{a}\left[x_t' x_t - (x_t' X_t)(X_t' X_t + aI)^{-1}(X_t' x_t)\right] \quad . \end{aligned}$$

We can apply the matrix identity $A(BA + I)^{-1} = (AB + I)^{-1} A$ (it holds if both the inversions can be performed and can be proven by multiplying both the sides by $BA + I$ and $AB + I$ and opening up the brackets) and further obtain

$$\begin{aligned}\frac{d_t}{a} &= \frac{1}{a}\left[x_t' x_t - x_t'(X_t X_t' + aI)^{-1} X_t X_t' x_t\right] \\ &= \frac{1}{a}\left[x_t'(I - (X_t X_t' + aI)^{-1} X_t X_t') x_t\right] \\ &= x_t'(X_t X_t' + aI)^{-1} x_t\end{aligned}$$

We will denote $X_t X_t' + aI$ by A_t. One can easily see that

$$A_t = aI + \sum_{i=1}^{t-1} x_i x_i' = a\sum_{i=1}^{n} e_i e_i' + \sum_{i=1}^{t-1} x_i x_i' \ ,$$

where e_i are unit vectors from the standard basis. If one assumes that the norms $\|x_t\|$, $t = 1, 2, \ldots$ are bounded, one can apply Lemma A.1 from [12] and infer that $x_t' A_t^{-1} x_t \to 0$ as $t \to \infty$. Note that this convergence does not hold in the general kernel case. Indeed, if $\mathcal{K} = \delta$ defined above and all x_t are different, we get $d_t = 1$.

The leftmost side of the identity is thus asymptotically close to the cumulative loss of on-line ridge regression and the regularised loss of the retrospectively best regressor in the linear case. We will reproduce a corollary from [18] formalising this intuition.

Corollary 4. *Let $x_t \in \mathbb{R}^n$, $t = 1, 2, \ldots$ and $\sup_{t=1,2,\ldots} \|x_t\| < \infty$; let γ_t^{RR} be the predictions output by on-line ridge regression with the linear kernel and a parameter $a > 0$. Then*

1. *if there is $\theta \in \mathbb{R}^n$ such that $\sum_{t=1}^{\infty}(y_t - \theta' x_t)^2 < +\infty$ then*

$$\sum_{t=1}^{\infty}(y_t - \gamma_t^{\mathrm{RR}})^2 < +\infty \ ;$$

2. *if for all $\theta \in \mathbb{R}^n$ we have $\sum_{t=1}^{\infty}(y_t - \theta' x_t)^2 = +\infty$, then*

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

Proof. Part 1 follows from bound (1).

Let us prove Part 2. First note that $x_t' A_t^{-1} x_t \geq 0$ implies

$$\sum_{t=1}^{T}(y_t - \gamma_t^{\mathrm{RR}})^2 \geq \sum_{t=1}^{T}\frac{(y_t - \gamma_t^{\mathrm{RR}})^2}{1 + x_t' A_t^{-1} x_t} = \min_{\theta\in\mathbb{R}^n}\left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2\right)$$

and thus the fraction in (2) is always greater than or equal to 1.

Let us show that $\min_{\theta\in\mathbb{R}^n}\left(\sum_{t=1}^{T}(y_t - \theta' x_t)^2 + a\|\theta\|^2\right) \to +\infty$ as $t \to \infty$. Suppose that this does not hold. Then there is a sequence T_k and θ_{T_k} such that the expressions $\sum_{t=1}^{T_k}(y_t - \theta_{T_k}' x_t)^2 + a\|\theta_{T_k}\|^2$ are bounded. Hence there

