An Identity for Kernel Ridge Regression[☆]

Fedor Zhdanov[a], Yuri Kalnishkan[a]

[a] *Computer Learning Research Centre and Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, United Kingdom*

Abstract

This paper derives an identity connecting the square loss of ridge regression in on-line mode with the loss of the retrospectively best regressor. Some corollaries about the properties of the cumulative loss of on-line ridge regression are also obtained.

1. Introduction

Ridge regression is a powerful technique of machine learning. It was introduced in [2]; the kernel version of it is derived in [3].

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 [4] for the linear case (it can be obtained by summing (4.21) from [4] in an exact rather than estimated form). In [5] Bayesian estimation was applied to the analysis of regression in a fashion very similar to this paper. However [5] focuses on probabilistic statements and stops one step short of formulating the identity. The right-hand side of the identity providing a short explicit formula was obtained in [6] (see Lemma 14).

In this paper we put it all together, explicitly formulate the identity in terms of ridge regression, and give two proofs for the kernel case. The first proof obtains the terms of the identity calculating the same likelihood in a Gaussian processes model by three different methods. Remarkably, a probabilistic argument yields a result that holds in the worst case along any sequence of signals and outcomes with no probabilistic assumptions. The other proof is based on the analysis of a Bayesian-type algorithm for prediction with expert advice; it is reproduced from unpublished technical report [1].

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 the 'clipped' ridge regression. The later reproduces Theorem 4.6 from [4] (this result is often confused with Theorem 4 in [7], which, in fact, provides a similar bound for an essentially different algorithm). Corollary 4 shows that for continuous kernels on compact domains the loss of (unclipped) on-line ridge regression is asymptotically close to the loss of the retrospectively best regressor. This result cannot be generalised to non-compact domains.

In the literature there is a range of specially designed regression-type algorithms with better worst-case bounds or bounds applicable to more general scenarios. Aggregating algorithm regression (also known as Vovk-Azoury-Warmuth predictor) is described in [7], [4], and Section 11.8 of [8]. Theorem 1 in [7] provides an upper bound for aggregating algorithm regression; the bound is better than the bound given by Corollary 3 for clipped ridge regression. The bound from [7] has also been shown to be optimal in a strong sense. The exact relation between the performance of ridge regression and the performance of aggregating algorithm regression is not known. Theorem 3 in [7] provides an example where aggregating algorithm regression

[☆] Earlier versions of this paper appeared in Proceedings of ALT 2010, LNCS 6331, Springer, 2010 and as technical report abs/1112.1390 at *CoRR*. This paper also reproduces some results from technical report [1].

Email addresses: fedor@cs.rhul.ac.uk (Fedor Zhdanov), yura@cs.rhul.ac.uk (Yuri Kalnishkan)

performs better, but the signals in the example are unbounded. An important class of regression-type algorithms achieving different bounds is based on the gradient descent idea; see [9], [10], and Section 11 in [8]. In [11] and [12] regression-type algorithms dealing with changing dependencies are constructed. In [13] regression is considered within the framework of discounted loss, which decays with time.

The paper is organised as follows. Section 2 introduces kernels and kernel ridge regression in batch and on-line settings. We use an explicit formula to introduce ridge regression; Appendix A contains a proof that this formula specifies a function with a certain optimality property. Section 3 contains the statement of the identity and Subsection 3.1 shows that the identity remains true (in a way) for the case of zero ridge.

Section 4 discusses corollaries of the identity. Section 5 contains the proof based on a probabilistic interpretation of ridge regression in the context of Gaussian fields. Section 6 contains an alternative proof based on prediction with expert advice. The proof has been reproduced from [1].

Appendixes A–C to the paper contain proofs of some known results; they have been included for completeness and to clarify the intuition behind other proofs in the paper. Appendix D contains a technical lemma.

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-semidefinite 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 $x_1, x_2, \dots, x_T \in X$ and any real numbers $\alpha_1, \alpha_2, \dots, \alpha_T \in \mathbb{R}$ 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. A function $\mathcal{K} : X \times X \to \mathbb{R}$ is a kernel if 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 [14].

2.2. Regression in Batch and On-line Settings

Suppose that we are given a sample of pairs (sometimes called *a training set*[1])

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

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

[1] Strictly speaking it is an array rather than a set: the same pair may appear in the training set twice so that $((x_0, y_0))$ and $((x_0, y_0), (x_0, y_0))$ are two essentially different training sets.

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 (*ridge*) $a > 0$ suggests the function $f_{\mathrm{RR}}(x) = Y'(K + aI)^{-1}k(x)$, where $Y = (y_1, y_2, \dots, y_T)'$ is the column vector[2] of outcomes,

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

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.

If the sample S is empty, i.e., $T = 0$ or no examples have been given to us yet, we assume that $f_{\mathrm{RR}}(x) = 0$ for all x.

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 f_{RR} the minimum of the expression $\sum_{t=1}^{T}(f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2$ (where $\|\cdot\|_{\mathcal{F}}$ is the norm in $\mathcal{F}$) over all f from the RKHS $\mathcal{F}$ is achieved. For completeness, we include a proof in Appendix A.

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-1} from the $t-1$ known examples $(x_1, y_1), (x_2, y_2), \dots, (x_{t-1}, y_{t-1})$ and output the prediction suggested by the ridge regression function for this sample.

For the on-line scenario on step t we will use the same notation as in the batch mode but with the index $t-1$ denoting the time[3]. Thus K_{t-1} is the kernel matrix on step t (its size is $(t-1) \times (t-1)$), Y_{t-1} is the vector of outcomes $(y_1, y_2, \dots, y_{t-1})'$, and $k_{t-1}(x_t) = (\mathcal{K}(x_1, x_t), \mathcal{K}(x_2, x_t), \dots, \mathcal{K}(x_{t-1}, x_t))'$ is $k(x_t)$ 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 $(x_1, y_1), (x_2, y_2), \dots, (x_T, y_T)$ let $\gamma_1^{\mathrm{RR}}, \gamma_2^{\mathrm{RR}}, \dots, \gamma_T^{\mathrm{RR}}$ be the predictions output by ridge regression*

[2]Throughout this paper M' denotes the transpose of a matrix M.

[3]The conference version of this paper used t rather than $t-1$. This paper uses $t-1$ because it coincides with the size and for compatibility with earlier papers.

with the kernel $\mathcal{K}$ and the parameter a in the on-line mode. Then

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

where $d_t = \mathcal{K}(x_t, x_t) - k_{t-1}'(x_t)(K_{t-1} + aI)^{-1} k_{t-1}(x_t) \geq 0$ and all other notation is as above.

The left-hand side term in this equality is *close* to the cumulative square 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. Lemma 2 shows that $d_t \to 0$ as $t \to \infty$ for continuous kernels on compact domains. The terms of the identity thus become close to the cumulative square loss asymptotically; this intuition is formalised by Corollary 4.

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 square loss of the *retrospectively* best fit $f \in \mathcal{F}$.

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

3.1. The Case of Zero Ridge

In this subsection we show that the identity essentially remains true for $a = 0$.

Let the parameter a in the identity approach 0. One may think that the third term of the identity should tend to zero. On the other hand, the value of the middle term of the identity for $a = 0$ depends on Y_T; the values of y_t can be chosen (at least in some cases) so that there is no exact fit in the RKHS (i.e., no $f \in \mathcal{F}$ such that $f(x_t) = y_t$, $t = 1, 2, \ldots, T$) and the middle term is not equal to 0. This section resolves the apparent contradiction.

As a matter of fact, the limit of the identity as $a \to 0$ does not have to be 0. The situation when there is no exact fit in the RKHS is only possible when the matrix K_T is singular, and in this situation the right-hand side does not always tend to 0.

The expression on the left-hand side of the identity is formally undefined for $a = 0$. The expression on the right-hand side is undefined when $a = 0$ and K_T is singular. The expression in the centre, by contrast, always makes sense. The following theorem clarifies the situation.

Corollary 1. *Under the conditions of Theorem 1, as $a \to 0$, the terms of the identity*

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

tend to the squared norm of the projection of the vector Y_T to the null space of the matrix K_T. This coincides with the value of the middle term of the identity for $a = 0$.

The *null space* (also called the *kernel*) of a matrix S is the space of vectors x such that $Sx = 0$. It is easy to see that the dimension of the null space and the rank of S (equal to the dimension of the span of the columns of S) sum up to the number of columns of S. If, moreover, S is square and symmetric, the null space of S is the orthogonal complement of the span of the columns of S.

PROOF. For every $a \geq 0$ let $m_a = \inf_{f \in \mathcal{F}} \left(\sum_{t=1}^{T} (f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2 \right)$. Proposition 2 implies that if $a > 0$ then the infimum is achieved on the ridge regression function with the parameter a. Throughout this proof we will denote this function by f_a.

Let us calculate the value of $m_0 = \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (f(x_t) - y_t)^2$. It follows from the representer theorem (see Proposition 3) that it is sufficient to consider the functions f of the form $f(\cdot) = \sum_{i=1}^{T} c_i \mathcal{K}(x_i, \cdot)$.

For $f(\cdot) = \sum_{i=1}^{T} c_i \mathcal{K}(x_i, \cdot)$ the sum $\sum_{t=1}^{T} (f(x_t) - y_t)^2$ equals the squared norm $\|K_T C - Y_T\|^2$, where $C = (c_1, c_2, \ldots, c_T)'$ is the vector of coefficients of the linear combination. If C_0 minimises this expression, then $K_T C_0$ is the projection of Y_T to the linear span of the columns of K_T. The vector $Y_T - K_T C_0$ is then

the projection of Y_T to the orthogonal complement of the span and the orthogonal complement is the null space of K_T.

