Non-asymptotic calibration and resolution

Vladimir Vovk
vovk@cs.rhul.ac.uk
http://vovk.net

February 7, 2008

Abstract

We analyze a new algorithm for probability forecasting of binary observations on the basis of the available data, without making any assumptions about the way the observations are generated. The algorithm is shown to be well calibrated and to have good resolution for long enough sequences of observations and for a suitable choice of its parameter, a kernel on the Cartesian product of the forecast space $[0, 1]$ and the data space. Our main results are non-asymptotic: we establish explicit inequalities, shown to be tight, for the performance of the algorithm.

1 Introduction

We consider the problem of forecasting a new observation from the available data, which may include, e.g., all or some of the previous observations and the values of some explanatory variables. To make the process of forecasting more vivid, we imagine that the data and observations are chosen by a player called Reality and the forecasts are made by a player called Forecaster. To establish properties of forecasting algorithms, the traditional theory of machine learning makes some assumptions about the way Reality generates the observations; e.g., statistical learning theory [28] assumes that the data and observations are generated independently from the same probability distribution. A more recent approach, prediction with expert advice (see, e.g., [5]), replaces the assumptions about Reality by a comparison class of prediction strategies; a typical result of this theory asserts that Forecaster can perform almost as well as the best strategies in the comparison class. This paper further explores a third possibility, suggested in [11], which requires neither assumptions about Reality nor a comparison class of Forecaster's strategies. It is shown in [11] that there exists a forecasting strategy which is automatically well calibrated; this result has been further developed in, e.g., [14, 20]. Almost all known calibration results, however, are asymptotic (see [22] and [21] for a critique of the standard asymptotic notion of calibration); a non-asymptotic result about calibration is given in [19], Proposition 2, but even this result involves unspecified constants

and randomization. The main results of this paper (Theorems 1 and 2) establish simple explicit inequalities characterizing calibration and resolution of our deterministic forecasting algorithm.

Next we briefly describe the main features of our proof techniques and their connections with the literature. The proofs rely on the game-theoretic approach to probability suggested in [24]. The forecasting protocol is complemented by another player, Skeptic, whose role is to gamble at the odds given by Forecaster's probabilities. It can be said that our approach to forecasting is Skeptic-based, whereas the traditional approach is Reality-based and prediction with expert advice is Forecaster-based. The two most popular formalizations of gambling are subsequence selection rules (going back to von Mises's collectives) and martingales (going back to Ville's critique [29] of von Mises's collectives and described in detail in [24]). The pioneering paper [11] on what we call the Skeptic-based approach, as well as the numerous papers developing it, used von Mises's notion of gambling; [33] appears to be the first paper in this direction to use Ville's notion of gambling. Another ingredient of this paper's approach, considering Skeptic's continuous strategies and thus avoiding randomization by Forecaster (which was the standard feature of the previous work) goes back to [15] and is also described in [12]; however, I learned it from Akimichi Takemura in June 2004 (whose observation was prompted by Glenn Shafer's talk at the University of Tokyo).

It should be noted that, although our approach was inspired by [11] and papers further developing [11], precise statements of our results and our proof techniques are completely different: they are more in the spirit of Levin's [15] result about the existence of neutral measures (see [32] for details).

This version (version 4) of this technical report differs from the previous one in that it incorporates the changes made in response to the comments of the reviewers of its journal version (to be published in *Theoretical Computer Science*).

2 The algorithms of large numbers

In this section we describe our learning protocol and the general forecasting algorithm studied in this paper. The protocol is:

FOR $n = 1, 2, \ldots$:
  Reality I announces $x_n \in \mathbf{X}$.
  Forecaster announces $p_n \in [0, 1]$.
  Reality II announces $y_n \in \{0, 1\}$.
END FOR.

On each round, Reality chooses the datum x_n, then Forecaster gives his forecast p_n for the next observation, and finally Reality discloses the actual observation $y_n \in \{0, 1\}$. Reality chooses x_n from a *data space* $\mathbf{X}$ and y_n from the two-element set $\{0, 1\}$; intuitively, Forecaster's move p_n is the probability he attaches to the

event $y_n = 1$. *Forecasting algorithm* is Forecaster's strategy in this protocol. For convenience in stating the results of §6, we split Reality into two players, Reality I and Reality II.

Our learning protocol is a perfect-information protocol; in particular, Reality may take into account the forecast p_n when deciding on her move y_n. (This feature is unusual for probability forecasting but it extends the domain of applicability of our results and we have it for free.)

Next we describe the general forecasting algorithm that we study in this paper (it was derived informally in [34]). A function $\mathbf{K} : Z^2 \to \mathbb{R}$, where Z is an arbitrary set and $\mathbb{R}$ is the set of real numbers, is a *kernel on* Z if it is symmetric ($\mathbf{K}(z, z') = \mathbf{K}(z', z)$ for all $z, z' \in Z$) and positive definite ($\sum_{i=1}^m \sum_{j=1}^m \lambda_i \lambda_j \mathbf{K}(z_i, z_j) \geq 0$ for all $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ and all $(z_1, \ldots, z_m) \in Z^m$). The usual interpretation of a kernel $\mathbf{K}(z, z')$ is as a measure of similarity between z and z' (see, e.g., [23], §1.1). Our algorithm has one parameter, which is a kernel on the Cartesian product $[0,1] \times \mathbf{X}$. The most straightforward way of constructing such kernels from kernels on $[0,1]$ and kernels on $\mathbf{X}$ is the operation of tensor product. (See, e.g., [3, 28, 23].) Let us say that a kernel $\mathbf{K}$ on $[0,1] \times \mathbf{X}$ is *forecast-continuous* if the function $\mathbf{K}((p, x), (p', x'))$, where $p, p' \in [0,1]$ and $x, x' \in \mathbf{X}$, is continuous in (p, p') for any fixed $(x, x') \in \mathbf{X}^2$.

K29* ALGORITHM
Parameter: forecast-continuous kernel $\mathbf{K}$ on $[0,1] \times \mathbf{X}$
FOR $n = 1, 2, \ldots$:
  Read $x_n \in \mathbf{X}$.
  Set $S_n(p) := \sum_{i=1}^{n-1} \mathbf{K}((p, x_n), (p_i, x_i))(y_i - p_i) + \frac{1}{2}\mathbf{K}((p, x_n), (p, x_n))(1 - 2p)$
    for $p \in [0,1]$.
  If $\operatorname{sign} S_n(0) = \operatorname{sign} S_n(1) \neq 0$, output $p_n := (1 + \operatorname{sign} S_n(0))/2$;
    otherwise, output any root p of $S_n(p) = 0$ as p_n.
  Read $y_n \in \{0, 1\}$.
END FOR.