is $C < +\infty$ such that $\sum_{t=1}^{T_k}(y_t - \theta'_{T_k} x_t)^2 \leq C$ for all $k = 1, 2, \ldots$ and the norms of θ_{T_k} are also bounded uniformly in k. Therefore the sequence θ_{T_k} has a converging subsequence. Let θ_0 be the limit of this subsequence. Let us show that $\sum_{i=1}^{T_k}(y_t - \theta'_0 x_t)^2 \leq C$. Indeed, let $\sum_{i=1}^{T_k}(y_t - \theta'_0 x_t)^2 > C$. For sufficiently large m the sum $\sum_{i=1}^{T_k}(y_t - \theta'_{T_m} x_t)^2$ is sufficiently close to $\sum_{i=1}^{T_k}(y_t - \theta'_0 x_t)^2$ so that

$$\sum_{i=1}^{T_m}(y_t - \theta'_{T_m} x_t)^2 \geq \sum_{i=1}^{T_k}(y_t - \theta'_{T_m} x_t)^2 > C \ ,$$

which contradicts $\sum_{t=1}^{T_m}(y_t - \theta'_{T_m} x_t)^2 \leq C$. In the limit we get $\sum_{i=1}^{\infty}(y_t - \theta'_0 x_t)^2 \leq C < +\infty$, which contradicts the condition of Part 2.

Take $\varepsilon > 0$. There is T_0 such that for all $T \geq T_0$ we have $1 + x'_T A_T^{-1} x_T \leq 1 + \varepsilon$ and

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

Therefore for all sufficiently large T the fraction in (2) does not exceed $1 + \varepsilon$. □

5 Probabilistic Interpretation

We will prove the identity by means of the probabilistic interpretation of ridge regression.

Suppose that we have a Gaussian random field[1] z_x with the means of 0 and the covariances $\mathrm{cov}(z_{x_1}, z_{x_2}) = \mathcal{K}(x_1, x_2)$. Such a field exists. Indeed, for any finite set of $x_1, x_2, \ldots, x_T$ our requirements imply the Gaussian distribution with the mean of 0 and the covariance matrix of K. These distributions satisfy the consistency requirements and thus the Kolmogorov extension (or existence) theorem (see, e.g., [13], Appendix 1 for a proof sketch[2]) can be applied to construct a field over X.

Let ε_x be a Gaussian field of mutually independent and independent of z_x random values with the variance σ^2. The existence of such a field can be shown using the same Kolmogorov theorem. Now let $y_x = z_x + \varepsilon_x$. Intuitively, ε_x can

[1] We use the term 'field' rather than 'process' to emphasise the fact that X is not necessarily a subset of $\mathbb{R}$ and its elements do not have to be moments of time; some textbooks still use the word 'process' in this case.

[2] Strictly speaking, we do not need to construct the field for the whole X in order to prove the theorem; is suffices to consider a finite-dimensional Gaussian distribution of $(z_{x_1}, z_{x_2}, \ldots, z_{x_T})$.

be thought of as random noise introduced by measurements of the original field z_x.

The learning process can be thought of as estimating the values of the field y_t given the values of the field at sample points. One can show that the conditional distribution of z_x given a sample $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T))$ is Gaussian with the mean of $\gamma_x^{\mathrm{RR}} = Y'(K + \sigma^2 I)^{-1} k(x)$ and the variance $d_x = \mathcal{K}(x, x) - k'(x)(K + \sigma^2 I)^{-1} k(x)$. The conditional distribution of y_x is Gaussian with the same mean and the variance $\sigma^2 + \mathcal{K}(x, x) - k'(x)(K + \sigma^2 I)^{-1} k(x)$ (see [14], Section 2.2, p. 17).

If we let $a = \sigma^2$, we see that γ_t^{RR} and $a + d_t$ are, respectively, the mean and the variance of the conditional distributions for y_{x_t} given the sample S_t.

Remark 5. Note that in the statement of the theorem there is no assumption that the signals x_t are pairwise different. Some of them may coincide. In the probabilistic picture all xs must be different though, or the corresponding probabilities make no sense. This obstacle may be overcome in the following way. Let us replace the domain X by $X' = X \times \mathbb{N}$, where $\mathbb{N}$ is the set of positive integers $\{1, 2, \ldots\}$, and replace x_t by $x'_t = (x_t, t) \in X'$. For X' there is a Gaussian field with the covariance function $\mathcal{K}'((x_1, t_1), (x_2, t_2)) = \mathcal{K}(x_1, x_2)$. The argument concerning the probabilistic meaning of ridge regression stays for $\mathcal{K}'$ on X'. We can thus assume that all x_t are different.