Let us show that m_a is continuous at $a = 0$. Fix some f_0 such that the infimum m_0 is achieved on f_0 (if K_T is singular there can be more than one such function). Substituting f_0 into the formula for m_a yields $m_a \le m_0 + a\|f_0\|^2_{\mathcal{F}} = m_0 + o(1)$ as $a \to 0$. Substituting f_a into the definition of m_0 yields $m_0 \le m_a$. We thus get that $m_a \to m_0$ as $a \to 0$. □

4. Corollaries

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

4.1. A Multiplicative Bound

It is easy to obtain a basic multiplicative bound on the loss of on-line ridge regression. The matrix $(K_{t-1} + aI)^{-1}$ is positive-definite as the inverse of a positive-definite. Therefore $k'_{t-1}(x_t)(K_{t-1} + aI)^{-1}k_{t-1}(x_t) \ge 0$ and $d_t \le \mathcal{K}(x_t, x_t)$. Assuming that there is $c_{\mathcal{F}} > 0$ such that $\mathcal{K}(x, x) \le c^2_{\mathcal{F}}$ on X (i.e., the evaluation functional on $\mathcal{F}$ is uniformly bounded by $c_{\mathcal{F}}$), we get

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

4.2. Additive Bounds for Clipped Regression

Some less trivial bounds can be obtained under 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*

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

where K_T 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 4 proved below yields

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

□

There is no sub-linear upper bound on the regret term

$$4Y^2 \ln\det\left(I + \frac{1}{a}K_T\right)$$

in the general case; indeed, consider the kernel

$$\delta(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 = x_2; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

However we can get good bounds in special cases.

It is shown in [15] that for the Gaussian 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 $E\ln\det\left(I + \frac{1}{a}K_T\right) = O((\ln T)^{d+1})$ (see Section IV.B in [15]). 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 is the *design matrix* made up of column vectors $x_1, x_2, \ldots, x_T$. The Sylvester determinant identity $\det(I + UV) = \det(I + VU)$ (see, e.g., [16], Eq. (6)) implies that

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

We will use an upper bound from [17] for this determinant[4]; a proof is given in Appendix C for completeness. We get the following corollary.

Corollary 3. *For a sample $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$, where $\|x_t\| \le B$ 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*

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

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 [7] 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 [7] provides a stronger lower bound, but at the cost of allowing unbounded xs.

4.3. An Asymptotic Comparison

The inequalities we have considered so far hold for finite time horizons T. We shall now let T tend to infinity.

Let us analyse the behaviour of the quantity

$$d_t = \mathcal{K}(x_t, x_t) - k_{t-1}'(x_t)(K_{t-1} + aI)^{-1} k_{t-1}(x_t) \quad .$$

[4]The authors would like to thank the anonymous reviewer who suggested this bound strengthening the corollary.

According to the probabilistic interpretation discussed in Subsection 5.1, d_t has the meaning of the variance of the prediction output by kernel ridge regression on step t and therefore it is always non-negative.

The probabilistic interpretation suggests that the variance should go down with time as we learn the data better. In the general case this is not true. Indeed, if $\mathcal{K} = \delta$ (defined by (2)) and all x_i are different then $d_t = 1$ for all $t = 1, 2, \ldots$ However under natural assumptions that hold in most reasonable cases the following lemma holds. The lemma generalises Lemma A.1 from [18] because for the linear kernel d_t can be represented as shown in (14) below.

Lemma 2. *Let X be a compact metric space and a kernel $\mathcal{K} : X^2 \to \mathbb{R}$ be continuous in both arguments. Then for any sequence $x_1, x_2, \ldots \in X$ and $a > 0$ we have $d_t \to 0$ as $t \to \infty$.*

Proof. As discussed in Subsection 5.1, d_t has the meaning of a variance under a certain probabilistic interpretation and therefore $d_t \geq 0$. One can easily see that $k'_{t-1}(x_t)(K_{t-1} + aI)^{-1}k_{t-1}(x_t) \geq 0$. Indeed, the matrix $(K_{t-1} + aI)^{-1}$ is positive-definite as the inverse of a positive-definite. We get

$$0 \leq k'_{t-1}(x_t)(K_{t-1} + aI)^{-1}k_{t-1}(x_t) \leq \mathcal{K}(x_t, x_t) \ . \tag{3}$$

Let us start by considering a special case. Suppose that the sequence $x_1, x_2, \ldots$ converges and let $\lim_{t\to\infty} x_t = x_0 \in X$. The continuity of $\mathcal{K}$ implies that $\mathcal{K}(x_t, x_t) \to \mathcal{K}(x_0, x_0)$ and

$$0 \leq k'_{t-1}(x_t)(K_{t-1} + aI)^{-1}k_{t-1}(x_t) \leq \mathcal{K}(x_0, x_0) + o(1) \tag{4}$$

as $t \to \infty$. We will obtain a lower bound on $k'_{t-1}(x_t)(K_{t-1} + aI)^{-1}k_{t-1}(x_t)$ and show that it converges to $\mathcal{K}(x_0, x_0)$ thus proving the lemma in the special case.

For every symmetric matrix M and a vector x of the matching size we have $\lambda_{\min}\|x\|^2 \leq x'Mx$, where $\lambda_{\min}$ is the smallest eigenvalue of M (this can be shown by considering an orthonormal base where M diagonalises). The smallest eigenvalue of $(K_{t-1} + aI)^{-1}$ equals $1/(\tilde{\lambda} + a)$, where $\tilde{\lambda}$ is the largest eigenvalue of K_{t-1}. The value of $\tilde{\lambda}$ is bounded from above by the trace of K_{t-1}:

$$\tilde{\lambda} \leq \sum_{i=1}^{t-1} \mathcal{K}(x_i, x_i)$$

and this yields a lower bound on the smallest eigenvalue of $(K_{t-1} + aI)^{-1}$.

The squared norm of k_{t-1} equals

$$\|k_{t-1}(x_t)\|^2 = \sum_{i=1}^{t-1} (\mathcal{K}(x_i, x_t))^2 \ .$$

Combining this with the above estimates we get

$$k'_{t-1}(x_t)(K_{t-1} + aI)^{-1}k_{t-1}(x_t) \geq \frac{\sum_{i=1}^{t-1}(\mathcal{K}(x_i, x_t))^2}{a + \sum_{i=1}^{t-1}\mathcal{K}(x_i, x_i)} \ .$$

Let us assume $\mathcal{K}(x_0, x_0) \neq 0$ and show that the right-hand size of the inequality tends to $\mathcal{K}(x_0, x_0)$ (if $\mathcal{K}(x_0, x_0) = 0$, then (4) implies that $k'_{t-1}(x_t)(K_{t-1} + aI)^{-1}k_{t-1}(x_t) \to 0$). Dividing the numerator and the denominator by $t - 1$ yields

$$\frac{\sum_{i=1}^{t-1}(\mathcal{K}(x_i, x_t))^2}{a + \sum_{i=1}^{t-1}\mathcal{K}(x_i, x_i)} = \frac{\frac{\sum_{i=1}^{t-1}(\mathcal{K}(x_i, x_t))^2}{t-1}}{\frac{a}{t-1} + \frac{\sum_{i=1}^{t-1}\mathcal{K}(x_i, x_i)}{t-1}} \ .$$

Clearly, as t goes to infinity, most terms in the sums become arbitrary close to $(\mathcal{K}(x_0, x_0))^2$ and $\mathcal{K}(x_0, x_0)$ and thus the 'averages' tend to $(\mathcal{K}(x_0, x_0))^2$ and $\mathcal{K}(x_0, x_0)$, respectively. Therefore the fraction tends to

$\mathcal{K}(x_0, x_0)$ and (4) implies that $k'_{t-1}(x_t)(K_{t-1} + aI)^{-1}k_{t-1}(x_t) \to \mathcal{K}(x_0, x_0)$. We have shown that $d_t \to 0$ for the special case when the sequence $x_1, x_2, \ldots$ converges.

Let us prove the lemma for an arbitrary sequence $x_1, x_2, \ldots \in X$. If $d_t \not\to 0$, there is a subsequence of indexes $\tau_1 < \tau_2 < \ldots < \tau_k < \ldots$ such that d_{τ_k} is separated from 0. Since X is compact, we can choose a sub-subsequence $t_1 < t_2 < \ldots < t_n < \ldots$ such that d_{t_n} is separated from 0 and x_{t_n} converges. If we show that $d_{t_n} \to 0$, we get a contradiction and prove the lemma. Thus it is sufficient to show that $d_{t_n} \to 0$ where $\lim_{n\to\infty} x_{t_n} = x_0 \in X$.

Clearly, we have inequalities (4) for $t = t_n$:

$$0 \le k'_{t_n-1}(x_{t_n})(K_{t_n-1} + aI)^{-1}k_{t_n-1}(x_{t_n}) \le \mathcal{K}(x_0, x_0) + o(1) \ . \tag{5}$$

We proceed by obtaining a lower bound on the middle term as before.

Fix some n and the corresponding t_n. One can rearrange the order of the elements of the finite sequence $x_1, x_2, \ldots, x_{t_n-1}$ to put the elements of the subsequence to the front and consider the sequence (still of length $t_n - 1$) $x_{t_1}, x_{t_2}, \ldots, x_{t_{n-1}}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{t_n-n}$, where $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{t_n-n}$ are the elements of the original sequence with indexes not in the set $\{t_1, t_2, \ldots, t_{n-1}\}$.

One can write

$$k'_{t_n-1}(x_{t_n})(K_{t_n-1} + aI)^{-1}k_{t_n-1}(x_{t_n}) = \tilde{k}'_{t_n-1}(x_{t_n})(\widetilde{K}_{t_n-1} + aI)^{-1}\tilde{k}_{t_n-1}(x_{t_n}) \ ,$$

where

$$\tilde{k}'_{t_n-1}(x_{t_n}) = \begin{pmatrix} \mathcal{K}(x_{t_1}, x_{t_n}) \\ \vdots \\ \mathcal{K}(x_{t_{n-1}}, x_{t_n}) \\ \mathcal{K}(\bar{x}_1, x_{t_n}) \\ \vdots \\ \mathcal{K}(\bar{x}_{t_n-n}, x_{t_n}) \end{pmatrix}$$

and

$$\widetilde{K}_{t_n-1} = \begin{pmatrix} \mathcal{K}(x_{t_1}, x_{t_1}) & \ldots & \mathcal{K}(x_{t_1}, x_{t_{n-1}}) & \mathcal{K}(x_{t_1}, \bar{x}_1) & \ldots & \mathcal{K}(x_{t_1}, \bar{x}_{t_n-n}) \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathcal{K}(x_{t_{n-1}}, x_{t_1}) & \ldots & \mathcal{K}(x_{t_{n-1}}, x_{t_{n-1}}) & \mathcal{K}(x_{t_{n-1}}, \bar{x}_1) & \ldots & \mathcal{K}(x_{t_{n-1}}, \bar{x}_{t_n-n}) \\ \mathcal{K}(\bar{x}_1, x_{t_1}) & \ldots & \mathcal{K}(\bar{x}_1, x_{t_{n-1}}) & \mathcal{K}(\bar{x}_1, \bar{x}_1) & \ldots & \mathcal{K}(\bar{x}_1, \bar{x}_{t_n-n}) \\ \vdots & & \vdots & \vdots & & \vdots \\ \mathcal{K}(\bar{x}_{t_n-n}, x_{t_1}) & \ldots & \mathcal{K}(\bar{x}_{t_n-n}, x_{t_{n-1}}) & \mathcal{K}(\bar{x}_{t_n-n}, \bar{x}_1) & \ldots & \mathcal{K}(\bar{x}_{t_n-n}, \bar{x}_{t_n-n}) \end{pmatrix} .$$

Indeed, we can consider the matrix product $k_{t_n-1}(x_{t_n})(K_{t_n-1}+aI)^{-1}k_{t_n-1}(x_{t_n})$ in the rearranged orthonormal base where the base vectors with the indexes $t_1, t_2, \ldots, t_{n-1}$ are at the front of the list. (Alternatively one can check that rearranging the elements of the training set does not affect ridge regression prediction and its variance.)