(Since the function $S_n(p)$ is continuous, the equation $S_n(p) = 0$ indeed has a solution when $\operatorname{sign} S_n(0) = \operatorname{sign} S_n(1) \neq 0$ does not hold; remember that $\operatorname{sign} S$ is 1 for S positive, -1 for S negative, and 0 for $S = 0$.) The main term in the expression for $S_n(p)$ is $\sum_{i=1}^{n-1} \mathbf{K}((p, x_n), (p_i, x_i))(y_i - p_i)$. Ignoring the other term for a moment, we can describe the intuition behind this algorithm by saying that p_n is chosen so that p_i are unbiased forecasts for y_i on the rounds $i = 1, \ldots, n-1$ for which (p_i, x_i) is similar to (p_n, x_n). The term $\frac{1}{2}\mathbf{K}((p, x_n), (p, x_n))(1 - 2p)$, which can be rewritten as $\mathbf{K}((p, x_n), (p, x_n))(0.5 - p)$, adds an element of regularization, i.e., bias towards the "neutral" value $p_n = 0.5$.

The K29* algorithm requires solving the equation $S_n(p) = 0$, but this can be easily done using the bisection method or one of the numerous more sophisticated methods (see, e.g., [18], Chapter 9).

It is well known (see [10], Theorem II.3.1, for a simple proof) that there exists a function $\Phi : [0,1] \times \mathbf{X} \to \mathcal{H}$ (a *feature mapping* taking values in a

Hilbert space[1] $\mathcal{H}$ called the *feature space*) such that

$$\mathbf{K}(a,b) = \langle \Phi(a), \Phi(b) \rangle_{\mathcal{H}}, \quad \forall a, b \in [0,1] \times \mathbf{X} \tag{1}$$

($\langle \cdot, \cdot \rangle_{\mathcal{H}}$ standing for the inner product in $\mathcal{H}$). It is known that, for any $\mathbf{K}$ and Φ connected by (1), $\mathbf{K}$ is forecast-continuous if and only if Φ is a continuous function of p for each fixed $x \in \mathbf{X}$ (see Appendix B).

Now we can state the basic result about K29* (proved in Appendix A).

Theorem 1 *Let $\mathbf{K}$ be the kernel defined by (1) for a feature mapping $\Phi : [0,1] \times \mathbf{X} \to \mathcal{H}$ continuous in its first argument. The K29* algorithm with parameter $\mathbf{K}$ ensures*

$$\left\| \sum_{n=1}^{N} (y_n - p_n) \Phi(p_n, x_n) \right\|_{\mathcal{H}}^2 \le \sum_{n=1}^{N} p_n (1 - p_n) \left\| \Phi(p_n, x_n) \right\|_{\mathcal{H}}^2, \quad \forall N \in \{1, 2, \ldots\}. \tag{2}$$

Let us assume, for simplicity, that

$$\mathbf{c}_{\mathbf{K}} := \sup_{p,x} \left\| \Phi(p,x) \right\|_{\mathcal{H}} < \infty \tag{3}$$

(it is often a good idea to use kernels with $\left\| \Phi(p,x) \right\|_{\mathcal{H}} \equiv 1$ and, therefore, $\mathbf{c}_{\mathbf{K}} = 1$). Equation (2) then implies

$$\left\| \sum_{n=1}^{N} (y_n - p_n) \Phi(p_n, x_n) \right\|_{\mathcal{H}} \le \frac{\mathbf{c}_{\mathbf{K}}}{2} \sqrt{N}, \quad \forall N \in \{1, 2, \ldots\}. \tag{4}$$

When Φ is absent (in the sense $\Phi \equiv 1$), this shows that the forecasts p_n are unbiased, in the sense that they are close to y_n on average; the presence of Φ implies, for a suitable kernel, "local unbiasedness". This is further discussed in the first part of §5.

In the conference version [31] of this paper we also considered the K29 algorithm, which differs from K29* in that $S_n(p)$ is defined as

$$S_n(p) := \sum_{i=1}^{n-1} \mathbf{K}((p, x_n), (p_i, x_i))(y_i - p_i)$$

and that the requirement that $\mathbf{K}$ should be forecast-continuous is slightly relaxed (the joint continuity in (p, p') is replaced by the separate continuity in p and p'). For the K29 algorithm, the inequality (2) continues to hold if $p_n(1 - p_n)$ is removed; therefore, (4) continues to hold if the denominator 2 is removed. We will sometimes use "algorithms of large numbers" as generic name for the K29 and K29* algorithms; the motivation for these names is that the main properties of these algorithms are easy corollaries of Kolmogorov's 1929 proof [13] of the weak law of large numbers.

[1] Hilbert spaces in this paper are allowed to be non-separable or finite dimensional; we, however, always assume that their dimension is at least 1.

3 Reproducing kernel Hilbert spaces

A *reproducing kernel Hilbert space* (RKHS) on a set Z is a Hilbert space $\mathcal{F}$ of real-valued functions on Z such that the evaluation functional $f \in \mathcal{F} \mapsto f(z)$ is continuous for each $z \in Z$. By the Riesz–Fischer theorem, for each $z \in Z$ there exists a function $\mathbf{K}_z \in \mathcal{F}$ such that

$$f(z) = \langle \mathbf{K}_z, f \rangle_{\mathcal{F}}, \quad \forall f \in \mathcal{F}. \tag{5}$$

The *kernel of RKHS* $\mathcal{F}$ is

$$\mathbf{K}(z, z') := \langle \mathbf{K}_z, \mathbf{K}_{z'} \rangle_{\mathcal{F}} \tag{6}$$

(equivalently, we could define $\mathbf{K}(z, z')$ as $\mathbf{K}_z(z')$ or as $\mathbf{K}_{z'}(z)$). Since (6) is a special case of (1), the function $\mathbf{K}$ defined by (6) is indeed a kernel on Z, as defined earlier. On the other hand, for every kernel $\mathbf{K}$ on Z there exists a unique RKHS $\mathcal{F}$ on Z such that $\mathbf{K}$ is the kernel of $\mathcal{F}$ (see, e.g., [2], Theorem 2).

A long list of RKHS and the corresponding kernels is given in [4], §7.4. Perhaps the most interesting RKHS in our current context are various Sobolev spaces $W^{m,p}(\Omega)$ ([1] is the standard reference for the latter). We will be interested in the especially simple space $W^{1,2}([0,1])$, to be defined shortly; but first let us make a brief terminological remark. The term "Sobolev space" is usually treated as the name for a topological vector space. All these spaces are normable, but different norms are not considered to lead to different Sobolev spaces as long as the topology does not change.

