Understanding Gaussian Process Regression Using the Equivalent Kernel

Peter Sollich[1] and Christopher K. I. Williams[2]

[1] Dept of Mathematics, King's College London,
Strand, London WC2R 2LS, U.K.
peter.sollich@kcl.ac.uk
[2] School of Informatics, University of Edinburgh,
5 Forrest Hill, Edinburgh EH1 2QL, U.K.
c.k.i.williams@ed.ac.uk

Abstract. The equivalent kernel [1] is a way of understanding how Gaussian process regression works for large sample sizes based on a continuum limit. In this paper we show how to approximate the equivalent kernel of the widely-used squared exponential (or Gaussian) kernel and related kernels. This is easiest for uniform input densities, but we also discuss the generalization to the non-uniform case. We show further that the equivalent kernel can be used to understand the learning curves for Gaussian processes, and investigate how kernel smoothing using the equivalent kernel compares to full Gaussian process regression.

1 Introduction

Consider the supervised regression problem for a dataset $\mathcal{D}$ with entries $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$. Under Gaussian Process (GP) assumptions the predictive mean at a test point $\mathbf{x}_*$ is given by

$$\bar{f}(\mathbf{x}_*) = \mathbf{k}^\top(\mathbf{x}_*)(K + \sigma^2 I)^{-1}\mathbf{y}, \tag{1}$$

where K denotes the $n \times n$ matrix of covariances between the training points with entries $k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}(\mathbf{x}_*)$ is the vector of covariances $k(\mathbf{x}_i, \mathbf{x}_*)$, σ^2 is the noise variance on the observations and $\mathbf{y}$ is a $n \times 1$ vector holding the training targets. See e.g. [2] for further details.

We can define a vector of functions $\mathbf{h}(\mathbf{x}_*) = (K + \sigma^2 I)^{-1}\mathbf{k}(\mathbf{x}_*)$. Thus we have $\bar{f}(\mathbf{x}_*) = \mathbf{h}^\top(\mathbf{x}_*)\mathbf{y}$, making it clear that the mean prediction at a point $\mathbf{x}_*$ is a linear combination of the target values $\mathbf{y}$. Gaussian process regression is thus a *linear smoother*, see [3, section 2.8] for further details. For a fixed test point $\mathbf{x}_*$, $\mathbf{h}(\mathbf{x}_*)$ gives the vector of weights applied to targets $\mathbf{y}$. Silverman [1] called $\mathbf{h}^\top(\mathbf{x}_*)$ the *weight function*.

Understanding the form of the weight function is made complicated by the matrix inversion of $K+\sigma^2 I$ and the fact that K depends on the specific locations of the n datapoints. Idealizing the situation one can consider the observations to be "smeared out" in $\mathbf{x}$-space at some constant density of observations. In this

case analytic tools can be brought to bear on the problem, as shown below. By analogy to kernel smoothing Silverman [1] called the idealized weight function the *equivalent kernel* (EK).

The structure of the remainder of the paper is as follows: In section 2 we describe how to derive the equivalent kernel in Fourier space. Section 3 derives approximations for the EK for the squared exponential and other kernels. In section 4 we show how use the EK approach to estimate learning curves for GP regression, and compare GP regression to kernel regression using the EK. A summary of our key results can be found in the short proceedings paper [4].

2 Gaussian Process Regression and the Equivalent Kernel

It is well known (see e.g. [5]) that the posterior mean for GP regression can be obtained as the function which minimizes the functional

$$J[f] = \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2, \tag{2}$$

where $\|f\|_{\mathcal{H}}$ is the RKHS norm corresponding to kernel k. (However, note that the GP framework gives much more than just this mean prediction, for example the predictive variance and the marginal likelihood $p(\mathbf{y})$ of the data under the model.)