The upper left corner of $\widetilde{K}_{t_n-1}$ and the upper part of $\tilde{k}_{t_n-1}(x_{t_n})$ consist of values of the kernel on elements of the subsequence, $\mathcal{K}(x_{t_i}, x_{t_j}), i, j = 1, 2, \ldots, n$. We shall use this observation and reduce the proof to the special case considered above.

Let us single out the left upper corner of size $(n-1) \times (n-1)$ in $\widetilde{K}_{t_n-1} + aI$ and apply Lemma 10 from Appendix D. The special case considered above implies that $\tilde{k}'_{t_n-1}(x_{t_n})(\widetilde{K}_{t_n-1} + aI)^{-1}\tilde{k}_{t_n-1}(x_{t_n}) \ge \mathcal{K}(x_0, x_0) + o(1)$ as $n \to \infty$. Combined with (5) this proves that $d_{t_n} \to 0$ as $n \to 0$. □

Remark 1. Lemma 10 from Appendix D implies (and is essentially equivalent) to the following statement in terms of the probabilistic interpretation from Subsection 5.1. Let y_x be a Gaussian field on a domain X with the means of 0. Let a sample of pairs $(x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_n, y'_n) \in X \times \mathbb{R}$ contain all the pairs from a sample $(x''_1, y''_1), (x''_2, y''_2), \ldots, (x''_m, y''_m) \in X \times \mathbb{R}$. Then the conditional variance of y_x given that $y_{x'_1} = y'_1$, $y_{x'_2} = y'_2$, …, and $y_{x'_n} = y'_n$ does not exceed the conditional variance of of y_x given that $y_{x''_1} = y''_1$, $y_{x''_2} = y''_2$, …, and $y_{x''_m} = y''_m$. (Note that by Remark 3 we can always assume that all x_t are different.)

We shall apply Lemma 2 to establish asymptotic equivalence between the losses in on-line and batch cases. The following corollary generalises Corollary 3 from [1].

Corollary 4. *Let X be a compact metric space and a kernel $\mathcal{K} : X^2 \to \mathbb{R}$ be continuous in both arguments; let $\mathcal{F}$ be the RKHS corresponding to the kernel $\mathcal{K}$. For a sequence $(x_1, y_1), (x_2, y_2), \ldots \in X \times \mathbb{R}$ let γ_t^{RR} be the predictions output by on-line ridge regression with a parameter $a > 0$. Then*

1. *if there is $f \in \mathcal{F}$ such that $\sum_{t=1}^{\infty}(y_t - f(x_t))^2 < +\infty$ then*

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

2. *if for all $f \in \mathcal{F}$ we have $\sum_{t=1}^{\infty}(y_t - f(x_t))^2 = +\infty$, then*

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T}(y_t - \gamma_t^{\mathrm{RR}})^2}{\min_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(y_t - f(x_t))^2 + a\|f\|_{\mathcal{F}}^2\right)} = 1 \ . \tag{6}$$

Proof. Part 1 follows from bound (1). Indeed, the continuous function $\mathcal{K}(x, x)$ is uniformly bounded on X and one can take a finite constant $c_{\mathcal{F}}$.

Let us prove Part 2. We observed above that $d_t \geq 0$. The identity implies

$$\sum_{t=1}^{T}(y_t - \gamma_t^{\mathrm{RR}})^2 \geq \sum_{t=1}^{T}\frac{(y_t - \gamma_t^{\mathrm{RR}})^2}{1 + d_t/a} = \min_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(y_t - f(x_t))^2 + a\|f\|_{\mathcal{F}}^2\right)$$

and thus the fraction in (6) is always greater than or equal to 1.

Let $m_t = \min_{f\in\mathcal{F}}\left(\sum_{\tau=1}^{t}(y_\tau - f(x_\tau))^2 + a\|f\|_{\mathcal{F}}^2\right)$. The sequence m_t is non-decreasing. Indeed let the minimum in the definition of m_t be achieved on f_t. If $m_{t+1} < m_t$ then one can substitute f_{t+1} into the definition of m_t and decrease the minimum.

Let us prove that $m_t \to +\infty$ as $t \to \infty$. A monotonic sequence must have a limit; let $\lim_{t\to\infty} m_t = m_\infty$. We have $m_1 \leq m_2 \leq \ldots \leq m_\infty$. We will assume that $m_\infty < +\infty$ and find $f_\infty \in \mathcal{F}$ such that $\sum_{t=1}^{\infty}(y_t - f_\infty(x_t))^2 \leq m_\infty < +\infty$ contrary to the condition of Part 2.

Proposition 2 implies that $f_t(\cdot)$ is the ridge regression function and belongs to the linear span of $\mathcal{K}(x_i, \cdot)$, $i = 1, 2, \ldots, t$, which we will denote by X_t. (For uniformity let $X_0 = \varnothing$ and $m_0 = 0$.) The squared norm of f_t does not exceed $m_t/a \leq m_\infty/a$. Thus all f_t belong to the ball of radius $\sqrt{m_\infty/a}$ centred at the origin.

Let X_∞ be the closure of the linear span of $\bigcup_{t=0}^{\infty} X_t$. If X_∞ happens to have a finite dimension, then all f_t belong to a ball in a finite-dimensional space; this ball is a compact set. If X_∞ is of infinite dimension, the ball is not compact but we will construct a different compact set containing all f_t.

Take $0 \leq s < t$. The function f_t can be uniquely decomposed as $f_t = g + h$, where g belongs to X_s and h is orthogonal to X_s. The Pythagoras theorem implies that $\|f\|_{\mathcal{F}}^2 = \|g\|_{\mathcal{F}}^2 + \|h\|_{\mathcal{F}}^2$. The function $h(\cdot)$ is orthogonal to all $\mathcal{K}(x_i, \cdot)$, $i = 1, 2, \ldots, s$; thus $h(x_i) = \langle h, \mathcal{K}(x_i, \cdot)\rangle = 0$ and $f_t(x_i) = g(x_i)$, $i = 1, 2, \ldots, s$ (recall the proof of the representer theorem, Proposition 3). Note that g cannot outperform f_s, which achieves the minimum m_s. We get

$$\begin{aligned} m_t &= \sum_{i=1}^{t}(y_i - f_t(x_i))^2 + a\|f_t\|_{\mathcal{F}}^2 \\ &= \sum_{i=1}^{s}(y_i - g(x_i))^2 + a\|g\|_{\mathcal{F}}^2 + \sum_{i=s+1}^{t}(y_i - f_t(x_i))^2 + a\|h\|_{\mathcal{F}}^2 \\ &\geq m_s + \sum_{i=s+1}^{t}(y_i - f_t(x_i))^2 + a\|h\|_{\mathcal{F}}^2 \ . \end{aligned}$$

This inequality implies that $\|h\|_{\mathcal{F}}^2 \le (m_t - m_s)/a \le (m_\infty - m_s)/a$.

Consider the set B of functions $f \in X_\infty \subseteq \mathcal{F}$ satisfying the following property for every $s = 0, 1, 2, \ldots$: let $f = g + h$ be the unique decomposition such that $g \in X_s$ and h is orthogonal to X_s; then the norm of h satisfies $\|h\|_{\mathcal{F}}^2 \le (m_\infty - m_s)/a$.

We have shown that all f_t belong to B, $t = 1, 2, \ldots$ Let us show that B is compact. It is closed because the projection in Hilbert spaces is a continuous operator. Let us show that B is totally bounded. We shall fix $\varepsilon > 0$ and construct a finite ε-net of points in B such that B is covered by closed balls of radius ε centred at the points from the net.

There is $s > 0$ such that $(m_\infty - m_s)/a \le \varepsilon^2/2$ because $m_s \to m_\infty$. The ball of radius $\sqrt{(m_\infty - m_0)/a}$ in X_s is compact and therefore it contains a finite $\varepsilon/\sqrt{2}$-net $g_1, g_2, \ldots, g_k$. Every $f \in B$ can be represented as $f = g + h$, where g belongs to X_s and h is orthogonal to X_s. Since $\|g\|_{\mathcal{F}}^2 \le \|f\|_{\mathcal{F}}^2 \le (m_\infty - m_0)/a$, the function g belongs to the ball of radius $\sqrt{(m_\infty - m_0)/a}$ in X_s and therefore $\|g - g_i\|_{\mathcal{F}} \le \varepsilon/\sqrt{2}$ for some g_i from the $\varepsilon/\sqrt{2}$-net. The definition of B implies that $\|h\|_{\mathcal{F}}^2 \le (m_\infty - m_s)/a \le \varepsilon^2/2$. The Pythagoras theorem yields

$$\|f - g_i\|_{\mathcal{F}}^2 = \|g - g_i\|_{\mathcal{F}}^2 + \|h\|_{\mathcal{F}}^2 \le \varepsilon^2/2 + \varepsilon^2/2 = \varepsilon^2 \ .$$

Thus the net we have constructed is an ε-net for B.