6 Proof of the Identity

The proof is based on the Gaussian field interpretation. Let us calculate the density of the joint distribution of the variables $(y_{x_1}, y_{x_2}, \ldots, y_{x_T})$ at the point $(y_1, y_2, \ldots, y_T)$. We will do this in three different ways: by decomposing the density into a chain of conditional densities, marginalisation, and, finally, direct calculation. Each method will give us a different expression corresponding to a term in the identity. Since all the three terms express the same density, they must be equal.

6.1 Conditional Probabilities

We have

$$p_{y_{x_1}, y_{x_2}, \ldots, y_{x_T}}(y_1, y_2, \ldots, y_T) = p_{y_{x_T}}(y_T \mid y_{x_1} = y_1, y_{x_2} = y_2, \ldots, y_{x_{T-1}} = y_{T-1}) \cdot p_{y_{x_1}, y_{x_2}, \ldots, y_{x_{T-1}}}(y_1, y_2, \ldots, y_{T-1}) \ .$$

Expanding this further yields

$$p_{y_{x_1}, y_{x_2}, \ldots, y_{x_T}}(y_1, y_2, \ldots, y_T) = p_{y_{x_T}}(y_T \mid y_{x_1} = y_1, y_{x_2} = y_2, \ldots, y_{x_{T-1}} = y_{T-1}) \cdot p_{y_{x_{T-1}}}(y_T \mid y_{x_1} = y_1, y_{x_2} = y_2, \ldots, y_{x_{T-1}} = y_{T-2}) \cdots p_{y_{x_1}}(y_1) \ .$$

As we have seen before, the distribution for y_{x_t} given that $y_{x_1} = y_1, y_{x_2} = y_2, \ldots, y_{x_{t-1}} = y_{t-1}$ is Gaussian with the mean of γ_t^{RR} and the variance of $d_t + \sigma^2$. Thus

$$p_{y_{x_T}}(y_t \mid y_{x_1} = y_1, y_{x_2} = y_2, \ldots, y_{x_{t-1}} = y_{t-1}) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{d_t + \sigma^2}} e^{-\frac{1}{2}\frac{(y_t - \gamma_t^{\mathrm{RR}})^2}{d_t + \sigma^2}}$$

and

$$p_{y_{x_1}, y_{x_2}, \ldots, y_{x_T}}(y_1, y_2, \ldots, y_T) = \frac{1}{(2\pi)^{T/2}\sqrt{(d_1 + \sigma^2)(d_2 + \sigma^2)\ldots(d_T + \sigma^2)}} e^{-\frac{1}{2}\sum_{t=1}^{T}\frac{(\gamma_t^{\mathrm{RR}} - y_t)^2}{d_t + \sigma^2}} .$$

6.2 Dealing with Singular Kernel Matrix

The expression for the second case looks particularly simple for non-singular K. Let us show that this is sufficient to prove the identity.

All the terms in the identity are in fact some continuous functions of $T(T+1)/2$ values of $\mathcal{K}$ at the pairs of points x_i, x_j, $i, j = 1, 2, \ldots, T$. Indeed, the values of γ_t^{RR} in the left-hand side expression are ridge regression predictions given by respective analytic formula. Note that the coefficients of the inverse matrix are continuous functions of the original matrix.