Let $\eta(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ be the target function for our regression problem and write $\mathbb{E}[(y - f(\mathbf{x}))^2] = \mathbb{E}[(y - \eta(\mathbf{x}))^2] + (\eta(\mathbf{x}) - f(\mathbf{x}))^2$. Using the fact that the first term on the RHS is independent of f motivates considering a smoothed version of equation (2),

$$J_\rho[f] = \frac{\rho}{2\sigma^2}\int(\eta(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x} + \frac{1}{2}\|f\|_{\mathcal{H}}^2,$$

where ρ has dimensions of the number of observations per unit of $\mathbf{x}$-space (length/area/volume etc. as appropriate). If we consider kernels that are stationary, $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}')$, the natural basis in which to analyse equation (2) is the Fourier basis of complex sinusoids so that $f(\mathbf{x})$ is represented as $\int \tilde{f}(\mathbf{s})e^{2\pi i\mathbf{s}\cdot\mathbf{x}}d\mathbf{s}$ and similarly for $\eta(\mathbf{x})$. Thus we obtain

$$J_\rho[f] = \frac{1}{2}\int\left(\frac{\rho}{\sigma^2}|\tilde{f}(\mathbf{s}) - \tilde{\eta}(\mathbf{s})|^2 + \frac{|\tilde{f}(\mathbf{s})|^2}{S(\mathbf{s})}\right)d\mathbf{s}, \tag{3}$$

as $\|f\|_{\mathcal{H}}^2 = \int|\tilde{f}(\mathbf{s})|^2/S(\mathbf{s})d\mathbf{s}$ where $S(\mathbf{s})$ is the power spectrum of the kernel k, $S(\mathbf{s}) = \int k(\mathbf{x})e^{-2\pi i\mathbf{s}\cdot\mathbf{x}}d\mathbf{x}$. $J_\rho[f]$ can be minimized using calculus of variations to obtain $\tilde{f}(\mathbf{s}) = S(\mathbf{s})\eta(\mathbf{s})/(\sigma^2/\rho + S(\mathbf{s}))$ which is recognized as the convolution

$$f(\mathbf{x}_*) = \int h(\mathbf{x}_* - \mathbf{x})\eta(\mathbf{x})d\mathbf{x} \tag{4}$$

Here the Fourier transform of the equivalent kernel $h(\mathbf{x})$ is

$$\tilde{h}(\mathbf{s}) = \frac{S(\mathbf{s})}{S(\mathbf{s}) + \sigma^2/\rho} = \frac{1}{1 + \sigma^2/(\rho S(\mathbf{s}))}. \tag{5}$$

The term σ^2/ρ in the first expression for $\tilde{h}(\mathbf{s})$ corresponds to the power spectrum of a white noise process, whose delta-function covariance function becomes a constant in the Fourier domain. This analysis is known as Wiener filtering; see, e.g. [6, §14-1]. Notice that as $\rho \to \infty$, $h(\mathbf{x})$ tends to the delta function.

To see the relation between the EK and the weights $\mathbf{h}(\mathbf{x}_*)$ for prediction from a finite data set, one notes that the integral in equation (4) can be approximated by the discrete sum $(1/\rho)\sum_i h(\mathbf{x}_* - \mathbf{x}_i)y_i$; the factor $1/\rho$ represents the average volume element associated with each of the discrete training inputs. The EK $h(\mathbf{x}_* - \mathbf{x}_i)$ should therefore approximate the scaled weights $\rho\mathbf{h}(\mathbf{x}_*)$. We will see this confirmed below.

3 The EK for the Squared Exponential and Related Kernels

For certain kernels/covariance functions the EK $h(\mathbf{x})$ can be computed exactly by Fourier inversion. Examples include the Ornstein-Uhlenbeck process in $D = 1$ with covariance $k(x) = e^{-\alpha|x|}$ (see [6, p. 326]), splines in $D = 1$ corresponding to the regularizer $\|Pf\|^2 = \int (f^{(m)})^2 dx$ [1, 7], and the regularizer $\|Pf\|^2 = \int (\nabla^2 f)^2 d\mathbf{x}$ in two dimensions, where the EK is given in terms of the Kelvin function kei [8].

We now consider the commonly used squared exponential (SE) kernel $k(r) = \exp(-r^2/2\ell^2)$, where $r^2 = ||\mathbf{x} - \mathbf{x}'||^2$. (This is sometimes called the Gaussian or radial basis function kernel.) Its Fourier transform is given by $S(\mathbf{s}) = (2\pi\ell^2)^{D/2}\exp(-2\pi^2\ell^2|\mathbf{s}|^2)$, where D denotes the dimensionality of $\mathbf{x}$ (and $\mathbf{s}$) space.