Since the functions f_t belong to a compact set, there is a converging sub-sequence f_{t_k}; let $\lim_{k\to\infty} f_{t_k} = f_\infty$. We have $\sum_{t=1}^{\infty}(y_t - f_\infty(x_t))^2 + a\|f_\infty\|_{\mathcal{F}}^2 \le m_\infty$. Indeed, if $\sum_{t=1}^{\infty}(y_t - f_\infty(x_t))^2 + a\|f_\infty\|_{\mathcal{F}}^2 > m_\infty$ then for a sufficiently large T_0 we have $\sum_{t=1}^{T_0}(y_t - f_\infty(x_t))^2 + a\|f_\infty\|_{\mathcal{F}}^2 > m_\infty$. Since $f_{t_k} \to f_\infty$ we get $f_{t_k}(x) \to f_\infty(x)$ for all $x \in X$ and for sufficiently large k all $f_{t_k}(x_t)$ are sufficiently close to $f_\infty(x_t)$, $t = 1, 2, \ldots, T_0$ and $\|f_{t_k}\|_{\mathcal{F}}$ is sufficiently close to $\|f_\infty\|_{\mathcal{F}}$ so that $\sum_{t=1}^{T_0}(y_t - f_{t_k}(x_t))^2 + a\|f_{t_k}\|_{\mathcal{F}}^2 > m_\infty$.

We have proved that under the conditions of Part 2 we have $m_t \to +\infty$ as $t \to \infty$.

Take $\varepsilon > 0$. Since by Lemma 2 we have $d_T \to 0$, there is T_0 such that for all $T \ge T_0$ we have $1 + d_T/a \le 1 + \varepsilon$ and

$$\begin{aligned}
\sum_{t=1}^{T}(y_t - \gamma_t^{\mathrm{RR}})^2 &= \sum_{t=1}^{T_0}(y_t - \gamma_t^{\mathrm{RR}})^2 + \sum_{t=T_0+1}^{T}(y_t - \gamma_t^{\mathrm{RR}})^2 \\
&\le \sum_{t=1}^{T_0}(y_t - \gamma_t^{\mathrm{RR}})^2 + (1+\epsilon)\sum_{t=1}^{T}\frac{(y_t - \gamma_t^{\mathrm{RR}})^2}{1 + d_t/a} \\
&= \sum_{t=1}^{T_0}(y_t - \gamma_t^{\mathrm{RR}})^2 + (1+\epsilon)\min_{f\in\mathcal{F}}\left(\sum_{t=1}^{T}(y_t - f(x_t))^2 + a\,\|f\|_{\mathcal{F}}^2\right) \ .
\end{aligned}$$

Therefore for all sufficiently large T the fraction in (6) does not exceed $1 + 2\varepsilon$. □

Remark 2. The proof of compactness above is based on the following general result (cf. [19], Chapter 4, exercise 7 on p. 172). Let B be a subset of l_2. Then B is totally bounded if and only if there is a sequence of nonnegative numbers $\alpha_1, \alpha_2, \ldots \ge 0$ converging to 0, i.e., $\lim_{t\to\infty}\alpha_t = 0$, such that for every $x = (x_1, x_2, \ldots) \in B$ and every $t = 1, 2, \ldots$ the inequality $\sum_{i=t}^{\infty} x_i^2 \le \alpha_t$ holds. This result generalises the well-known construction of the Hilbert cube (also known as the Hilbert brick).

The corollary does not hold for a non-compact domain. Let us construct a counterexample.

Let X be the unit ball in l_2, i.e., $X = \{x \in l_2 \mid \|x\|_{l_2} = 1\}$. Let the kernel on X be the scalar product in l_2, i.e., for $u = (u_1, u_2, \ldots)$ and $v = (v_1, v_2, \ldots)$ from X we have $\mathcal{K}(u, v) = \langle u, v\rangle_{l_2} = \sum_{i=1}^{\infty} u_i v_i$.

Consider the following sequence of elements $x_t \in X$. Let $x_{2i-1} = x_{2i}$ have one at position i and zeroes elsewhere, $i = 1, 2, \ldots$ Consider the sequence of outcomes where odd elements equal 1 and even elements equal 0, i.e., $y_{2i-1} = 1$ and $y_{2i} = 0$ for $i = 1, 2, \ldots$ We get

t	x_t	y_t
1	$(1,0,0,\ldots)$	1
2	$(1,0,0,\ldots)$	0
3	$(0,1,0,\ldots)$	1
4	$(0,1,0,\ldots)$	0
	$\vdots$	

Fix $a > 0$. Let us work out the predictions γ_t^{RR} output by on-line ridge regression on this sequence. The definition implies that $\gamma_1^{\mathrm{RR}} = 0$ and $\gamma_2^{\mathrm{RR}} = 1/(1+a)$. To obtain further predictions we need the following lemma stating that examples with signals orthogonal to *all other* signals and x_0 where we want to obtain a prediction can be dropped from the sample.

Lemma 3. *Let $\mathcal{K} : X \times X \to \mathbb{R}$ be a kernel on a domain X; let $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)) \in (X \times \mathbb{R})^*$ be a sample of pairs and let $x_0 \in X$. If there is a subset $(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2}), \ldots, (x_{i_k}, y_{i_k})$ of S such that the signals of the examples from this subset are orthogonal w.r.t. $\mathcal{K}$ to all other signals, i.e., $\mathcal{K}(x_{i_j}, x_m) = 0$ for all $j = 1, 2, \ldots, k$ and $m \neq i_1, i_2, \ldots, i_k$, and orthogonal to x_0 w.r.t. $\mathcal{K}$, i.e., $\mathcal{K}(x_{i_j}, x_0) = 0$ for all $j = 1, 2, \ldots, k$, then all elements of this subset can be removed from the sample S without affecting the ridge regression prediction $f_{\mathrm{RR}}(x_0)$.*

Proof. Let the subset coincide with the whole of S. Then $k(x_0) = 0$ and the ridge regression formula implies that ridge regression outputs $\gamma = 0$. Dropping the whole sample S leads to the same prediction by definition. For the rest of the proof assume that the subset is proper.

The main part of the proof relies on the optimality of ridge regression given by Proposition 2. Let $\mathcal{F}$ be the RKHS of functions on X corresponding to $\mathcal{K}$. The ridge regression function for the sample S minimises $\sum_{t=1}^{T}(f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2$ and by the representer theorem (Proposition 3) it is a linear combination of $\mathcal{K}(x_i, \cdot)$, $i = 1, 2, \ldots, T$.

Let us represent a linear combination f as $f_1 + f_2$, where f_1 is a linear combination of $\mathcal{K}(x_{i_j}, \cdot)$, $j = 1, 2, \ldots, k$ corresponding to signals from the subset and f_2 is a linear combination of the remaining signals. The functions f_1 and f_2 are orthogonal in $\mathcal{F}$ and this representation is unique. For every $j = 1, 2, \ldots, k$ and $m \neq i_1, i_2, \ldots, i_k$ we have $f(x_{i_j}) = f_1(x_{i_j})$ and $f(x_m) = f_2(x_m)$ and therefore

$$\sum_{t=1}^{T}(f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2 = \sum_{j=1}^{k}(f_1(x_{i_j}) - y_{i_j})^2 + \sum_{m \neq i_1, i_2, \ldots, i_k}(f_2(x_m) - y_m)^2 + a\|f_1\|_{\mathcal{F}}^2 + a\|f_2\|_{\mathcal{F}}^2 \ .$$

This expression splits into two terms depending only on f_1 and f_2. We can minimise it independently over f_1 and f_2. Note that $f_1(x_0) = 0$ by assumption and therefore $f_{\mathrm{RR}}(x_0) = \tilde{f}_2(x_0)$, where $\tilde{f}_2$ minimises

$$\sum_{m \neq i_1, i_2, \ldots, i_k}(f(x_m) - y_m)^2 + a\|f\|_{\mathcal{F}}^2$$

over $\mathcal{F}$. The optimality property implies that $\tilde{f}_2$ is the ridge regression function for the smaller sample. □

The lemma implies that $\gamma_{2i-1} = \gamma_1 = 0$ and $\gamma_{2i} = \gamma_2 = 1/(1+a)$ for all $i = 1, 2, \ldots$ It is easy to see that Corollary 4 is violated. For $f_0 = 0$ we have

$$\frac{\sum_{t=1}^{2i}(\gamma_t^{\mathrm{RR}} - y_i)^2}{\sum_{t=1}^{2i}(f_0(x_t) - y_t)^2} = \frac{i(1 + 1/(1+a)^2)}{i} = 1 + \frac{1}{(1+a)^2} > 1 \ .$$

The actual minimiser[5] gives an even smaller denominator and an even larger fraction.

We have shown that compactness is necessary in Corollary 4. It is easy to modify the counterexample to show that compactness without the continuity of $\mathcal{K}$ is not sufficient. Indeed, take an arbitrary compact

[5]It can easily be calculated but we do not really need it.

metric space X containing an infinite sequence $\tilde{x}_0, \tilde{x}_1, \tilde{x}_2, \ldots$ where $\tilde{x}_i \neq \tilde{x}_j$ for $i \neq j$. Let $\Phi : X \to l_2$ be such that $\tilde{x}_i$ is mapped to x_i from the counterexample for every $i = 1, 2, \ldots$ and x_0 is mapped to 0. Define the kernel $\mathcal{K}$ on X^2 by $\mathcal{K}(u, v) = \langle \Phi(u), \Phi(v) \rangle_{l_2}$ (this kernel cannot be continuous). Take y_i as in the counterexample. All predictions and losses on $(\tilde{x}_1, y_1), (\tilde{x}_2, y_2), \ldots$ will be as in the counterexample with $\mathcal{K}(x_0, \cdot)$ playing the part of f_0.

5. Gaussian Fields and a Proof of the Identity

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

5.1. Probabilistic Interpretation

Suppose that we have a Gaussian random field[6] z_x with the means of 0 and the covariances $\operatorname{cov}(z_{x_1}, z_{x_2}) = \mathcal{K}(x_1, x_2)$. 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., [20], Appendix 1 for a proof sketch[7]) 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 mean of 0 and 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 be thought of as random noise introduced by measurements of the original field z_x. The field z_x is not observable directly and we can possibly obtain only the values of y_x.

The learning process can be thought of as estimating the values of the field y_x 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 [21], 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 3. 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.

The proof of the identity 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.

5.2. Conditional Probabilities

We have

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

[6] 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.

[7] 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 $(z_{x_1}, z_{x_2}, \ldots, z_{x_T})$.

Expanding this further yields

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

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

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

5.3. Dealing with a 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 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 δ is as in (2) and $\alpha>0$. Clearly, δ is a kernel and thus $\mathcal{K}_\alpha$ is a kernel. If all x_t are different (recall Remark 3), the kernel matrix for $\mathcal{K}_\alpha$ equals $K+\alpha I$ and therefore it is non-singular.