The *Fermi–Sobolev norm* $\|f\|_{\mathrm{FS}}$ of a smooth function $f : [0,1] \to \mathbb{R}$ is defined by

$$\|f\|_{\mathrm{FS}}^2 := \left(\int_0^1 f(t) \, \mathrm{d}t \right)^2 + \int_0^1 \left(f'(t) \right)^2 \, \mathrm{d}t. \tag{7}$$

The *Fermi–Sobolev space* on $[0,1]$ is the completion of the set of smooth $f : [0,1] \to \mathbb{R}$ satisfying $\|f\|_{\mathrm{FS}} < \infty$ with respect to the norm $\|\cdot\|_{\mathrm{FS}}$. It is easy to see that it is in fact an RKHS (indeed, if $\|f\|_{\mathrm{FS}} = c < \infty$, the mean of f is bounded by c in absolute value and $|f(b) - f(a)| \le \int_a^b |f'(t)| \, \mathrm{d}t \le c$ for all $0 \le a < b \le 1$). As a topological vector space, it coincides with the Sobolev space $W^{1,2}([0,1])$. The *Fermi–Sobolev space* on $[0,1]^k$ is the tensor product of k copies of the Fermi–Sobolev space on $[0,1]$.

The kernel of the Fermi–Sobolev space on $[0,1]$ was found in [6] (see also [35], §10.2); it is given by

$$\begin{aligned} \mathbf{K}(t,t') &= k_0(t)k_0(t') + k_1(t)k_1(t') + k_2(|t-t'|) \\ &= 1 + \left(t - \frac{1}{2}\right)\left(t' - \frac{1}{2}\right) + \frac{1}{2}\left(|t-t'|^2 - |t-t'| + \frac{1}{6}\right) \\ &= \frac{1}{2}\min^2(t,t') + \frac{1}{2}\min^2(1-t,1-t') + \frac{5}{6}, \end{aligned} \tag{8}$$

where $k_l := B_l/l!$ are scaled Bernoulli polynomials B_l. We will derive the final expression for $\mathbf{K}(t,t')$ in (8) in Appendix C. For the Fermi–Sobolev space on

$[0,1]^k$ we have

$$\mathbf{K}\left((t_1,\dots,t_k),(t'_1,\dots,t'_k)\right) = \prod_{i=1}^{k}\left(\frac{1}{2}\min^2(t_i,t'_i) + \frac{1}{2}\min^2(1-t_i,1-t'_i) + \frac{5}{6}\right) \tag{9}$$

and, therefore,

$$\mathbf{c}_{\mathbf{K}}^2 = \max_{t\in[0,1]}\left(\frac{1}{2}t^2 + \frac{1}{2}(1-t)^2 + \frac{5}{6}\right)^k = \left(\frac{4}{3}\right)^k. \tag{10}$$

For further information about the Fermi–Sobolev spaces, see [31].

4 The K29* algorithm in RKHS

We can now deduce the following corollary from Theorem 1.

Theorem 2 *Let $\mathcal{F}$ be an RKHS on $[0,1]\times\mathbf{X}$ with a forecast-continuous kernel $\mathbf{K}$. The K29* algorithm with parameter $\mathbf{K}$ ensures*

$$\left|\sum_{n=1}^{N}(y_n-p_n)f(p_n,x_n)\right| \le \|f\|_{\mathcal{F}}\sqrt{\sum_{n=1}^{N}p_n(1-p_n)\mathbf{K}((p_n,x_n),(p_n,x_n))} \tag{11}$$

for all N and all $f\in\mathcal{F}$.

Proof Applying K29* to the feature mapping $(p,x)\in[0,1]\times\mathbf{X}\mapsto\mathbf{K}_{p,x}\in\mathcal{F}$ and using (2), we obtain, for any $f\in\mathcal{F}$:

$$\begin{aligned}\left|\sum_{n=1}^{N}(y_n-p_n)f(p_n,x_n)\right| &= \left|\sum_{n=1}^{N}(y_n-p_n)\left\langle\mathbf{K}_{p_n,x_n},f\right\rangle_{\mathcal{F}}\right| \\ &= \left|\left\langle\sum_{n=1}^{N}(y_n-p_n)\mathbf{K}_{p_n,x_n},f\right\rangle_{\mathcal{F}}\right| \le \left\|\sum_{n=1}^{N}(y_n-p_n)\mathbf{K}_{p_n,x_n}\right\|_{\mathcal{F}}\|f\|_{\mathcal{F}} \\ &\le \|f\|_{\mathcal{F}}\sqrt{\sum_{n=1}^{N}p_n(1-p_n)\mathbf{K}((p_n,x_n),(p_n,x_n))}. \quad \blacksquare\end{aligned}$$

When $\mathbf{c}_{\mathbf{K}}$ in (3) is finite, (11) implies

$$\left|\sum_{n=1}^{N}(y_n-p_n)f(p_n,x_n)\right| \le \frac{\mathbf{c}_{\mathbf{K}}}{2}\|f\|_{\mathcal{F}}\sqrt{N}. \tag{12}$$

5 Informal discussion

In this section we explain why the inequalities in Theorems 1 and 2 can be interpreted as results about calibration and resolution, and then briefly discuss a puzzling aspect of the algorithms of large numbers. For concreteness, we usually talk about the K29* algorithm, but all we say can also be applied, with obvious modifications, to K29.

Calibration, resolution, and calibration-cum-resolution

We start from the intuitive notion of calibration (for further details, see [9] and [11]). The forecasts p_n, $n = 1, \dots, N$, are said to be "well calibrated" (or "unbiased in the small", or "reliable", or "valid") if, for any $p^* \in [0, 1]$,

$$\frac{\sum_{n=1,\dots,N:p_n \approx p^*} y_n}{\sum_{n=1,\dots,N:p_n \approx p^*} 1} \approx p^* \tag{13}$$

provided $\sum_{n=1,\dots,N:p_n \approx p^*} 1$ is not too small. The interpretation of (13) is that the forecasts should be in agreement with the observed frequencies. It will be convenient to rewrite (13) as

$$\frac{\sum_{n=1,\dots,N:p_n \approx p^*} (y_n - p_n)}{\sum_{n=1,\dots,N:p_n \approx p^*} 1} \approx 0. \tag{14}$$

The fact that good calibration is only a necessary condition for good forecasting performance can be seen from the following standard example [9, 11]: if

$$(y_1, y_2, y_3, y_4, \dots) = (1, 0, 1, 0, \dots),$$

the forecasts $p_n = 1/2$, $n = 1, 2, \dots$, are well calibrated but rather poor; it would be better to forecast with

$$(p_1, p_2, p_3, p_4, \dots) = (1, 0, 1, 0, \dots).$$

Assuming that each datum x_n contains the information about the parity of n (which can always be added to x_n), we can see that the problem with the forecasting strategy $p_n \equiv 1/2$ is its lack of resolution: it does not distinguish between the data with odd and even n. In general, we would like each forecast p_n to be as specific as possible to the current datum x_n; the resolution of a forecasting algorithm is the degree to which it achieves this goal (taking it for granted that x_n contains all relevant information).