The optimal function minimising the second expression is in fact $f_{\mathrm{RR}}(x) = \sum_{t=1}^{T} c_t \mathcal{K}(x_t, x)$, where the coefficients c_t are continuous functions of the values of $\mathcal{K}$. The reproducing property implies that

$$\|f_{\mathrm{RR}}\|^2 = \sum_{i,j=1}^{T} c_i c_j \langle \mathcal{K}(x_i, \cdot), \mathcal{K}(x_j, \cdot)\rangle_{\mathcal{F}} = \sum_{i,j=1}^{T} c_i c_j \mathcal{K}(x_i, x_j) \ .$$

We can thus conclude that all the expressions are continuous in the values of $\mathcal{K}$. Consider the kernel $\mathcal{K}_\alpha(x_1, x_2) = \mathcal{K}(x_1, x_2) + \alpha\delta(x_1, x_2)$, where

$$\delta(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 = x_2; \\ 0 & \text{otherwise} \end{cases}$$

and $\alpha > 0$. Clearly, δ is a kernel and thus $\mathcal{K}_\alpha$ is a kernel. If all x_t are different (recall Remark 5), kernel matrix for $\mathcal{K}_\alpha$ equals $K + \alpha I$ and therefore it is nonsingular.

However the values of $\mathcal{K}_\alpha$ tend to the corresponding values of $\mathcal{K}$ as $\alpha \to 0$.

6.3 Marginalisation

The method of marginalisation consists of introducing extra variables to obtain the joint density in some manageable form and then integrating over the

extra variables to get rid of them. The variables we are going to consider are $z_{x_1}, z_{x_2}, \ldots, z_{x_T}$.

Given the values of $z_{x_1}, z_{x_2}, \ldots, z_{x_T}$, the density of $y_{x_1}, y_{x_2}, \ldots, y_{x_T}$ is easy to calculate. Indeed, given zs all ys are independent and have the means of corresponding zs and variances of σ^2, i.e.,

$$p_{y_{x_1},y_{x_2},\ldots,y_{x_T}}(y_1, y_2, \ldots, y_T \mid z_{x_1} = z_1, z_{x_2} = z_2, \ldots, z_{x_{T-1}} = z_{T-1}) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{-\frac{1}{2}\frac{(y_1-z_1)^2}{\sigma^2}}\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{-\frac{1}{2}\frac{(y_2-z_2)^2}{\sigma^2}}\cdots\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{-\frac{1}{2}\frac{(y_T-z_T)^2}{\sigma^2}} = \frac{1}{(2\pi)^{T/2}\sigma^T}e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t-z_t)^2}$$

The density of $z_{x_1}, z_{x_2}, \ldots, z_{x_T}$ is given by

$$p_{z_{x_1},z_{x_2},\ldots,z_{x_T}}(z_1, z_2, \ldots, z_T) = \frac{1}{(2\pi)^{T/2}\sqrt{\det K_{T+1}}}e^{-\frac{1}{2}Z'K_{T+1}^{-1}Z} \ ,$$

where $Z = (z_1, z_2, \ldots, z_T)$, provided K_{T+1} is nonsingular.

Using

$$p_{y_{x_1},y_{x_2},\ldots,y_{x_T},z_{x_1},z_{x_2},\ldots,z_{x_T}}(y_1, y_2, \ldots, y_T, z_1, z_2, \ldots, z_T) = p_{y_{x_1},y_{x_2},\ldots,y_{x_T}}(y_1, y_2, \ldots, y_T \mid z_{x_1} = z_1, z_{x_2} = z_2, \ldots, z_{x_{T-1}} = z_{T-1})\cdot p_{z_{x_1},z_{x_2},\ldots,z_{x_T}}(z_1, z_2, \ldots, z_T)$$

and

$$p_{y_{x_1},y_{x_2},\ldots,y_{x_T}}(y_1, y_2, \ldots, y_T) = \int_{\mathbb{R}^T} p_{y_{x_1},y_{x_2},\ldots,y_{x_T},z_{x_1},z_{x_2},\ldots,z_{x_T}}(y_1, y_2, \ldots, y_T, z_1, z_2, \ldots, z_T)dZ$$

we get

$$p_{y_{x_1},y_{x_2},\ldots,y_{x_T}}(y_1, y_2, \ldots, y_T) = \frac{1}{(2\pi)^{T/2}\sigma^T}\frac{1}{(2\pi)^{T/2}\sqrt{\det K_{T+1}}}\int_{\mathbb{R}^T} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t-z_t)^2-\frac{1}{2}Z'K_{T+1}^{-1}Z}dZ \ .$$

To evaluate the integral we need the following lemma (see [3], Theorem 3 of Chapter 2) .