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

5.4. 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.,

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

Since $z_{x_1}, z_{x_2},\ldots,z_{x_T}$ have a joint Gaussian distribution with the mean of 0 and covariance matrix K_T, their density is given by

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

where $Z = (z_1, z_2, \ldots, z_T)'$, provided K_T is non-singular.

Using

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

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

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

To evaluate the integral we need the following proposition (see [22], Theorem 3 of Chapter 2) .

Proposition 1. *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 $\frac{1}{2}K_T^{-1} + \frac{1}{2\sigma^2} I$ and therefore

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

We have

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

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 (see Proposition 3) implies that the minimum from the definition of M is achieved 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 $Z(x) = K_T C$, where $C = (c_1, c_2, \ldots, c_T)'$. Since K_T is supposed to be non-singular, there is a one-to-one correspondence between C and $Z(x)$; we have $C = K_T^{-1} Z(x)$ and $\|f\|_{\mathcal{F}}^2 = C' K_T C = Z'(x) K_T^{-1} Z(x)$. We can minimise by Z instead of C and therefore

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

For the density we get the expression

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

5.5. Direct Calculation

One can easily calculate the covariances of ys:

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

Therefore, one can write down the expression

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

5.6. Equating the Terms

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

Lemma 4.

$$(d_1 + \sigma^2)(d_2 + \sigma^2)\ldots(d_T + \sigma^2) = \det(K_T + \sigma^2 I)$$

Proof. The lemma follows from Frobenius's identity (see, e.g., [16]):

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

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

We have

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

□

We get

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

The theorem follows.

6. Bayesian Merging Algorithm and an Alternative Proof of the Identity

In this section we reproduce an alternative way (after [1]) of obtaining the identity.

An advantage of this approach is that we do not need to consider random fields. The use of probability is minimal; all probabilities in this approach are no more than weights or predictions. This provides an additional intuition to the proof.

6.1. Prediction with Expert Advice

Consider the standard prediction with expert advice framework. Let outcomes $y_1, y_2, \ldots$ from an *outcome set* Ω occur successively in discrete time. A *learner* tries to predict each outcome and outputs a prediction γ_t from a *prediction set* Γ each time before it sees the outcome y_t. There is also a pool Θ of *experts*; experts try to predict the outcomes from the same sequence and their predictions γ_t^θ are made available to the learner. The quality of predictions is assessed by means of a *loss function* $\lambda : \Gamma \times \Omega \to [0, +\infty]$.

The framework can be summarised in the following protocol:

Protocol 2.

`for` $t = 1, 2, \ldots$
  `experts` $\theta \in \Theta$ `announce predictions` $\gamma_t^\theta \in \Gamma$
  `learner outputs` $\gamma_t \in \Gamma$
  `reality announces` $y_t \in \Omega$
  `each expert` $\theta \in \Theta$ `suffers loss` $\lambda(\gamma_t^\theta, y_t)$
  `learner suffers loss` $\lambda(\gamma_t, y_t)$
`endfor`

The goal of the learner in this framework is to suffer the cumulative loss $\mathrm{Loss}_T = \sum_{t=1}^T \lambda(\gamma_t, y_t)$ not much larger than the cumulative loss of each expert $\mathrm{Loss}_T(\theta) = \sum_{t=1}^T \lambda(\gamma_t^\theta, y_t)$.

In this paper we consider the game with the outcome set $\Omega = \mathbb{R}$ and the prediction set Γ of all continuous density functions on $\mathbb{R}$, i.e., continuous functions $\xi : \mathbb{R} \to [0, +\infty)$ such that $\int_{-\infty}^{+\infty} \xi(y)dy = 1$. The loss function is negative logarithmic likelihood, i.e., $\lambda(\xi, y) = -\ln \xi(y)$.

6.2. Bayesian Merging Algorithm

Consider the following merging algorithm for the learner. The algorithm takes an initial distribution P_0 on the pool of experts Θ as a parameter and maintains weights P_t for experts θ.

Protocol 3.

`let` $P_0^* = P_0$
`for` $t = 1, 2, \ldots$
  `read experts' predictions` $\xi_t^\theta \in \Gamma$, $\theta \in \Theta$
  `predict` $\xi_t = \int_\Theta \xi_t^\theta P_{t-1}^*(d\theta)$
  `read` y_t
  `update the weights` $P_t(d\theta) = \xi_t^\theta(y_t) P_{t-1}(d\theta)$
  `normalise the weights` $P_t^*(d\theta) = P_t(d\theta) / \int_\Theta P_t(d\theta)$
`endfor`

If we consider an expert θ as a probabilistic hypothesis, this algorithm becomes the Bayesian strategy for merging hypotheses. The weights P_t^* relate to P_{t-1}^* as posterior probabilities to prior probabilities assigned to the hypotheses. We will refer to the algorithm as the Bayesian Algorithm (BA).

The algorithm can also be considered as a special case of the Aggregating Algorithm ([23, 24], see also [7]) going back to [25]. It is easy to check that the Aggregating Algorithm for these outcome set, prediction set, and the loss function and the learning rate $\eta = 1$ reduces to Protocol 3. However we will not be using the results proved for the Aggregating Algorithm in this paper.

After t steps the weights become

$$P_t(d\theta) = e^{-\mathrm{Loss}_t(\theta)} P_0(d\theta) \ . \tag{7}$$

The following lemma is a special case of Lemma 1 in Vovk [7]. It shows that the cumulative loss of the BA is an average of the experts' cumulative losses in a generalised sense (as in, e.g., Chapter 3 of [26]).

Lemma 5. *For any prior P_0 and any $t = 1, 2, \ldots$, the cumulative loss of the BA can be expressed as*

$$\mathrm{Loss}_t = -\ln \int_\Theta e^{-\mathrm{Loss}_t(\theta)} P_0(d\theta).$$

PROOF. The proof is by induction on t. For $t = 0$ the equality is obvious and for $t > 0$ we have

$$\begin{aligned}\operatorname{Loss}_t = \operatorname{Loss}_{t-1} - \ln \xi_t(y_t) &= -\ln \int_\Theta e^{-\operatorname{Loss}_{t-1}(\theta)} P_0(d\theta) - \ln \int_\Theta \xi_t^\theta(y_t) \frac{e^{-\operatorname{Loss}_{t-1}(\theta)}}{\int_\Theta e^{-\operatorname{Loss}_{t-1}(\theta)} P_0(d\theta)} P_0(d\theta) \\ &= -\ln \int_\Theta e^{-(-\ln \xi_t^\theta(y_t) + \operatorname{Loss}_{t-1}(\theta))} P_0(d\theta) = -\ln \int_\Theta e^{-\operatorname{Loss}_t(\theta)} P_0(d\theta)\end{aligned}$$

(the second equality follows from the inductive assumption, the definition of ξ_t, and (7)). □

6.3. Linear Ridge Regression as a Mixture

The above protocols can incorporate signals as in Protocol 1. Indeed let the reality announce a signal x_t on each step t; the signal can be used by both the experts and the learner.

Suppose that signals come from $\mathbb{R}^n$. Take a pool of *Gaussian experts* $\Theta = \mathbb{R}^n$. Fix some $\sigma > 0$ and let expert θ output the density of Gaussian distribution $\mathcal{N}(\theta' x_t, \sigma^2)$, i.e.,

$$\xi_t^\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta' x_t - y)^2}{2\sigma^2}} \quad , \tag{8}$$

on step t.

Let us assume the multivariate Gaussian distribution $\mathcal{N}(0, I)$ with the density

$$p_0(\theta) = \frac{1}{(2\pi)^{n/2}} e^{-\|\theta\|^2/2} \tag{9}$$

as the initial distribution over the pool of experts. We will show that the learner using the Bayesian merging algorithm with this initial distribution will be outputting a Gaussian density with the mean of the ridge regression prediction. Note that there is no assumption on the mechanism generating outcomes y_t.

Let Y_t be the vector of outcomes $y_1, y_2, \dots, y_t$. Let X_t be the design matrix made up of column vectors $x_1, x_2, \dots, x_t$ and $A_t = X_t X_t' + \sigma^2 I$, $t = 1, 2, \dots$.

Lemma 6. *The learner using the Bayesian merging algorithm with the initial distribution (9) on the pool of experts $\mathbb{R}^n$ predicting according to (8) will be outputting on step $T = 1, 2, \dots$ the density*

$$\xi_T(y) = \frac{1}{\sqrt{2\pi\sigma_T^2}} e^{-\frac{(\gamma_T^{\mathrm{RR}} - y)^2}{2\sigma_T^2}} \quad ,$$

where

$$\begin{aligned}\gamma_T^{\mathrm{RR}} &= Y_{T-1}' X_{T-1}' A_{T-1}^{-1} x_T \\ \sigma_T^2 &= \sigma^2 x_T' A_{T-1}^{-1} x_T + \sigma^2 \quad .\end{aligned}$$

We have $\gamma_T^{\mathrm{RR}} = \left(\theta_T^{\mathrm{RR}}\right)' x_T$, where $\theta_T^{\mathrm{RR}} = A_{T-1}^{-1} X_{T-1} Y_{T-1}$. At θ_T^{RR} the minimum

$$\min_{\theta \in \mathbb{R}^n} \left(\sum_{t=1}^{T-1} (\theta' x_t - y_t)^2 + \sigma^2 \|\theta\|^2 \right)$$

is achieved. This can be checked directly by differentiation or by reducing to Proposition 2 (see Subsection 6.5 for a discussion of linear ridge regression as a special case of kernel ridge regression). We will refer to the function $\left(\theta_T^{\mathrm{RR}}\right)' x$ as the *linear ridge regression* with the parameter σ^2. We are considering the on-line mode, but linear ridge regression can also be applied in the batch mode just like the general kernel ridge regression.

Let us prove the lemma.

PROOF. To evaluate the integral

$$\xi_T(v) = \int_{\mathbb{R}^n} \xi_T^\theta(v) P_{T-1}^*(d\theta) \tag{10}$$

we will use a probabilistic interpretation.