Analogously to (14), the forecasts p_n, $n = 1, \dots, N$, may be said to have good resolution if, for any $x^* \in \mathbf{X}$,

$$\frac{\sum_{n=1,\dots,N:x_n \approx x^*} (y_n - p_n)}{\sum_{n=1,\dots,N:x_n \approx x^*} 1} \approx 0$$

provided the denominator is not too small. We can also require that the forecasts p_n, $n = 1, \dots, N$, should have good "calibration-cum-resolution": for any $(p^*, x^*) \in [0,1] \times \mathbf{X}$,

$$\frac{\sum_{n=1,\dots,N:(p_n,x_n)\approx(p^*,x^*)} (y_n - p_n)}{\sum_{n=1,\dots,N:(p_n,x_n)\approx(p^*,x^*)} 1} \approx 0$$

provided the denominator is not too small. Notice that even if forecasts have both good calibration and good resolution, they can still have poor calibration-cum-resolution.

It is easy to see that (4) implies good calibration-cum-resolution for a suitable Φ and large N: indeed, (4) shows that the forecasts p_n are unbiased in the neighborhood of each (p^*, x^*) for functions Φ that map distant (p, x) and (p', x') to almost orthogonal elements of the feature space (such as Φ corresponding to the Gaussian kernel

$$\mathbf{K}\left((p,x),(p',x')\right) := \exp\left(\frac{(p-p')^2 + \|x - x'\|^2}{2\sigma^2}\right) \tag{15}$$

for a small "kernel width" $\sigma > 0$).

In general, to make sense of the $\approx$ in the numerator and denominator of, say, (14), we replace each "crisp" point p^* by a "fuzzy point" $I_{p^*} : [0,1] \to [0,1]$; I_{p^*} is required to be continuous, and we might also want to have $I_{p^*}(p^*) = 1$ and $I_{p^*}(p) = 0$ for all p outside a small neighborhood of p^*. The alternative of choosing $I_{p^*} := \mathbb{I}_{[p_-,p_+]}$, where $[p_-, p_+]$ is a short interval containing p^* and $\mathbb{I}_{[p_-,p_+]}$ is its indicator function, does not work because of Oakes's and Dawid's examples [17, 8]; I_{p^*} can, however, be arbitrarily close to $\mathbb{I}_{[p_-,p_+]}$.

Consider, e.g., the following approximation to the indicator function of a short interval $[p_-, p_+]$ containing p^*:

$$f(p) := \begin{cases} 1 & \text{if } p_- + \epsilon \le p \le p_+ - \epsilon \\ 0 & \text{if } p \le p_- - \epsilon \text{ or } p \ge p_+ + \epsilon \\ \frac{1}{2} + \frac{1}{2\epsilon}(p - p_-) & \text{if } p_- - \epsilon \le p \le p_- + \epsilon \\ \frac{1}{2} + \frac{1}{2\epsilon}(p_+ - p) & \text{if } p_+ - \epsilon \le p \le p_+ + \epsilon; \end{cases} \tag{16}$$

we assume that $\epsilon > 0$ satisfies

$$0 < p_- - \epsilon < p_- + \epsilon < p_+ - \epsilon < p_+ + \epsilon < 1.$$

It is clear that this approximation belongs to the Fermi–Sobolev space. An easy computation shows that (12) and (10) imply

$$\left|\sum_{n=1}^{N} (y_n - p_n) f(p_n)\right| \le \frac{1}{\sqrt{3}} \sqrt{\left(\frac{1}{\epsilon} + (p_+ - p_-)^2\right) N} \tag{17}$$

for all N. We can see that (14), in the form

$$\frac{\sum_{n=1,\dots,N} f(p_n)(y_n - p_n)}{\sum_{n=1,\dots,N} f(p_n)} \approx 0,$$

will hold if

$$\sum_{n=1}^{N} f(p_n) \gg \sqrt{N}$$

(roughly, if significantly more than $\sqrt{N}$ forecasts fall in the neighborhood $[p^-, p^+]$ of p^*).

It is clear that inequalities analogous to (17) can also be proved for "soft neighborhoods" of points (p^*, x^*) in $[0,1] \times \mathbf{X}$ (at least when $\mathbf{X}$ is a domain in a Euclidean space), and so Theorem 2 also implies good calibration-cum-resolution for large N. Convenient neighborhoods in $[0,1] \times [0,1]^K$ can be constructed as tensor products of neighborhoods (16).

Inequality (17) and analogous inequalities expressing resolution and calibration-cum-resolution are explicit in the sense that they do not involve limits, o, O, unspecified constants, etc. The price to pay is their relative complexity; therefore, we also state a simple asymptotic result about calibration-cum-resolution.

Corollary 1 *If $\mathbf{X}$ is a compact metric space, some forecasting algorithm guarantees*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} (y_n - p_n) f(p_n, x_n) = 0 \tag{18}$$

for all continuous functions $f : [0,1] \times \mathbf{X} \to \mathbb{R}$.

Calibration corresponds to the case where $f(p,x) = I_{p^*}(p)$ does not depend on x and resolution to the case where $f(p,x) = I_{x^*}(x)$ does not depend on p. This result was proved in [12] in the case of calibration (there are no x_n) and Lipschitz functions f.

Proof of Corollary 1 Let $\mathcal{F}$ be an RKHS on $[0,1] \times \mathbf{X}$ which is *universal*, i.e., dense in the space $C([0,1] \times \mathbf{X})$, and whose kernel $\mathbf{K}$ is continuous and satisfies $\mathbf{c_K} < \infty$. The notion of universality is introduced in [25], Definition 4, and the existence of such an $\mathcal{F}$ is shown in [26], Theorem 2. For any continuous function $f : [0,1] \times \mathbf{X} \to \mathbb{R}$ there is a $g \in \mathcal{F}$ that is ϵ-close to f in the metric $C([0,1] \times \mathbf{X})$, and so, by (12),

$$\begin{aligned}\limsup_{N\to\infty} \left| \frac{1}{N} \sum_{n=1}^{N} (y_n - p_n) f(p_n, x_n) \right| &\le \limsup_{N\to\infty} \left| \frac{1}{N} \sum_{n=1}^{N} (y_n - p_n) g(p_n, x_n) \right| + \epsilon \\ &\le \limsup_{N\to\infty} \frac{1}{N} \frac{\mathbf{c_K}}{2} \|g\|_{\mathcal{F}} \sqrt{N} + \epsilon = \epsilon;\end{aligned}$$

since this holds for any $\epsilon > 0$, (18) also holds. ∎

One of the algorithms achieving (18) for $\mathbf{X} = [0,1]^k$ is K29* applied to the Fermi–Sobolev kernel (9). It is interesting, and somewhat counterintuitive, that K29* applied to the Gaussian kernel (15) (with any $\sigma > 0$) also achieves (18); the universality of the Gaussian kernels is proved in [25] (Example 1).