Lemma 6. *Let $Q(\theta)$ be a quadratic form of $\theta \in \mathbb{R}^n$ with the positive definite quadratic part, i.e., $Q(\theta) = \theta' A\theta + \theta' b + c$, where the matrix A is symmetric positive definite. Then*

$$\int_{\mathbb{R}^n} e^{-Q(\theta)}d\theta = e^{-Q(\theta_0)}\frac{\pi^{n/2}}{\sqrt{\det A}} \ ,$$

where $\theta_0 = \arg\min_{\mathbb{R}^n} Q$.

The quadratic part of the form in our integral has the matrix $\frac{1}{2}K_{T+1}^{-1} + \frac{1}{2\sigma^2}I$ and therefore

$$p_{y_{x_1},y_{x_2},\ldots,y_{x_T}}(y_1, y_2, \ldots, y_T) = \frac{1}{(2\pi)^T \sigma^T \sqrt{\det K_{T+1}}} \frac{\pi^{T/2}}{\sqrt{\det(\frac{1}{2}K_{T+1}^{-1} + \frac{1}{2\sigma^2}I)}} \times e^{-\min_Z\left(\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - z_t)^2 - \frac{1}{2}Z' K_{T+1}^{-1} Z\right)}$$

We have

$$\sqrt{\det K_{T+1}}\sqrt{\det\left(\frac{1}{2}K_{T+1}^{-1} + \frac{1}{2\sigma^2}I\right)} = \sqrt{\det\left(\frac{1}{2}I + \frac{1}{2\sigma^2}K_{T+1}\right)} = \frac{1}{2^{T/2}\sigma^T}\sqrt{\det(K_{T+1} + \sigma^2 I)} \ .$$

Let us deal with the minimum. We will link it to

$$M = \min_{f \in \mathcal{F}}\left(\sum_{t=1}^{T}(f(x_t) - y_t)^2 + \sigma^2 \|f\|_{\mathcal{F}}^2\right) \ .$$

The representer theorem implies that the minimum in the definition of M is achieved on f from the linear span of $\mathcal{K}(x_1, \cdot), \mathcal{K}(x_2, \cdot), \ldots, \mathcal{K}(x_T, \cdot)$, i.e., on a function of the form $f(x) = \sum_{t=1}^{T} c_t \mathcal{K}(x_t, \cdot)$. For the column vector $Z(x) = (f(x_1), f(x_2), \ldots, f(x_T))'$ we have $Z(x) = K_{T+1}C$, where $C = (c_1, c_2, \ldots, c_T)'$. Since K_{T+1} is supposed to be non-singular, there is a one-to-one correspondence between C and $Z(x)$; we have $C = K_{T+1}^{-1}Z(x)$ and $\|f\|_{\mathcal{F}}^2 = C' K_{T+1} C = Z'(x) K_{T+1}^{-1} Z(x)$. Thus

$$\min_Z\left(\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - z_t)^2 + \frac{1}{2}Z' K_{T+1}^{-1} Z\right) = \frac{1}{2\sigma^2}M \ .$$

For the density we get the expression

$$p_{y_{x_1},y_{x_2},\ldots,y_{x_T}}(y_1, y_2, \ldots, y_T) = \frac{1}{(2\pi)^{T/2}\sqrt{\det(K_{T+1} + \sigma^2 I)}} e^{-\frac{1}{2\sigma^2}M} \ .$$

6.4 Direct Calculation

One can easily calculate the covariances of ys:

$$\begin{aligned} \operatorname{cov}(y_{x_1}, y_{x_2}) &= E(z_{x_1} + \varepsilon_{x_1})(z_{x_2} + \varepsilon_{x_2}) \\ &= E z_{x_1} z_{x_2} + E\varepsilon_{x_1}\varepsilon_{x_2} \\ &= \mathcal{K}(x_1, x_2) + \sigma^2 \delta(x_1, x_2) \ . \end{aligned}$$