From equation (5) we obtain

$$\tilde{h}_{\mathrm{SE}}(\mathbf{s}) = \frac{1}{1 + b\exp(2\pi^2\ell^2|\mathbf{s}|^2)},$$

where $b = \sigma^2/\rho(2\pi\ell^2)^{D/2}$. We are unaware of an exact result in this case, but the following initial approximation is simple but effective. For large ρ, b will be small. Thus for small $s = |\mathbf{s}|$ we have that $\tilde{h}_{\mathrm{SE}} \simeq 1$, but for large s it is approximately 0. The change takes place around the point s_c where $b\exp(2\pi^2\ell^2 s_c^2) = 1$, i.e. $s_c^2 = \log(1/b)/2\pi^2\ell^2$. As $\exp(2\pi^2\ell^2 s^2)$ grows quickly with s, the transition of $\tilde{h}_{\mathrm{SE}}$ between 1 and 0 can be expected to be rapid, and thus be well-approximated by a step function.

Proposition 1 *The approximate form of the equivalent kernel for the squared-exponential kernel in D-dimensions is given by*

$$h_{\mathrm{SE}}(r) = \left(\frac{s_c}{r}\right)^{D/2} J_{D/2}(2\pi s_c r).$$

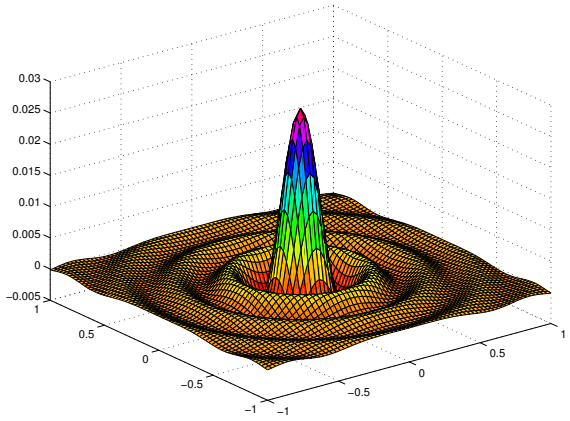

Fig. 1. Plot of the asymptotic form of the EK $(s_c/r)J_1(2\pi s_c r)$ for $D = 2$ and $\rho = 1225$.

Proof: $h_{\mathrm{SE}}(\mathbf{s})$ is a function of $s = |\mathbf{s}|$ only, and for $D > 1$ the Fourier integral can be simplified by changing to spherical polar coordinates and integrating out the angular variables to give

$$h_{\mathrm{SE}}(r) = 2\pi r \int_0^\infty \left(\frac{s}{r}\right)^{\nu+1} J_\nu(2\pi r s)\tilde{h}_{\mathrm{SE}}(s)\, ds \tag{6}$$
$$\simeq 2\pi r \int_0^{s_c} \left(\frac{s}{r}\right)^{\nu+1} J_\nu(2\pi r s)\, ds = \left(\frac{s_c}{r}\right)^{D/2} J_{D/2}(2\pi s_c r).$$

where $\nu = D/2 - 1$, $J_\nu(z)$ is a Bessel function of the first kind and we have used the identity $z^{\nu+1}J_\nu(z) = (d/dz)[z^{\nu+1}J_{\nu+1}(z)]$. □

Note that in $D = 1$ by computing the Fourier transform of the boxcar function we obtain $h_{\mathrm{SE}}(x) = 2s_c\mathrm{sinc}(2\pi s_c x)$ where $\mathrm{sinc}(z) = \sin(z)/z$. This is consistent with Proposition 1 and $J_{1/2}(z) = (2/\pi z)^{1/2}\sin(z)$. The asymptotic form of the EK in $D = 2$ is shown in Figure 1.

Notice that s_c scales as $(\log(\rho))^{1/2}$ so that the width of the EK (which is proportional to $1/s_c$) will decay very slowly as ρ increases. In contrast for a spline of order m (with power spectrum $\propto |\mathbf{s}|^{-2m}$) the width of the EK scales as $\rho^{-1/2m}$ [1].