Let θ be a random value distributed according to P_{T-1}^*, i.e., having the density

$$p_\theta(u) \sim e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{T-1}(u'x_t - y_t)^2 - \frac{1}{2}\|u\|^2} \ .$$

Clearly, θ has a multivariate Gaussian distribution. The mean of a Gaussian distribution coincides with its mode and thus the mean of θ equals

$$\theta_T^{\mathrm{RR}} = \arg\min_{u\in\mathbb{R}^n} \left(\sum_{t=1}^{T-1} (u'x_t - y_t)^2 + \sigma^2 \|u\|^2 \right) = A_{T-1}^{-1} X_{T-1} Y_{T-1} \ .$$

The covariance matrix

$$\Sigma = \left(\frac{1}{\sigma^2} \sum_{t=1}^{T-1} x_t x_t' + I \right)^{-1} = \sigma^2 A_{T-1}^{-1}$$

can be obtained by singling out the quadratic part of the quadratic form in u.

Let y be a random value given by $y = \theta' x_T + \varepsilon$, where ε is independent of θ and has a Gaussian distribution with the mean of 0 and variance of σ^2. Clearly, given that $\theta = u$, the distribution of y is $\mathcal{N}(u'x_T, \sigma^2)$. The marginal density of y is just $\xi_T(v)$ we need to evaluate.

We will use the following statement from [27], Section 2.3.3. Let η have the (multivariate) Gaussian distribution $\mathcal{N}(\mu, \Lambda^{-1})$ and ζ have the (multivariate) Gaussian distribution $\mathcal{N}(A\eta + b, L^{-1})$, where A is a fixed matrix and b is a fixed vector. Then the marginal distribution of ζ is $\mathcal{N}(A\mu + b, L^{-1} + A\Lambda^{-1}A')$.

We get that the mean of y is $x_T' \theta_T^{\mathrm{RR}}$ and the variance is $\sigma^2 + x_T' \sigma^2 A_{T-1}^{-1} x_T$. □

The lemma is essentially equivalent to the following statement from Bayesian statistics. Let $y_t = x_t'\theta + \varepsilon_t$, where ε_t are independent Gaussian values with the means of 0 and variances σ^2, and x_t are not stochastic. Let the prior distribution for θ be $\mathcal{N}(0, I)$. Then the distribution for y_T given the observations $x_1, y_1, x_2, y_2, \ldots, x_{T-1}, y_{T-1}, x_T$ is $\mathcal{N}(\gamma_T^{\mathrm{RR}}, \sigma_T)$; see, e.g., [27], Section 3.3.2 or [28].

6.4. The Identity in the Linear Case

The following theorem is a special case of Theorem 1

Theorem 7. *Take $a > 0$. For a sample $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$, where $x_1, x_2, \ldots, x_T \in \mathbb{R}^n$ and $y_1, y_2, \ldots, y_T \in \mathbb{R}^n$, let $\gamma_1^{\mathrm{RR}}, \gamma_2^{\mathrm{RR}}, \ldots, \gamma_T^{\mathrm{RR}}$ be the predictions output by linear ridge regression with the parameter a in the on-line mode. Then*

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

where $A_t = \sum_{i=1}^{t} x_i' x_i + aI = X_t' X_t$, X_t is the design matrix consisting of column vectors $x_1, x_2, \ldots, x_t$, and $Y_t = (y_1, y_2, \ldots, y_t)'$.

PROOF. We start by showing that the first two terms are equal and then proceed to the last term.

Consider the pool of Gaussian experts with the variance $\sigma^2 = a$ and the learner following the Bayesian merging algorithm with the initial distribution $\mathcal{N}(0, I)$ on the experts.

It follows from Lemma 6, that the total loss of the learner over T steps is given by

$$\mathrm{Loss}_T = -\sum_{t=1}^{T} \ln \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{(\gamma_t - y_t)^2}{2\sigma_t^2}} \tag{11}$$

$$= \sum_{t=1}^{T} \frac{(\gamma_t - y_t)^2}{2\sigma_t^2} + \ln \prod_{t=1}^{T} \sigma_t + \frac{T}{2} \ln(2\pi) \ , \tag{12}$$

where $\sigma_t^2 = \sigma^2(1 + x_t' A_{t-1}^{-1} x_t)$.

Lemma 5 implies that

$$\mathrm{Loss}_T = -\ln\left(\frac{1}{(2\pi\sigma^2)^{T/2}(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^T (\theta' x_t - y_t)^2 - \frac{1}{2}\|\theta\|^2} d\theta\right) \ .$$

It follows from Proposition 1 that the integral evaluates to

$$e^{-\min_{\theta\in\mathbb{R}_n}\left(\frac{1}{2\sigma^2}\sum_{t=1}^T (\theta' x_t - y_t)^2 + \frac{1}{2}\|\theta\|^2\right)} \frac{\pi^{n/2}}{\sqrt{\det(A_T/(2\sigma^2))}} = e^{-\frac{1}{2\sigma^2}\min_{\theta\in\mathbb{R}_n}\left(\sum_{t=1}^T (\theta' x_t - y_t)^2 + \sigma^2\|\theta\|^2\right)} \frac{(2\pi)^{n/2}}{\sqrt{\det(A_T/\sigma^2)}}$$

and thus

$$\mathrm{Loss}_T = \frac{1}{2\sigma^2}\min_{\theta\in\mathbb{R}_n}\left(\sum_{t=1}^T (\theta' x_t - y_t)^2 + \sigma^2\|\theta\|^2\right) + \frac{T}{2}\ln(2\pi) + T\ln\sigma + \frac{1}{2}\ln\det\frac{A_T}{\sigma^2} \ . \tag{13}$$

Let us equate the expressions for the loss provided by (12) and (13). To prove the identity we need to show that

$$\frac{1}{2}\ln\prod_{t=1}^T \sigma_t^2 = \frac{1}{2}T\ln\sigma^2 + \frac{1}{2}\ln\det\frac{A_T}{\sigma^2} \ .$$

This equality follows from the lemma.

Lemma 8. *For any $a > 0$ and positive integer T we have*

$$\det\frac{A_T}{a} = \prod_{t=1}^T (1 + x_t' A_{t-1}^{-1} x_t) \ ,$$

where $A_t = \sum_{i=1}^t x_i' x_i + aI$.

Proof. We will use the matrix determinant lemma

$$\det(A + uv') = (1 + v' A^{-1} u)\det A \ ,$$

which holds for any non-singular square matrix A and vectors u and v (see, e.g., [29], Theorem 18.1.1). We get

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

□

Remark 4. The lemma is in fact a special case of Lemma 4 with the linear kernel $\mathcal{K}(u_1, u_2) = u_1' u_2$ and $a = \sigma^2$. As shown in Subsection 6.5, $d_t/a = x_t' A_{t-1}^{-1} x_t$. The Sylvester identity implies that $\det(aI + K_T) = \det(aI + X_T' X_T) = \det(aI + X_T X_T') = \det A_T$.

We have shown that the left-hand side and the middle terms in the identity are equal. Let us proceed to the equality between the middle and the right-hand side terms.

The minimum in the middle term is achieved on $\theta_{T+1}^{\mathrm{RR}} = A_T^{-1} X_T Y_T = (X_T X_T' + aI)^{-1} X_T Y_T$ as shown in Subsection 6.3. Using Lemma 9 we can also write $\theta_{T+1}^{\mathrm{RR}} = X_T (X_T' X_T + aI)^{-1} Y_T$. The proof is by direct substitution of these expressions for $\theta_{T+1}^{\mathrm{RR}}$. We have

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

$$\left(\theta_{T+1}^{\mathrm{RR}}\right)' (X_T X_T' + aI) \theta_{T+1}^{\mathrm{RR}} - 2 \left(\theta_{T+1}^{\mathrm{RR}}\right)' X_T Y_T + Y_T' Y_T \ .$$

Substituting the first expression for the second appearance of $\theta_{T+1}^{\mathrm{RR}}$ and cancelling out $X_T X_T' + aI$ we get

$$M = (-\theta_{T+1}^{\mathrm{RR}} X_T + Y_T') Y_T \ .$$

Substituting the second expression for $\theta_{T+1}^{\mathrm{RR}}$ yields

$$M = Y_T' (-(X_T' X_T + aI)^{-1} X_T' X_T + I) Y_T$$

It remains to carry $(X_T' X_T + aI)^{-1}$ out of the brackets and cancel out the remaining terms. □

6.5. Kernelisation

Let us derive Theorem 1 from Theorem 7.

First, let us show that Theorem 7 is really a special case of Theorem 1 for the linear kernel $\mathcal{K}(x_1, x_2) = x_1' x_2$. We will consider the identity term by term. By Lemma 9 the prediction output by linear ridge regression on step t equals

$$\begin{aligned} \left(\theta_t^{\mathrm{RR}}\right)' x_t &= Y_{t-1}' X_{t-1}' (X_{t-1} X_{t-1}' + aI)^{-1} x_t \\ &= Y_{t-1}' (X_{t-1}' X_{t-1} + aI)^{-1} X_{t-1} x_t \\ &= Y_{t-1}' (K_{t-1} + aI)^{-1} k(x_t) \ . \end{aligned}$$

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

$$\begin{aligned} \frac{d_t}{a} &= \frac{1}{a} \left[\mathcal{K}(x_t, x_t) - k_{t-1}'(x_t) (K_{t-1} + aI)^{-1} k_{t-1}(x_t) \right] \\ &= \frac{1}{a} \left[x_t' x_t - (x_t' X_{t-1}) (X_{t-1}' X_{t-1} + aI)^{-1} (X_{t-1}' x_t) \right] \ . \end{aligned}$$

We can apply Lemma 9 and further obtain

$$\begin{aligned} \frac{d_t}{a} &= \frac{1}{a} \left[x_t' x_t - x_t' (X_{t-1} X_{t-1}' + aI)^{-1} X_{t-1} X_{t-1}' x_t \right] \\ &= \frac{1}{a} \left[x_t' (I - (X_{t-1} X_{t-1}' + aI)^{-1} X_{t-1} X_{t-1}') x_t \right] \\ &= x_t' (X_{t-1} X_{t-1}' + aI)^{-1} x_t \\ &= x_t' A_{t-1}^{-1} x_t \ . \end{aligned} \tag{14}$$

Let us proceed to the middle term in the identity. The set of functions $f_\theta(x) = \theta' x$ on $\mathbb{R}^n$ with the scalar product $\langle f_{\theta_1}, f_{\theta_2} \rangle = \theta_1' \theta_2$ is a Hilbert space. It contains all functions $\mathcal{K}(u, \cdot) = f_u$ and the reproducing

property for $\mathcal{K}$ holds: $\langle f_\theta, \mathcal{K}(x,\cdot)\rangle = \langle f_\theta, f_x\rangle = \theta' x = f_\theta(x)$. The minimum in the middle term of Theorem 7 is thus the same as in the middle term of Theorem 1.