Our discussion of calibration and resolution in this subsection has been somewhat speculative, and the reader might ask whether these two properties are really useful. This question is answered, to some degree, in [30, 32], which show that probability forecasts satisfying these properties lead to good decisions (at least in the simple decision protocols considered in those papers).

Puzzle of the iterated logarithm

Theorems 1 and 2 imply that the forecasts produced by the K29* algorithm are even closer to the actual observations on average than in the case of "genuine randomness", where Reality produces the data and observations from a probability distribution on $(\mathbf{X} \times \{0,1\})^\infty$ and each p_n is the conditional probability that $y_n = 1$ given $x_1, \dots, x_n$, $y_1, \dots, y_{n-1}$, and whatever further information may be available at this point. Indeed, let us take, for simplicity, $\Phi \equiv 1$ (and $\mathcal{H} := \mathbb{R}$) in Theorem 1. According to the martingale law of the iterated logarithm (see, e.g., [27] or Chapter 5 of [24]), we would expect

$$\limsup_{N \to \infty} \frac{\left| \sum_{n=1}^N (y_n - p_n) \right|}{\sqrt{2 A_N \ln \ln A_N}} = 1, \tag{19}$$

where $A_N := \sum_{n=1}^N p_n (1 - p_n)$ is assumed to tend to ∞ as $N \to \infty$, and so expect, contrary to (4),

$$\sup_{N \in \{1,2,\dots\}} \frac{\left\| \sum_{n=1}^N (y_n - p_n) \Phi(p_n, x_n) \right\|_{\mathcal{H}}}{\sqrt{N}}$$

to be infinite for p_n not consistently very close to 0 or 1. Actually, in this case ($\Phi \equiv 1$) Forecaster can even make sure that

$$\left\| \sum_{n=1}^N (y_n - p_n) \Phi(p_n, x_n) \right\|_{\mathcal{H}} = \frac{1}{2}, \quad \forall N \in \{1, 2, \dots\}$$

(choosing $p_1 := 1/2$ and $p_n := y_{n-1}$, $n = 2, 3, \dots$).

For a general Φ, we can also expect that the probabilities p_n contrived by the algorithms of large numbers (K29 or K29*) will have better calibration and resolution than the true probabilities. There is, however, little doubt that the true probabilities are more useful than any probabilities we are able to come up with. The true probabilities are not as good at calibration and resolution, so they must be better in some other equally important respects. It remains unclear what these other respects may be, and this is what we call the puzzle of the iterated logarithm.

6 Optimality of the K29* algorithm

In this section we establish that the inequalities in Theorems 1 and 2 are tight, in a natural sense.

Equation (2) says that the differences $y_n - p_n$ are small on average, even when scattered in a Hilbert space by multiplying by $\Phi(p_n, x_n)$. The next result says that it is the best Forecaster can do.

Theorem 3 *Let $\Phi : [0,1] \times \mathbf{X} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space. There is a strategy for Reality II which guarantees that*

$$\left\| \sum_{n=1}^{N} (y_n - p_n)\Phi(p_n, x_n) \right\|_{\mathcal{H}}^2 \geq \sum_{n=1}^{N} p_n(1-p_n) \left\| \Phi(p_n, x_n) \right\|_{\mathcal{H}}^2 \tag{20}$$

always holds for all $N = 1, 2, \ldots$, regardless of what the other players do.

Proof Set

$$R_N := \left\| \sum_{n=1}^{N} (y_n - p_n)\Phi(p_n, x_n) \right\|_{\mathcal{H}}, \quad N = 1, 2, \ldots;$$

it is sufficient to show that on the Nth round, $N = 1, 2, \ldots$, Reality II can ensure that

$$R_N^2 - R_{N-1}^2 \geq p_N(1-p_N)\Phi_N^2, \tag{21}$$

where

$$\Phi_N := \left\| \Phi(p_N, x_N) \right\|_{\mathcal{H}}.$$

Fix an N. Define points $A, C, D \in \mathcal{H}$ as

$$\begin{aligned} C &:= \sum_{n=1}^{N-1} (y_n - p_n)\Phi(p_n, x_n), \\ A &:= \sum_{n=1}^{N-1} (y_n - p_n)\Phi(p_n, x_n) + (1 - p_N)\Phi(p_N, x_N), \\ D &:= \sum_{n=1}^{N-1} (y_n - p_n)\Phi(p_n, x_n) + (-p_N)\Phi(p_N, x_N); \end{aligned}$$

it is up to Reality II whether make R_N equal to $|OA|$ or $|OD|$, where O is the origin. Assuming, without loss of generality, that $R_N = \max(|OA|, |OD|)$, we reduce our task to showing that the maximal value of R_{N-1} for fixed R_N, Φ_N, and p_N satisfies (21). It is geometrically obvious (see the last paragraph of this proof for a rigorous argument) that R_{N-1} attains its maximal value when $|OA| = |OD|$; this is illustrated in Figure 1 (remember that all four points, O, A, C, and D, lie in the same plane). Let B be the base of the perpendicular dropped from O onto the interval AD and $h := |OB|$. Since the triangles OBD and OBC are right-angled,

$$\begin{aligned} R_N^2 &= h^2 + \left(\frac{1}{2}\Phi_N \right)^2, \\ R_{N-1}^2 &= h^2 + \left(\frac{1}{2}\Phi_N - p_N\Phi_N \right)^2. \end{aligned}$$

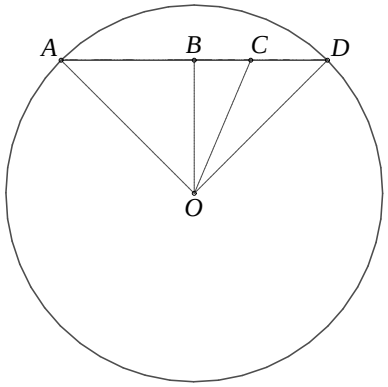

Figure 1: The worst case for Reality II; $|OA| = |OD| = R_N$, $|OC| = R_{N-1}$, $|AC| = (1 - p_N)\Phi_N$, $|CD| = p_N\Phi_N$, $|OB| = h$.