If instead of $\mathbb{R}^D$ we consider the input set to be the unit circle, a stationary kernel can be periodized by the construction $k_p(x, x') = \sum_{n\in\mathbb{Z}} k(x - x' + 2n\pi)$. This kernel will be represented as a Fourier series (rather than with a Fourier transform) because of the periodicity. In this case the step function in Fourier space approximation would give rise to a Dirichlet kernel as the EK (see [9, section 4.4.3] for further details on the Dirichlet kernel).

We now show that the result of Proposition 1 is asymptotically exact for $\rho \to \infty$, and calculate the leading corrections for finite ρ. The scaling of the

width of the EK as $1/s_c$ suggests writing $h_{\rm SE}(r) = (2\pi s_c)^D g(2\pi s_c r)$. Then from equation (6) and using the definition of s_c

$$\begin{aligned} g(z) &= \frac{z}{s_c(2\pi s_c)^D} \int_0^\infty \left(\frac{2\pi s_c s}{z}\right)^{\nu+1} \frac{J_\nu(zs/s_c)}{1+\exp[2\pi^2\ell^2(s^2-s_c^2)]}\, ds \\ &= z\int_0^\infty \left(\frac{u}{2\pi z}\right)^{\nu+1} \frac{J_\nu(zu)}{1+\exp[2\pi^2\ell^2 s_c^2(u^2-1)]}\, du \end{aligned} \tag{7}$$

where we have rescaled $s = s_c u$ in the second step. The value of s_c, and hence ρ, now enters only in the exponential via $a = 2\pi^2\ell^2 s_c^2$. For $a \to \infty$, the exponential tends to zero for $u < 1$ and to infinity for $u > 1$. The factor $1/[1+\exp(\ldots)]$ is therefore a step function $\Theta(1-u)$ in the limit and Proposition 1 becomes exact, with $g_\infty(z) \equiv \lim_{a\to\infty} g(z) = (2\pi z)^{-D/2} J_{D/2}(z)$. To calculate corrections to this, one uses that for large but finite a the difference $\Delta(u) = \{1+\exp[a(u^2-1)]\}^{-1} - \Theta(1-u)$ is non-negligible only in a range of order $1/a$ around $u = 1$. The other factors in the integrand of equation (7) can thus be Taylor-expanded around that point to give

$$g(z) = g_\infty(z) + z\sum_{k=0}^\infty \frac{I_k}{k!}\frac{d^k}{du^k}\left[\left(\frac{u}{2\pi z}\right)^{\nu+1} J_\nu(zu)\right]\Big|_{u=1}, \quad I_k = \int_0^\infty \Delta(u)(u-1)^k\, du$$

The problem is thus reduced to calculating the integrals I_k. Setting $u = 1 + v/a$ one has

$$\begin{aligned} a^{k+1} I_k &= \int_{-a}^0 \left[\frac{1}{1+\exp(v^2/a+2v)} - 1\right] v^k\, dv + \int_0^\infty \frac{v^k}{1+\exp(v^2/a+2v)}\, dv \\ &= \int_0^a \frac{(-1)^{k+1} v^k}{1+\exp(-v^2/a+2v)}\, dv + \int_0^\infty \frac{v^k}{1+\exp(v^2/a+2v)}\, dv \end{aligned}$$

In the first integral, extending the upper limit to ∞ gives an error that is exponentially small in a. Expanding the remaining $1/a$-dependence of the integrand one then gets, to leading order in $1/a$, $I_0 = c_0/a^2$, $I_1 = c_1/a^2$ while all I_k with $k \geq 2$ are smaller by at least $1/a^2$. The numerical constants are $-c_0 = c_1 = \pi^2/24$. This gives, using that $(d/dz)[z^{\nu+1}J_\nu(z)] = z^\nu J_\nu(z) + z^{\nu+1}J_{\nu-1}(z) = (2\nu+1)z^\nu J_\nu(z) - z^{\nu+1}J_{\nu+1}(z)$:

Proposition 2 *The equivalent kernel for the squared-exponential kernel is given for large ρ by $h_{\rm SE}(r) = (2\pi s_c)^D g(2\pi s_c r)$ with*

$$g(z) = \frac{1}{(2\pi z)^{\frac{D}{2}}}\left\{J_{D/2}(z) + \frac{z}{a^2}\left[(c_0 + c_1(D-1))J_{D/2-1}(z) - c_1 z J_{D/2}(z)\right]\right\}$$

within an expansion in $1/a$; the neglected terms are $O(1/a^4)$.