Therefore, one can write down the expression

$$p_{y_{x_1},y_{x_2},\ldots,y_{x_T}}(y_1,y_2,\ldots,y_T) = \frac{1}{(2\pi)^{T/2}\sqrt{\det(K_{T+1}+\sigma^2 I)}} e^{-\frac{1}{2}Y'_{T+1}(K_{T+1}+\sigma^2 I)^{-1}Y_{T+1}} \ .$$

6.5 Equating the Terms

It remains to take the logarithms of the densities calculated in different ways. We need the following matrix lemma.

Lemma 7

$$(d_1+\sigma^2)(d_2+\sigma^2)\ldots(d_T+\sigma^2) = \det(K_{T+1}+\sigma^2 I)$$

Proof. The lemma follows from Frobenius's identity (see, e.g., [7]):

$$\det\begin{pmatrix} A & u \\ v' & d \end{pmatrix} = (d - v'A^{-1}u)\det A \ ,$$

where d is a scalar and the submatrix A is non-singular.

We have

$$\begin{aligned}\det(K_{T+1}+\sigma^2 I) &= (\mathcal{K}(x_T,x_T)+\sigma^2 - k'_T(x_T)(K_T+\sigma^2 I)^{-1}k_T(x_T))\cdot \\ &\quad \det(K_T+\sigma^2 I) \\ &= (d_T+\sigma^2)\det(K_T+\sigma^2 I) \\ &= \ldots \\ &= (d_T+\sigma^2)(d_{T-1}+\sigma^2)\ldots(d_2+\sigma^2)(d_1+\sigma^2) \ . \end{aligned}$$ □

We get

$$\sum_{t=1}^{T}\frac{(\gamma_t^{\mathrm{RR}}-y_t)^2}{d_t+\sigma^2} = \frac{1}{\sigma^2}M = Y'(K_{T+1}+\sigma^2 I)^{-1}Y \ .$$

The theorem follows.

7 Alternative Derivations for the Linear Case

In this section we outline alternative ways of obtaining the identity in the linear case.

A Gaussian field z_x with the covariance function x'_1x_2 on $\mathbb{R}^n$ can be obtained as follows. Let θ be an n-dimensional Gaussian random variable with the mean of 0 and the covariance matrix I; let $z_x = \theta' x$ and $y_x = z_x + \varepsilon_x$, where ε_x is independent Gaussian with the mean of 0 and the variance of σ^2 (recall that we let $\sigma^2 = a$). Estimating y_x given a sample of pairs (x_t, y_t) can be thought of as going from the prior distribution for θ to a posterior distribution. The learning process described in Section 5 can thus be thought of as Bayesian estimation. It

can be performed in an on-line fashion (the term 'sequential' is more common in Bayesian statistics): the posterior distribution serves as the prior for the next step. This procedure leads to the Gaussian distribution for y with the mean equal to the on-line ridge regression prediction. The linear case is thus a special case of the kernel case.

There is an entirely different way to look at this procedure; it is based on the aggregating algorithm (described, e.g., in [17]). Consider the following game between a predictor and the reality. On step t the reality produces x_t; the predictor sees it and outputs a prediction, which is a Gaussian distribution on $\mathbb{R}$ with the density function p_t. Then the reality announces y_t and the predictor suffers loss $-\ln p_t(y_t)$. Suppose that there is a set of experts who play the same game and we are able to see their predictions before making ours. The aim of aggregating algorithm is to merge experts' predictions so as to suffer cumulative loss comparable to that of the best expert. The game we have described happens to be perfectly mixable, so the merging can be done relatively easily.