Subtracting the second equality from the first, we obtain

$$R_N^2 - R_{N-1}^2 = \left(\frac{1}{2}\Phi_N\right)^2 - \left(\frac{1}{2}\Phi_N - p_N\Phi_N\right)^2 = p_N(1 - p_N)\Phi_N^2.$$

In conclusion, let us see that the maximum of R_{N-1} is indeed attained when $|OA| = |OD|$. Assume that $|OA| = R_N$, with $|OD|$ now allowed to be less than R_N. Because of the compactness of the disk in Figure 1 (we are only interested in two-dimensional subspaces of $\mathcal{H}$, which are isometrically isomorphic to $\mathbb{R}^2$), the maximum of $|OC|$ is attained at some point C. Supposing $|OD| < R_N$, it is, however, easy to check that no C will be a point of local maximum for $|OC|$; the least trivial case is perhaps where O lies on the line AD and C is between O and D. ∎

The next result establishes the tightness of the bound in Theorem 2.

Theorem 4 *Let $\mathcal{F}$ be an RKHS on $[0,1] \times \mathbf{X}$ with kernel $\mathbf{K}$. Reality II has a strategy which ensures, regardless of what the other players do, that for each $N = 1, 2, \ldots$ there exists a non-zero $f \in \mathcal{F}$ such that*

$$\sum_{n=1}^{N}(y_n - p_n)f(p_n, x_n) \geq \|f\|_{\mathcal{F}} \sqrt{\sum_{n=1}^{N} p_n(1 - p_n)\mathbf{K}((p_n, x_n), (p_n, x_n))}. \tag{22}$$

Proof By Theorem 3 there exists a strategy for Reality II which ensures

$$\left\|\sum_{n=1}^{N}(y_n - p_n)\mathbf{K}_{p_n,x_n}\right\|_{\mathcal{F}} \geq \sqrt{\sum_{n=1}^{N} p_n(1 - p_n)\mathbf{K}((p_n, x_n), (p_n, x_n))}. \tag{23}$$

Taking

$$f := \sum_{n=1}^{N}(y_n - p_n)\mathbf{K}_{p_n,x_n},$$

we obtain:

$$\sum_{n=1}^{N}(y_n - p_n)f(p_n, x_n) = \sum_{n=1}^{N}(y_n - p_n)\left\langle \mathbf{K}_{p_n,x_n}, f\right\rangle_{\mathcal{F}}$$
$$= \left\langle \sum_{n=1}^{N}(y_n - p_n)\mathbf{K}_{p_n,x_n}, f \right\rangle_{\mathcal{F}} = \left\| \sum_{n=1}^{N}(y_n - p_n)\mathbf{K}_{p_n,x_n} \right\|_{\mathcal{F}} \|f\|_{\mathcal{F}}$$
$$\geq \|f\|_{\mathcal{F}} \sqrt{\sum_{n=1}^{N} p_n(1-p_n)\mathbf{K}((p_n,x_n),(p_n,x_n))}.$$

If $f \neq 0$, our task is accomplished. Otherwise, the right-hand side of (23) will also be zero, and we can take any $f \neq 0$. ∎

Acknowledgments

I am grateful to Ilia Nouretdinov for a discussion that lead to the proof of Theorem 3 and to the anonymous reviewers of the conference and journal versions of this paper for their comments. This work was partially supported by MRC (grant S505/65) and Royal Society.

References

[1] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, Amsterdam, second edition, 2003.

[2] Nachman Aronszajn. La théorie générale des noyaux reproduisants et ses applications, première partie. *Proceedings of the Cambridge Philosophical Society*, 39:133–153 (additional note: p. 205), 1944. The second part of this paper is [3].

[3] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[4] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, Boston, 2004.

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

[6] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[7] A. Philip Dawid. Calibration-based empirical probability (with discussion). *Annals of Statistics*, 13:1251–1285, 1985.

[8] A. Philip Dawid. Self-calibrating priors do not exist: Comment. *Journal of the American Statistical Association*, 80:340–341, 1985. This is a contribution to the discussion in [17].

[9] A. Philip Dawid. Probability forecasting. In Samuel Kotz, Norman L. Johnson, and Campbell B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley, New York, 1986.

[10] Joseph L. Doob. *Stochastic Processes*. Wiley, New York, 1953.

[11] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.

[12] Sham M. Kakade and Dean P. Foster. Deterministic calibration and Nash equilibrium. In John Shawe-Taylor and Yoram Singer, editors, *Proceedings of the Seventeenth Annual Conference on Learning Theory*, volume 3120 of *Lecture Notes in Computer Science*, pages 33–48, Heidelberg, 2004. Springer.

[13] Andrei N. Kolmogorov. Sur la loi des grands nombres. *Atti della Reale Accademia Nazionale dei Lincei. Classe di scienze fisiche, matematiche, e naturali. Rendiconti Serie VI*, 185:917–919, 1929.

[14] Ehud Lehrer. Any inspection is manipulable. *Econometrica*, 69:1333–1347, 2001.

[15] Leonid A. Levin. Uniform tests of randomness. *Soviet Mathematics Doklady*, 17:337–340, 1976.

[16] Herbert Meschkowski. *Hilbertsche Räume mit Kernfunktion*. Springer, Berlin, 1962.

[17] David Oakes. Self-calibrating priors do not exist (with discussion). *Journal of the American Statistical Association*, 80:339–342, 1985.

[18] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, Cambridge, second edition, 1992.

[19] Alvaro Sandroni. The reproducible properties of correct forecasts. *International Journal of Game Theory*, 32:151–159, 2003.

[20] Alvaro Sandroni, Rann Smorodinsky, and Rakesh V. Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 28:141–153, 2003.

[21] Mark J. Schervish. Contribution to the discussion in [7]. *Annals of Statistics*, 13:1274–1282, 1985.

[22] Mark J. Schervish. Self-calibrating priors do not exist: Comment. *Journal of the American Statistical Association*, 80:341–342, 1985. This is a contribution to the discussion in [17].

[23] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[24] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

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

[26] Ingo Steinwart, Don Hush, and Clint Scovel. Function classes that approximate the Bayes risk. In Gábor Lugosi and Hans Ulrich Simon, editors, *Proceedings of the Nineteenth Annual Conference on Learning Theory*, volume 4005 of *Lecture Notes in Artificial Intelligence*, pages 79–93, Berlin, 2006. Springer.

[27] William F. Stout. A martingale analogue of Kolmogorov's law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 15:279–290, 1970.