For e.g. $D = 1$ this becomes $g(z) = \pi^{-1}\{\sin(z)/z - \pi^2/(24a^2)[\cos(z) + z\sin(z)]\}$. Here and in general, by comparing the second part of the $1/a^2$ correction with the leading order term, one estimates that the correction is of relative size z^2/a^2. It will therefore provide a useful improvement as long as $z = 2\pi s_c r < a$; for larger z the expansion in powers of $1/a$ becomes a poor approximation because the correction terms (of all orders in $1/a$) are comparable to the leading order.

3.1 Accuracy of the approximation

To evaluate the accuracy of the approximation we can compute the EK numerically as follows: Consider a dense grid of points in $\mathbb{R}^D$ with a sampling density ρ_{grid}. For making predictions at the grid points we obtain the smoother matrix $K(K+\sigma^2_{\text{grid}}I)^{-1}$, where[3] $\sigma^2_{\text{grid}} = \sigma^2\rho_{\text{grid}}/\rho$, as per equation (1). Each row of this matrix is (up to a factor ρ_{grid}) an approximation to the EK at the appropriate location, as this is the response to a $\mathbf{y}$ vector which is zero at all points except one. Note that in theory one should use a grid over the whole of $\mathbb{R}^D$ but in practice one can obtain an excellent approximation to the EK by only considering a grid around the point of interest as the EK typically decays with distance. Also, by only considering a finite grid one can understand how the EK is affected by edge effects.

Figure 2 shows plots of the weight function for $\rho = 100$, the EK computed on the grid as described above and the analytical sinc approximation. These are computed for parameter values of $\ell^2 = 0.004$ and $\sigma^2 = 0.1$, with $\rho_{\text{grid}}/\rho = 5/3$. To reduce edge effects, the interval $[-3/2, 3/2]$ was used for computations, although only the centre of this is shown in the figure. There is quite good agreement between the numerical computation and the analytical approximation, although the sidelobes decay more rapidly for the numerically computed EK. This is not surprising because the absence of a truly hard cutoff in Fourier space means one should expect less "ringing" than the analytical approximation predicts. The figure also shows good agreement between the weight function (based on the finite sample) and the numerically computed EK. The insets show the approximation of Proposition 2 to $g(z)$ for $\rho = 100$ ($a = 5.67$, left) and $\rho = 10^4$ ($a = 9.67$, right). As expected, the addition of the $1/a^2$-correction gives better agreement with the numerical result for $z < a$. Numerical experiments also show that the mean squared error between the numerically computed EK and the sinc approximation decreases like $1/\log(\rho)$. The is larger than the naïve estimate $(1/a^2)^2 \sim 1/(\log(\rho))^4$ based on the first correction term from Proposition 2, because the dominant part of the error comes from the region $z > a$ where the $1/a$ expansion breaks down.

3.2 Other kernels

Our analysis is not in fact restricted to the SE kernel. First of all, it trivially extends to automatic-relevance determination kernels, which are obtained from

[3] To understand this scaling of σ^2_{grid} consider the case where $\rho_{\text{grid}} > \rho$ which means that the effective variance at each of the ρ_{grid} points per unit $\mathbf{x}$-space is larger, but as there are correspondingly more points this effect cancels out. This can be understood by imagining the situation where there are ρ_{grid}/ρ independent Gaussian observations with variance σ^2_{grid} at a single $\mathbf{x}$-point; this would be equivalent to one Gaussian observation with variance σ^2. In effect the ρ observations per unit $\mathbf{x}$-space have been smoothed out uniformly.