For the right-hand side term the equality is obvious.

Now take an arbitrary kernel $\mathcal{K}$ on a domain X and let $\mathcal{F}$ be the corresponding RKHS. We will apply a standard kernel trick. Consider a sample $(x_1, y_1), (x_2, y_2), \dots, (x_T, y_T)$, where $x_t \in X$ and $y_t \in \mathbb{R}$, $t = 1, 2, \dots, T$. It follows from the representer theorem (see Proposition 3) that the minimum in the middle term is achieved on a linear combination of the form $f(\cdot) = \sum_{t=1}^{T} c_t \mathcal{K}(x_t, \cdot)$, where $c_1, c_2, \dots, c_T \in \mathbb{R}$. These linear combinations form a finite-dimensional subspace in the RKHS $\mathcal{F}$. Let $e_1, e_2, \dots, e_m$, $m \le T$, be its orthonormal base and let $\mathcal{C}$ map each linear combination f into the (column) vector of its coordinates in $e_1, e_2, \dots, e_m$. Since the base is orthonormal, the scalar product does not change and $\langle f_1, f_2\rangle_{\mathcal{F}} = (\mathcal{C}(f_1))'\mathcal{C}(f_2)$. The reproducing property implies that

$$f(x_t) = \langle f, \mathcal{K}(x_t, \cdot)\rangle_{\mathcal{F}} = (\mathcal{C}(f))'\mathcal{C}(\mathcal{K}(x_t, \cdot))$$

for $t = 1, 2, \dots, T$. We also have

$$\mathcal{K}(x_i, x_j) = \langle \mathcal{K}(x_i, \cdot), \mathcal{K}(x_j, \cdot)\rangle_{\mathcal{F}} = (\mathcal{C}(\mathcal{K}(x_i, \cdot)))'\mathcal{C}(\mathcal{K}(x_j, \cdot)) \ ,$$

$i, j = 1, 2, \dots, T$. Note that $\mathcal{C}$ is a surjection: each $\theta \in \mathbb{R}^m$ is an image of some linear combination f.

Consider the sample $(\tilde{x}_1, y_1), (\tilde{x}_2, y_2), \dots, (\tilde{x}_T, y_T)$, where $\tilde{x}_t = \mathcal{C}(\mathcal{K}(x_t, \cdot)) \in \mathbb{R}^m$, $t = 1, 2, \dots, T$. Clearly, linear ridge regression in the on-line mode outputs the same predictions on this sample as the kernel ridge regression on the original sample and $\langle \tilde{x}_i, \tilde{x}_j\rangle = \mathcal{K}(x_i, x_j)$. The minimum from Theorem 1 on the original sample clearly coincides with the minimum from Theorem 7 on the new sample.

Theorem 1 follows.

Appendix A. Optimality of Kernel Ridge Regression

In this appendix we derive the optimality property for the ridge regression function function f_{RR}.

Proposition 2. *Let $\mathcal{K} : X \times X \to \mathbb{R}$ be a kernel on a domain X and $\mathcal{F}$ be the corresponding RKHS. For every non-negative integer T, every $x_1, x_2, \dots, x_T \in X$ and $y_1, y_2, \dots, y_T \in \mathbb{R}$, and every $a > 0$ the minimum*

$$\min_{f \in \mathcal{F}} \left(\sum_{t=1}^{T} (f(x_t) - y_t)^2 + a\|f\|_{\mathcal{F}}^2 \right) \tag{15}$$

is achieved on the unique function $f_{\mathrm{RR}}(x) = Y'(aI + K)^{-1}k(x)$ for $T > 0$, where Y, K, and $k(x)$ are as in Subsection 2.2, and $f_{\mathrm{RR}}(x) = 0$ identically for $T = 0$

Proof. If $T = 0$, i.e., the initial sample is empty, the sum in (15) contains no terms and the minimum is achieved on the unique function f with the norm $\|f\|_{\mathcal{F}} = 0$. This function is identically equal to zero and it coincides with f_{RR} for this case by definition. For the rest of the proof assume $T > 0$.

The representer theorem (see Proposition 3) implies that every minimum in (15) is achieved on a linear combination of the form $f(\cdot) = \sum_{t=1}^{T} c_t \mathcal{K}(x_t, \cdot)$.

The minimum in (15) thus can be taken over a finite-dimensional space. As $\|f\|_{\mathcal{F}} \to \infty$, the expression tends to $+\infty$, and thus the minimum can be taken over a bounded set of functions. The value $f(x) = \langle f, \mathcal{K}(x, \cdot)\rangle_{\mathcal{F}}$ is continuous in f for every $x \in X$. Therefore we are minimising a continuous function over a bounded set in a finite-dimensional space. The minimum must be achieved on some f.

Let $C = (c_1, c_2, \dots, c_T)'$ be the vector of coefficients of some optimal function $f(x) = \sum_{t=1}^{T} c_t \mathcal{K}(x_t, x) = C'k(x)$. It is easy to see that the vector $(f(x_1), f(x_2), \dots, f(x_T))'$ of values of f equals KC and

$$\|f\|_{\mathcal{F}}^2 = \sum_{i,j=1}^{T} c_i c_j \langle \mathcal{K}(x_i, \cdot), \mathcal{K}(x_j, \cdot)\rangle_{\mathcal{F}} = C'KC \ .$$

Thus

$$\sum_{t=1}^{T}(f(x)-y_t)^2 + a\|f\|_{\mathcal{F}}^2 = \|KC-Y\|^2 + aC'KC$$
$$= C'K^2C - 2Y'KC + \|Y\|^2 + aC'KC \ .$$

Since f is optimal, the derivative over C must vanish. By differentiation we obtain

$$2K^2C - 2KY + 2aKC = 0$$

and

$$K(K+aI)C = KY \ .$$

Hence

$$(K+aI)C = Y + v$$

and

$$C = (K+aI)^{-1}Y + (K+aI)^{-1}v \ ,$$

where v belongs to the null space of K, i.e., $Kv = 0$.

Let us show that $K(K+aI)^{-1}v = 0$. We need a simple matrix identity; as it occurs in this paper quite often, we formulate it explicitly.

Lemma 9. *For any (not necessarily square) matrices A and B and any constant a the identity*

$$A(BA+aI)^{-1} = (AB+aI)^{-1}A$$

holds provided the inversions can be performed. If $B = A'$ and $a > 0$, the matrices $AB + aI$ and $BA + aI$ are both positive-definite and therefore non-singular.

Proof. We have $ABA + aA = A(BA+aI) = (AB+aI)A$. If $AB+aI$ and $BA+aI$ are invertible, we can multiply the equality by the inverses. □

We get $K(K+aI)^{-1}v = (K+aI)^{-1}Kv = 0$. Therefore C has the form $C = (K+aI)^{-1}Y + u$, where $Ku = 0$.

Consider the function $f_u(x) = u'k(x)$. It is a linear combination of $\mathcal{K}(x_i,\cdot)$. On the other hand, it vanishes at every x_t, $t = 1, 2, \ldots, T$, because $Ku = 0$. We have

$$0 = f_u(x_t) = \langle f, \mathcal{K}(x_t,\cdot)\rangle_{\mathcal{F}}$$

and thus f_u is orthogonal to the space of linear combinations. This is only possible if $f_u = 0$.

Thus the minimum can only be achieved on a unique function that can be represented as $f_{\mathrm{RR}}(x) = Y'(K+aI)^{-1}k(x)$. Since it must be achieved somewhere, it is achieved on f_{RR}. □

Appendix B. Representer Theorem

In this appendix we formulate and prove a version of the reproducing property for RKHSs. See [30] for more details including a history of the theorem.

Proposition 3. *Let $\mathcal{K}$ be a kernel on a domain X, $\mathcal{F}$ be the corresponding RKHS and*

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

be a sample such that $x_t \in X$ and $y_t \in \mathbb{R}$, $t = 1, 2, \ldots, T$. Then for every $f \in \mathcal{F}$ there is a linear combination $\tilde{f}(\cdot) = \sum_{t=1}^{T} c_t \mathcal{K}(x_t, \cdot) \in \mathcal{F}$ such that

$$\sum_{t=1}^{T} (\tilde{f}(x_t) - y_i)^2 \leq \sum_{t=1}^{T} (f(x_t) - y_i)^2$$

and $\|\tilde{f}\|_{\mathcal{F}} \leq \|f\|_{\mathcal{F}}$. If f is not itself a linear combination of this type, there is a linear combination $\tilde{f}$ with this property such that $\|\tilde{f}\|_{\mathcal{F}} < \|f\|_{\mathcal{F}}$.

Proof. The linear combinations of $\mathcal{K}(x_t, \cdot)$ form a finite-dimensional (and therefore closed) subspace in the Hilbert space $\mathcal{F}$. Every $f \in \mathcal{F}$ can be represented as $f = h + g$, where h is a linear combination and g is orthogonal to the subspace of linear combinations. For every $t = 1, 2, \ldots, T$ we have $g(x_t) = \langle g, \mathcal{K}(x_t, \cdot) \rangle_{\mathcal{F}} = 0$ and the values of f and h on $x_1, x_2, \ldots, x_T$ coincide. On the other hand, the Pythagoras theorem implies that $\|f\|_{\mathcal{F}}^2 = \|h\|_{\mathcal{F}}^2 + \|g\|_{\mathcal{F}}^2 \geq \|h\|_{\mathcal{F}}^2$; if $g \neq 0$, the inequality is strict. □

Appendix C. An Upper Bound on a Determinant

In this appendix we reproduce an upper bound from [17].

Proposition 4. *Let the columns of a $n \times T$ matrix X be vectors $x_1, x_2, \ldots, x_T \in \mathbb{R}^n$ and $a > 0$. If $\|x_t\| \leq B$, $t = 1, 2, \ldots, T$, then*

$$\det\left(I + \frac{1}{a} XX'\right) = \det\left(I + \frac{1}{a} X'X\right) \leq \left(1 + \frac{TB^2}{an}\right)^n .$$

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$ be the eigenvalues (counting multiplicities) of the symmetric positive-definite matrix XX'. The eigenvalues of $I + \frac{1}{a} XX'$ are then $1 + \lambda_1/a, 1 + \lambda_2/a, \ldots, 1 + \lambda_n/a$ and $\det(I + \frac{1}{a} XX') = \prod_{i=1}^{n} (1 + \frac{\lambda_i}{a})$.