[28] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[29] Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

[30] Vladimir Vovk. Defensive forecasting with expert advice. Technical Report `arXiv:cs.LG/0506041`, `arXiv.org` e-Print archive, 2005. A short version of this technical report is published in the Proceedings of the Sixteenth International Conference on Algorithmic Learning Theory (ed. by Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita), *Lecture Notes in Artificial Intelligence*, vol. 3734, pp. 444–458. Springer, Berlin, 2005.

[31] Vladimir Vovk. Non-asymptotic calibration and resolution. Technical Report `arXiv:cs.LG/0506004` (version 2), `arXiv.org` e-Print archive, July 2005. A short version of this technical report is published in the Proceedings of the Sixteenth International Conference on Algorithmic Learning Theory (ed. by Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita), *Lecture Notes in Artificial Intelligence*, vol. 3734, pp. 429–443. Springer, Berlin, 2005.

[32] Vladimir Vovk. Predictions as statements and decisions. Technical Report `arXiv:cs.LG/0606093`, `arXiv.org` e-Print archive, June 2006.

[33] Vladimir Vovk and Glenn Shafer. Good randomized sequential probability forecasting is always possible. *Journal of the Royal Statistical Society* B, 67:747–763, 2005.

[34] Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting. Technical Report `arXiv:cs.LG/0505083`, `arXiv.org` e-Print archive, May 2005.

[35] Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, PA, 1990.

A Proof of Theorem 1

The proof of Theorem 1 is based on the game-theoretic approach to the foundations of probability proposed in [24]. A new player, called Skeptic, is added to the learning protocol of §2; the idea is that Skeptic is allowed to bet at the odds defined by Forecaster's probabilities. In this proof there is no need to distinguish between Reality I and Reality II.

BINARY FORECASTING GAME I
Players: Reality, Forecaster, Skeptic
Protocol:
$\mathcal{K}_0 := C$.
FOR $n = 1, 2, \ldots$:
    Reality announces $x_n \in \mathbf{X}$.
    Forecaster announces $p_n \in [0, 1]$.
    Skeptic announces $s_n \in \mathbb{R}$.
    Reality announces $y_n \in \{0, 1\}$.
    $\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - p_n)$.
END FOR.

The protocol describes not only the players' moves but also the changes in Skeptic's capital $\mathcal{K}_n$; its initial value is an arbitrary constant C.

The crucial (albeit very simple) observation [34] is that for any continuous strategy for Skeptic there exists a strategy for Forecaster that does not allow Skeptic's capital to grow, regardless of what Reality is doing (similar observations were made in [15] and [12]). To state this observation in its strongest form, we will make Skeptic announce his strategy for each round before Forecaster's move on that round rather than announce his full strategy at the beginning of the game. Therefore, we consider the following perfect-information game:

BINARY FORECASTING GAME II
Players: Reality, Forecaster, Skeptic
Protocol:
$\mathcal{K}_0 := C$.
FOR $n = 1, 2, \ldots$:
    Reality announces $x_n \in \mathbf{X}$.
    Skeptic announces continuous $S_n : [0, 1] \to \mathbb{R}$.
    Forecaster announces $p_n \in [0, 1]$.

Reality announces $y_n \in \{0,1\}$.
$\mathcal{K}_n := \mathcal{K}_{n-1} + S_n(p_n)(y_n - p_n)$.
END FOR.

Lemma 1 *Forecaster has a strategy in Binary Forecasting Game II that ensures* $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \mathcal{K}_2 \geq \cdots$.

Proof Forecaster can use the following strategy to ensure $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \cdots$:

- if $S_n(0)$ and $S_n(1)$ are both positive or both negative, take $p_n := (1 + \operatorname{sign} S_n(0))/2$;
- otherwise, choose p_n so that $S_n(p_n) = 0$ (such a p_n will exist). ∎

A measure-theoretic version of Lemma 1 (involving randomization) was proved in [19], Proposition 1.

Proof of the theorem

We start by noticing that

$$(y_n - p_n)^2 = p_n(1 - p_n) + (1 - 2p_n)(y_n - p_n) \tag{24}$$

both for $y_n = 0$ and for $y_n = 1$. Following K29*, Forecaster ensures that Skeptic will never increase his capital with the strategy

$$s_n := \sum_{i=1}^{n-1} \mathbf{K}\left((p_n, x_n), (p_i, x_i)\right)(y_i - p_i) + \frac{1}{2}\mathbf{K}\left((p_n, x_n), (p_n, x_n)\right)(1 - 2p_n) \tag{25}$$

(continuous in p_n by our assumptions). The increase in Skeptic's capital when he follows (25) is

$$\begin{aligned}
\mathcal{K}_N - \mathcal{K}_0 &= \sum_{n=1}^{N} s_n(y_n - p_n) \\
&= \sum_{n=1}^{N} \sum_{i=1}^{n-1} \mathbf{K}\left((p_n, x_n), (p_i, x_i)\right)(y_n - p_n)(y_i - p_i) \\
&\quad + \frac{1}{2} \sum_{n=1}^{N} \mathbf{K}\left((p_n, x_n), (p_n, x_n)\right)(1 - 2p_n)(y_n - p_n) \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} \mathbf{K}\left((p_n, x_n), (p_i, x_i)\right)(y_n - p_n)(y_i - p_i) \\
&\quad - \frac{1}{2} \sum_{n=1}^{N} \mathbf{K}\left((p_n, x_n), (p_n, x_n)\right)(y_n - p_n)^2
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{2} \sum_{n=1}^{N} \mathbf{K}\left((p_n, x_n), (p_n, x_n)\right) (1 - 2p_n)(y_n - p_n) \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} \mathbf{K}\left((p_n, x_n), (p_i, x_i)\right) (y_n - p_n)(y_i - p_i) \\
&- \frac{1}{2} \sum_{n=1}^{N} \mathbf{K}\left((p_n, x_n), (p_n, x_n)\right) p_n(1 - p_n)
\end{aligned}
$$

(we used (24) in the last equality). We can rewrite this as

$$
\mathcal{K}_N - \mathcal{K}_0 = \frac{1}{2} \left\| \sum_{n=1}^{N} (y_n - p_n) \Phi(p_n, x_n) \right\|_{\mathcal{H}}^2 - \frac{1}{2} \sum_{n=1}^{N} p_n(1 - p_n) \left\| \Phi(p_n, x_n) \right\|_{\mathcal{H}}^2 ,
$$

which immediately implies (2).