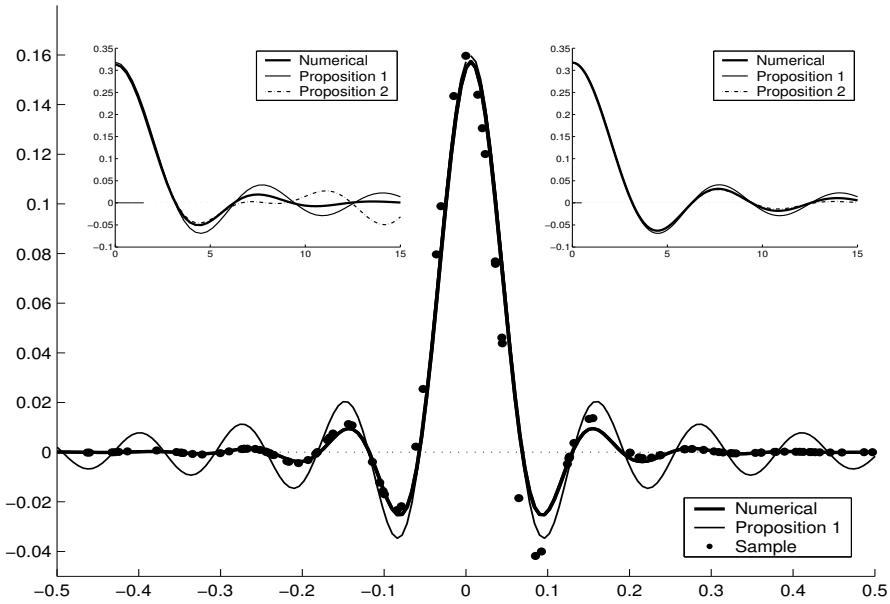

Fig. 2. Main figure: plot of the weight function $\rho \mathbf{h}(x_*)$ corresponding to $\rho = 100$ training points per unit length, plus the numerically computed equivalent kernel at $x_* = 0$ and the sinc approximation from Proposition 1. Insets: numerically evaluated $g(z)$ together with sinc and Proposition 2 approximations for $\rho = 100$ (left) and $\rho = 10^4$ (right).

the SE kernel by allowing separate lengthscales for each input dimension, i.e.

$$k(\mathbf{x}) = \exp\left(-\frac{x_1^2}{2\ell_1^2} - \cdots - \frac{x_D^2}{2\ell_D^2}\right)$$

One can write this as $k(\mathbf{x}) = \exp(-||\mathbf{L}^{-1}\mathbf{x}||^2/2)$ with $\mathbf{L}$ a diagonal matrix containing the lengthscales $\ell_1, \ldots \ell_D$. This could be further extended to arbitrary linear transformations on input space, where $\mathbf{L}$ also has nonzero off-diagonal elements. Making the replacement $\tilde{\mathbf{x}} = \mathbf{L}^{-1}\mathbf{x}$ in the Fourier transform defining the power spectrum $S(\mathbf{s})$, and following through how this affects the EK as defined in equation (5), one easily finds that

$$h(\mathbf{x}) = |\mathbf{L}|^{-1}\, h_1\big(\mathbf{L}^{-1}\mathbf{x}, \rho|\mathbf{L}|\big) \tag{8}$$

where $h_1(\mathbf{x}, \rho)$ is the EK for the isotropic case with $\ell = 1$ and data point density ρ. Equation (8) tells us that the EK is stretched or squashed by exactly the same transformation matrix $\mathbf{L}$ as the underlying covariance kernel. The scaling of the density ρ by the determinant $|\mathbf{L}|$ is also reasonable: what is relevant is the number of data points per "correlation volume". The latter can be defined e.g.

as the size of the region in $\mathbf{x}$-space where $k(\mathbf{x})$ is above some threshold value, and is proportional to $|\mathbf{L}|$. In the isotropic case one has $|\mathbf{L}| = \ell^D$, and the general result in equation (8) is consistent with the fact that, in our earlier results, ρ always appeared in the combination $\rho\ell^D$.

The identity (8) holds for linear transformations applied to any isotropic kernel. In the following we therefore only consider the latter; the power spectrum $S(\mathbf{s})$ then depends on $s = |\mathbf{s}|$ only. We can again define from equation (5) an effective cutoff s_c on the range of s in the EK via $\sigma^2/\rho = S(s_c)$, so that $\tilde{h}(s) = [1+S(s_c)/S(s)]^{-1}$. The EK will then have the limiting form given in Proposition 1 if $\tilde{h}(s)$ approaches a step function $\Theta(s_c - s)$, i.e. if it becomes infinitely "steep" around the point $s = s_c$ for $s_c \to \infty$. A quantitative criterion for this is that the slope $|\tilde{h}'(s_c)|$ should become much larger than $1/s_c$, the inverse of the range of the step function. Since $\tilde{h}'(s) = S'(s)S(s_c)S^{-2}(s)[1 + S(s_c)/S(s)]^{-2}$, this is equivalent to requiring that $-s_c S'(s_c)/4S(s_c) \propto -d\log S(s_c)/d\log s_c$ must diverge for $s_c \to \infty$. The result of Proposition 1 therefore applies to *any* kernel whose power spectrum $S(s)$ decays more rapidly than any positive power of $1/s$.