Let us consider a pool of experts $\mathcal{E}_\theta$, $\theta \in \mathbb{R}^n$, such that on step t expert $\mathcal{E}_\theta$ outputs the Gaussian distribution with the mean of $\theta' x_t$ and the variance σ^2. The aggregating algorithm requires a prior distribution on the experts. Let us take the Gaussian distribution with the mean of 0 and the covariance matrix I. The distribution is updated on each step; one can show that the update corresponds to the Bayesian update of the distribution for θ. Finally, it is possible to show that the distribution output by the aggregating algorithm on step t is the Gaussian distribution with the mean $\gamma_t = Y_t X_t A_t^{-1} x_t$, i.e., the ridge regression prediction, and the variance $\sigma^2 x_t A_t^{-1} x_t + \sigma^2$, i.e., the conditional variance of y_t in the estimation procedure.

The equality between the first two terms in the identity from Theorem 1 can be derived from a fundamental property of the aggregating algorithm, namely, Lemma 1 in [17], which links the cumulative loss of the predictor to experts' losses. For more details see [18].

An advantage of this approach is that we do not need to consider random fields, estimation, prior and posterior distributions etc. All probabilities are no more than weights or predictions. This is arguably more intuitive.

Acknowledgements

The authors have been supported through the EPSRC grant EP/F002998 'Practical competitive prediction'. The first author has also been supported by an ASPIDA grant from the Cyprus Research Promotion Foundation.

The authors are grateful to Vladimir Vovk and Alexey Chernov for useful discussions and to anonymous COLT and ALT reviewers for detailed comments.

References

[1] Aronszajn, N.: La théorie des noyaux reproduisants et ses applications. Première partie. Proceedings of the Cambridge Philosophical Society 39, 133–153 (1943)

[2] Azoury, K.S., Warmuth, M.K.: Relative loss bounds for on-line density estimation with the exponential family of distributions. Machine Learning 43, 211–246 (2001)

[3] Beckenbach, E.F., Bellman, R.E.: Inequalities. Springer, Heidelberg (1961)
[4] Busuttil, S., Kalnishkan, Y.: Online regression competitive with changing predictors. In: Proceedings of Algorithmic Learning Theory, 18th International Conference, pp. 181–195 (2007)
[5] Cesa-Bianchi, N., Long, P., Warmuth, M.K.: Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. IEEE Transactions on Neural Networks 7, 604–619 (1996)
[6] Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, Cambridge (2006)
[7] Henderson, H.V., Searle, S.R.: On deriving the inverse of a sum of matrices. SIAM Review 23(1) (1981)
[8] Herbster, M., Warmuth, M.K.: Tracking the best linear predictor. Journal of Machine Learning Research 1, 281–309 (2001)
[9] Hoerl, A.E.: Application of ridge analysis to regression problems. Chemical Engineering Progress 58, 54–59 (1962)
[10] Kakade, S.M., Seeger, M.W., Foster, D.P.: Worst-case bounds for Gaussian process models. In: Proceedings of the 19th Annual Conference on Neural Information Processing Systems (2005)
[11] Kivinen, J., Warmuth, M.K.: Exponentiated gradient versus gradient descent for linear predictors. Infornation and Computation 132(1), 1–63 (1997)
[12] Kumon, M., Takemura, A., Takeuchi, K.: Sequential optimizing strategy in multi-dimensional bounded forecasting games. CoRR abs/0911.3933v1 (2009)
[13] Lamperti, J.: Stochastic Processes: A Survey of the Mathematical Theory. Springer, Heidelberg (1977)
[14] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
[15] Saunders, C., Gammerman, A., Vovk, V.: Ridge regression learning algorithm in dual variables. In: Proceedings of the 15th International Conference on Machine Learning, pp. 515–521 (1998)
[16] Seeger, M.W., Kakade, S.M., Foster, D.P.: Information consistency of nonparametric Gaussian process methods. IEEE Transactions on Information Theory 54(5), 2376–2382 (2008)
[17] Vovk, V.: Competitive on-line statistics. International Statistical Review 69(2), 213–248 (2001)
[18] Zhdanov, F., Vovk, V.: Competing with gaussian linear experts. CoRR abs/0910.4683 (2009)