B Forecast-continuity of feature mappings and kernels

In this appendix we will prove, essentially following [25], Lemma 3, that the forecast-continuity of a kernel $\mathbf{K}$ on $[0,1] \times \mathbf{X}$ is equivalent to the continuity in p of a feature mapping $\Phi(p, x)$ satisfying (1). As a byproduct, we will also see that the forecast-continuity of a kernel $\mathbf{K}$ on $[0,1] \times \mathbf{X}$ can be equivalently defined by requiring that

- $\mathbf{K}((p, x), (p', x))$ should be continuous in p, for all $x \in \mathbf{X}$ and all $p' \in [0,1]$,
- and $\mathbf{K}((p, x), (p, x))$ should be continuous in p, for all $x \in \mathbf{X}$.

In one direction the statement is obvious: if $\Phi(p, x)$ is continuous in p, the continuity of the operation of taking the inner product immediately implies that $\mathbf{K}$ is forecast-continuous, in both senses.

Now suppose that $\mathbf{K}$ is forecast-continuous, as defined in the first paragraph of this appendix (this is the apparently weaker sense of forecast-continuity). To complete the proof, notice that

$$
\begin{aligned}
&\left\| \Phi(p, x) - \Phi(p_n, x) \right\|_{\mathcal{H}} \\
&= \sqrt{\mathbf{K}((p, x), (p, x)) - 2\mathbf{K}((p, x), (p_n, x)) + \mathbf{K}((p_n, x), (p_n, x))} \\
&\to \sqrt{\mathbf{K}((p, x), (p, x)) - 2\mathbf{K}((p, x), (p, x)) + \mathbf{K}((p, x), (p, x))} = 0
\end{aligned}
$$

when $p_n \to p$ $(n \to \infty)$.

C Derivation of the kernel of the Fermi–Sobolev space

We first describe the standard reduction of the problem of finding the kernel of an RKHS to a variational problem. Let $\mathbf{K}$ be the kernel of an RKHS $\mathcal{F}$ on Z.

Let $c \in Z$. According to [16] (Satz III.3), the minimum of $\|f\|_{\mathcal{F}}$ among the functions $f \in \mathcal{F}$ satisfying $f(c) = 1$ is attained by the function $\mathbf{K}(\cdot, c)/\mathbf{K}(c, c)$. Therefore, we obtain a function $k(\cdot, c)$ proportional to $\mathbf{K}(\cdot, c)$ by solving the optimization problem $\|f\|_{\mathcal{F}} \to \min$ under the constraint $f(c) = 1$ (or under the constraint $f(c) = d$, where d is any other constant). It remains to find the coefficient of proportionality in terms of $k(\cdot, c)$. If $\mathbf{K}(\cdot, \cdot) = \alpha k(\cdot, \cdot)$, we have:

$$\begin{aligned} \mathbf{K}(c, c) &= \|\mathbf{K}(\cdot, c)\|_{\mathcal{F}}^2; \\ \alpha k(c, c) &= \alpha^2 \|k(\cdot, c)\|_{\mathcal{F}}^2; \\ \alpha &= \frac{k(c, c)}{\|k(\cdot, c)\|_{\mathcal{F}}^2}. \end{aligned}$$

Therefore, the recipe for finding $\mathbf{K}$ is: for each $c \in Z$ solve the optimization problem $\|f\|_{\mathcal{F}} \to \min$ under the constraint $f(c) = 1$ (the completeness of RKHS implies that the minimum is attained) and set

$$\mathbf{K}(z, c) := \frac{k(z, c)k(c, c)}{\|k(\cdot, c)\|_{\mathcal{F}}^2}, \tag{26}$$

where $k(\cdot, c)$ is the solution.

Now let us apply this technique to finding the kernel corresponding to the Fermi–Sobolev space on $[0, 1]$ with the norm given by (7). Let $c \in [0, 1]$ and let f be the solution to the optimization problem $\|f\|_{\mathcal{F}} \to \min$ under the constraint $f(c) = 1$ (because of the convexity of the set $\{f \in \mathcal{F} \mid f(c) = 1\}$, there is only one solution). First we show that the derivative f' is a linear function on $[0, c]$ and on $[c, 1]$, arguing indirectly. Suppose, for concreteness, that f' is not linear on the interval $(0, c)$; in particular this interval is non-empty. There are three points $0 < t_1 < t_2 < t_3 < c$ such that

$$f'(t_2) \neq \frac{t_3 - t_2}{t_3 - t_1} f'(t_1) + \frac{t_2 - t_1}{t_3 - t_1} f'(t_3). \tag{27}$$

For a small constant $\epsilon > 0$ (in particular, we assume $2\epsilon < \min(t_1, t_2 - t_1, t_3 - t_2, c - t_3)$), let $g : [0, 1] \to \mathbb{R}$ be a smooth function such that $\int_0^1 g(t)\,\mathrm{d}t = 0$ and:

- $g(t) = 0$ for $t < t_1 - \epsilon$;
- $g(t)$ is increasing for $t_1 - \epsilon < t < t_1 + \epsilon$;
- $g(t) = t_3 - t_2$ for $t_1 + \epsilon < t < t_2 - \epsilon$;
- $g(t)$ is decreasing for $t_2 - \epsilon < t < t_2 + \epsilon$;

- $g(t) = -(t_2 - t_1)$ for $t_2 + \epsilon < t < t_3 - \epsilon$;
- $g(t)$ is increasing for $t_3 - \epsilon < t < t_3 + \epsilon$;
- $g(t) = 0$ for $t > t_3 + \epsilon$.

Since, for any $\delta \in \mathbb{R}$ (we are interested in nonzero δ small in absolute value),

$$\|f + \delta g\|_{FS}^2 = \|f\|_{FS}^2 + 2\delta \int_0^1 f'(t)g'(t)\,\mathrm{d}t + \delta^2 \int_0^1 (g'(t))^2\,\mathrm{d}t,$$

the definition of f implies

$$\int_0^1 f'(t)g'(t)\,\mathrm{d}t = 0.$$

However, as $\epsilon \to 0$, the last integral tends to

$$f'(t_1)(t_3 - t_2) - f'(t_2)(t_3 - t_1) + f'(t_3)(t_2 - t_1),$$

which cannot, by (27), be zero.

Once we know that f is a quadratic polynomial to the left and to the right of c, we can easily find (this can be done conveniently using a computer algebra system) that, ignoring a multiplicative constant,

$$f(t) = 3t^2 + 3c^2 - 6c + 8 = 3t^2 + 3(1 - c)^2 + 5$$

to the left of c and

$$f(t) = 3t^2 + 3c^2 - 6t + 8 = 3(1 - t)^2 + 3c^2 + 5$$

to the right of c. By (26), we can now find

$$\mathbf{K}(t, c) = \frac{f(t)f(c)}{\|f\|_{\mathcal{F}}^2} = f(t)/6,$$

which agrees with (8).