A trivial example of a kernel obeying this condition would be a superposition of finitely many SE kernels with different lengthscales ℓ^2; the asymptotic behaviour of s_c is then governed by the smallest ℓ. A less obvious case is the "rational quadratic" $k(r) = [1 + (r/l)^2]^{-(D+1)/2}$ which has an exponentially decaying power spectrum $S(s) \propto \exp(-2\pi\ell s)$. (This relationship is often used in the reverse direction, to obtain the power spectrum of the Ornstein-Uhlenbeck (OU) kernel $\exp(-r/\ell)$.) Proposition 1 then applies, with the width of the EK now scaling as $1/s_c \propto 1/\log(\rho)$.

The previous example is a special case of kernels which can be written as superpositions of SE kernels with a distribution $p(\ell)$ of lengthscales ℓ, $k(r) = \int \exp(-r^2/2\ell^2)p(\ell)\,d\ell$. This is in fact the most general representation for an isotropic kernel which defines a valid covariance function in any dimension D, see [10, §2.10]. Such a kernel has power spectrum

$$S(s) = (2\pi)^{D/2} \int_0^\infty \ell^D \exp(-2\pi^2\ell^2 s^2) p(\ell)\,d\ell \tag{9}$$

and one easily verifies that the rational quadratic kernel, which has $S(s) \propto \exp(-2\pi\ell_0 s)$, is obtained for $p(\ell) \propto \ell^{-D-2}\exp(-\ell_0^2/2\ell^2)$. More generally, because the exponential factor in equation (9) acts like a cutoff for $\ell > 1/s$, one estimates $S(s) \sim \int_0^{1/s} \ell^D p(\ell)\,d\ell$ for large s. This will decay more strongly than any power of $1/s$ for $s \to \infty$ if $p(\ell)$ itself decreases more strongly than any power of ℓ for $\ell \to 0$. Any such choice of $p(\ell)$ will therefore yield a kernel to which Proposition 1 applies.

3.3 Non-Uniform Input Densities

We next discuss how the above results generalize to the case where the input density $p(\mathbf{x})$ is not uniform. The smoothed version of the functional $J[f]$ in

equation (2) is now

$$J_\rho[f] = \frac{n}{2\sigma^2}\int(\eta(\mathbf{x}) - f(\mathbf{x}))^2 p(\mathbf{x})d\mathbf{x} + \frac{1}{2}\|f\|^2_{\mathcal{H}}, \tag{10}$$

To minimize this over f one decomposes the latter into eigenfunctions of the covariance kernel, rather than Fourier modes as before. The eigenfunctions are defined by the property

$$\int k(\mathbf{x},\mathbf{x}')\phi_s(\mathbf{x}')p(\mathbf{x}')d\mathbf{x}' = \lambda_s\phi_s(\mathbf{x})$$

where the λ_s are the associated eigenvalues. We index both by a subscript s to emphasize the similarity with the Fourier wavevector $\mathbf{s}$ we had so far; in particular, λ_s is the analogue of $S(\mathbf{s})$ above[4]. The eigenfunctions ϕ_s can always be chosen as normalized and orthogonal with respect to the input density, so that $\int \phi_s(\mathbf{x})\phi_{s'}(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \delta_{ss'}$ and the covariance function has the decomposition $k(\mathbf{x},\mathbf{x}') = \sum_s \lambda_s\phi_s(\mathbf{x})\phi_s(\mathbf{x}')$. In terms of the components of f and η along the eigenfunctions, $\tilde{f}_s = \int f(\mathbf{x})\phi_s(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ and similarly for $\tilde{\eta}_s$, the smoothed functional (10) can then be written as

$$J_\rho[f] = \frac{n}{2\sigma^2}\sum_s(\tilde{\eta}_s - \tilde{f}_s)^2 + \frac{1}{2}\sum\frac{\tilde{f}_s^2}{\lambda_s}\ .$$