The sum of eigenvalues $\lambda_1 + \lambda_2 + \ldots + \lambda_n$ equals the trace $\mathrm{tr}(XX')$ and $\mathrm{tr}(XX') = \mathrm{tr}(X'X)$. Indeed, the matrices AB and BA (provided they exist) have the same non-zero eigenvalues counting multiplicities while zero eigenvalues do not contribute to the trace. Alternatively one can verify the equality $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ by a straightforward calculation, see, e.g., [31], Proposition 10.9 (p. 219). The matrix $X'X$ is the Gram matrix of vectors $x_1, x_2, \ldots, x_T$ and the elements on its diagonal are the squared quadratic norms of the vectors not exceeding B^2. We get $\mathrm{tr}(XX') = \mathrm{tr}(X'X) \leq TB^2$.

The problem has reduced to obtaining an upper bound on the product of some positive numbers with a known sum. The inequality of arithmetic and geometric means implies that

$$\prod_{i=1}^{n} \left(1 + \frac{1}{a}\lambda_i\right) \leq \left(\frac{1}{n} \sum_{i=1}^{n} \left(1 + \frac{1}{a}\lambda_i\right)\right)^n = \left(1 + \frac{1}{an} \sum_{i=1}^{n} \lambda_i\right)^n .$$

Combining this with the bound on the trace obtained earlier proves the lemma. □

Appendix D. A Lemma about Partitioned Matrices

In this appendix we formulate and prove a matrix lemma for the proof of Lemma 2.

Lemma 10. *If a symmetric positive-definite matrix M is partitioned as*

$$M = \begin{pmatrix} A & B \\ B' & D \end{pmatrix} ,$$

where A and D are square matrices, then A is non-singular, and if a column vector x of the same height as M is partitioned as

$$x = \begin{pmatrix} u \\ v \end{pmatrix} ,$$

where u is of the same height as A, then $x'M^{-1}x \geq u'A^{-1}u \geq 0$.

PROOF. We shall rely on the following formula for inverting a partitioned matrix: if

$$M = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}$$

then the inverse can be written as

$$M^{-1} = \begin{pmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{pmatrix} \ ,$$

where

$$\begin{aligned} \tilde{P} &= P^{-1} + P^{-1}Q(S - RP^{-1}Q)^{-1}RP^{-1} \ , \\ \tilde{Q} &= -P^{-1}Q(S - RP^{-1}Q)^{-1} \ , \\ \tilde{R} &= -(S - RP^{-1}Q)^{-1}RP^{-1} \ , \\ \tilde{S} &= (S - RP^{-1}Q)^{-1} \ , \end{aligned}$$

provided all the inverses exist (see [32], Section 2.7.4, equation (2.7.25)). Applying these formulae to our partitioning of M we get

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BE^{-1}B'A^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}B'A^{-1} & E^{-1} \end{pmatrix} \ ,$$

where $E = D - B'A^{-1}B$.

The matrix A is symmetric positive-definite as a minor of a symmetric positive-definite matrix; therefore it is non-singular. Non-singularity of E follows from the identity

$$\det M = \det P \det(S - RP^{-1}Q) \ ,$$

where M and its blocks are as above (see [32], Section 2.7.4, equation (2.7.26) and [33], Section 0.8.5; the matrix $S - RP^{-1}Q$ is known as the Schur complement of P). Applying this identity to our matrices yields

$$\det M = \det A \det E$$

and since both M and A are non-singular, E is also non-singular. This justifies the use of the formula for the inverse of a partitioned matrix in this case.

Note also that E^{-1} is symmetric and positive-definite as a minor of a symmetric positive-definite matrix M^{-1}.

We can now write

$$x'Mx = u'A^{-1}u + u'A^{-1}BE^{-1}B'A^{-1}u - 2u'A^{-1}BE^{-1}v + v'E^{-1}v$$

(since $u'A^{-1}BE^{-1}v$ is a number, it equals its transpose). The first term in the sum is just what we need for the statement of the lemma. Let us show that the sum of the remaining three terms is non-negative. Let $w = B'A^{-1}u$. We have

$$\begin{aligned} &u'A^{-1}BE^{-1}B'A^{-1}u - 2u'A^{-1}BE^{-1}v + v'E^{-1}v = \\ &\qquad w'E^{-1}w - 2w'E^{-1}v + v'E^{-1}v = \begin{pmatrix} w' & v' \end{pmatrix} \begin{pmatrix} E^{-1} & -E^{-1} \\ -E^{-1} & E^{-1} \end{pmatrix} \begin{pmatrix} w \\ v \end{pmatrix} \ . \end{aligned}$$

To complete the proof, we need the following simple lemma.

Lemma 11. *If a matrix H is symmetric positive-semidefinite, then the matrix*

$$\begin{pmatrix} H & -H \\ -H & H \end{pmatrix}$$

is also symmetric positive-semidefinite.

PROOF. We will rely on the following criterion. A symmetric matrix H is positive-semidefinite if and only if it is has a symmetric square root L such that $H = L^2$ (the if part is trivial and the only if part can be proven by considering the orthonormal base where H diagonalises). We have

$$\begin{pmatrix} \frac{L}{\sqrt{2}} & -\frac{L}{\sqrt{2}} \\ -\frac{L}{\sqrt{2}} & \frac{L}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{L}{\sqrt{2}} & -\frac{L}{\sqrt{2}} \\ -\frac{L}{\sqrt{2}} & \frac{L}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} L^2 & -L^2 \\ -L^2 & L^2 \end{pmatrix} \ .$$

□

Thus

$$\begin{pmatrix} w' & v' \end{pmatrix} \begin{pmatrix} E^{-1} & -E^{-1} \\ -E^{-1} & E^{-1} \end{pmatrix} \begin{pmatrix} w \\ v \end{pmatrix} \geq 0 \ .$$

□

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, Alexey Chernov, and Wouter M. Koolen for useful discussions and to anonymous COLT, ALT, and Theoretical Computer Science reviewers for detailed comments.

References

[1] F. Zhdanov, V. Vovk, Competing with Gaussian linear experts, CoRR abs/0910.4683.
[2] A. E. Hoerl, Application of ridge analysis to regression problems, Chemical Engineering Progress 58 (1962) 54–59.
[3] C. Saunders, A. Gammerman, V. Vovk, Ridge regression learning algorithm in dual variables, in: Proceedings of the 15th International Conference on Machine Learning, 1998, pp. 515–521.
[4] K. S. Azoury, M. K. Warmuth, Relative loss bounds for on-line density estimation with the exponential family of distributions, Machine Learning 43 (2001) 211–246.
[5] S. M. Kakade, M. W. Seeger, D. P. Foster, Worst-case bounds for Gaussian process models, in: Proceedings of the 19th Annual Conference on Neural Information Processing Systems, 2005.
[6] C. A. Micchelli, M. Pontil, Learning the kernel function via regularization, Journal of Machine Learning Research 6 (2005) 1099–1125.
[7] V. Vovk, Competitive on-line statistics, International Statistical Review 69 (2) (2001) 213–248.
[8] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, 2006.
[9] N. Cesa-Bianchi, P. Long, M. K. Warmuth, Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent, IEEE Transactions on Neural Networks 7 (1996) 604–619.
[10] J. Kivinen, M. K. Warmuth, Exponentiated gradient versus gradient descent for linear predictors, Infornation and Computation 132 (1) (1997) 1–63.
[11] M. Herbster, M. K. Warmuth, Tracking the best linear predictor, Journal of Machine Learning Research 1 (2001) 281–309.
[12] S. Busuttil, Y. Kalnishkan, Online regression competitive with changing predictors, in: Algorithmic Learning Theory, 18th International Conference, Proceedings, 2007, pp. 181–195.
[13] A. V. Chernov, F. Zhdanov, Prediction with expert advice under discounted loss, in: Proccedings of ALT 2010, Vol. LNAI 6331, Springer, 2010, pp. 255–269.
[14] N. Aronszajn, La théorie des noyaux reproduisants et ses applications. Première partie, Proceedings of the Cambridge Philosophical Society 39 (1943) 133–153.
[15] M. W. Seeger, S. M. Kakade, D. P. Foster, Information consistency of nonparametric Gaussian process methods, IEEE Transactions on Information Theory 54 (5) (2008) 2376–2382.
[16] H. V. Henderson, S. R. Searle, On deriving the inverse of a sum of matrices, SIAM Review 23 (1).
[17] N. Cesa-Bianchi, A. Conconi, C. Gentile, A second-order perceptron algorithm, SIAM Journal on Computing 34 (3) (2005) 640–668.
[18] M. Kumon, A. Takemura, K. Takeuchi, Sequential optimizing strategy in multi-dimensional bounded forecasting games, Stochastic Processes and their Applications 121 (2011) 155–183.
[19] A. Brown, A. Page, Elements of Functional Analysis, Van Nostrand Reinhold, 1970.
[20] J. Lamperti, Stochastic Processes: A Survey of the Mathematical Theory, Springer, 1977.
[21] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, 2006.
[22] E. F. Beckenbach, R. E. Bellman, Inequalities, Springer, 1961.
[23] V. Vovk, Aggregating strategies, in: Proceedings of the 3rd Annual Workshop on Computational Learning Theory, Morgan Kaufmann, San Mateo, CA, 1990, pp. 371–383.

[24] V. Vovk, A game of prediction with expert advice, Journal of Computer and System Sciences 56 (1998) 153–173.
[25] A. DeSantis, G. Markowski, M. N. Weigman, Learning probabilistic prediction functions, in: Proceedings of the 1988 Workshop on Computational Learning Theory, 1988, pp. 312–328.
[26] G. H. Hardy, J. E. Littlewood, G. Pólya, Inequalities, 2nd Edition, Cambridge University Press, 1952.
[27] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[28] A. E. Hoerl, R. W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 42 (2000) 80–86.
[29] D. A. Harville, Matrix Algebra From a Statistician's Perspective, Springer, 1997.
[30] B. Schölkopf, R. Herbrich, A. J. Smola, A generalized representer theorem, in: Proceedings of COLT/EuroCOLT 2001, LNAI 2111, Springer, 2001, pp. 416–426.
[31] S. Axler, Linear Algebra Done Right, 2nd Edition, Springer, 1997.
[32] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, Numerical Recipies: The Art of Scientific Computing, 3rd Edition, Cambrige University Press, 2007.
[33] R. A. Horn, C. R. Johnson, Matrix analysis, Cambridge University Press, 1985.