Minimization over the $\tilde{f}_s$ then gives $\tilde{f}_s = \tilde{\eta}_s/[1+\sigma^2/(n\lambda_s)]$ or, after reassembling $f(\mathbf{x}) = \sum_s \tilde{f}_s\phi_s(\mathbf{x})$,

$$f(\mathbf{x}_*) = \sum_s \frac{\tilde{\eta}_s\phi_s(\mathbf{x}_*)}{1+\sigma^2/(n\lambda_s)} = \int h(\mathbf{x}_*,\mathbf{x})p(\mathbf{x})\eta(\mathbf{x})d\mathbf{x}' \tag{11}$$

where the equivalent kernel is now defined as

$$h(\mathbf{x}_*,\mathbf{x}) = \sum_s \frac{1}{1+\sigma^2/(n\lambda_s)}\,\phi_s(\mathbf{x}_*)\phi_s(\mathbf{x})\ . \tag{12}$$

One sees from this that, in general, the EK $h(\mathbf{x}_*,\mathbf{x})$ for non-uniform input densities is a function of its two arguments separately rather than just of their difference $\mathbf{x}_* - \mathbf{x}$ as in the uniform case. Also, the eigenfunctions $\phi_s(\mathbf{x})$ depend in a nontrivial manner on the input density and will not be known a priori.

To gain more insight into the behaviour of the EK for non-uniform input densities we now consider the specific case of one-dimensional inputs x with

[4] More precisely, if the input density $p(\mathbf{x})$ is uniform ($= 1/V$) over a large cubic box of size $V = L^D$, then the eigenfunctions $\phi_{\mathbf{s}}(\mathbf{x}) = \exp(2\pi i\mathbf{s}\cdot\mathbf{x})$ are indexed by Fourier wavevectors $\mathbf{s}$ whose components are multiples of $1/L$, and the eigenvalues are related to the power spectrum by $\lambda_{\mathbf{s}} = V^{-1}S(\mathbf{s})$. The resulting EK (12) reproduces the one derived earlier in (5) once it is multiplied by $p(\mathbf{x}) = 1/V$ and the limit $V \to \infty$ is taken; see also the discussion later in the text.

Gaussian density $p(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ and a SE kernel. In that case the eigenfunctions are known to be [11]

$$\phi_s(x) = c^{1/4}(2^{s-1}s!)^{-1/2} e^{-(c-1/4)x^2} H_s(\sqrt{2c}x) \qquad s = 0,\ 1, \ldots,$$

where $c = [1/16 + 1/(4l^2)]^{1/2}$ and $H_s(\cdot)$ is the s-th Hermite polynomial. The associated eigenvalues decay exponentially, $\lambda_s = l[2l(c - 1/4)]^{2k+1}$. The fraction in equation (12) again drops from $\simeq 1$ to $\simeq 0$, around the eigenfunction index s_c where $\lambda_{s_c} \simeq \sigma^2/n$. One can show that this drop becomes increasingly steep as $n\lambda_0/\sigma^2$ grows large, and imposing a hard cutoff at s_c according to $h(\mathbf{x}, \mathbf{x}') \approx \sum_{s=0}^{s_c} \phi_s(\mathbf{x})\phi_s(\mathbf{x}')$ should then give a good approximation to the EK. This approximation can in fact be evaluated in closed form because the eigenfunctions are, apart from s-independent factors, normalized orthogonal polynomials. The Christoffel-Darboux formula [12, eq. 22.12.1] then gives for the hard cutoff approximation to the EK

$$h(x, x') \approx \sqrt{\frac{s_c + 1}{4c}} \, \frac{\phi_{s_c+1}(x)\phi_{s_c}(x') - \phi_{s_c}(x)\phi_{s_c+1}(x')}{x - x'} \ . \tag{13}$$

In the results below, we have also calculated the unapproximated EK from equation (12), using the exact eigenfunctions and eigenvalues. We have compared this with a grid-based calculation: if K is the covariance matrix on a grid of density ρ_{grid}, one can show that the appropriate estimate for the product $h(\mathbf{x}, \mathbf{x}')p(\mathbf{x}')$ at the grid points is $\rho_{\text{grid}} K[K + \sigma^2 \rho_{\text{grid}}(nP)^{-1}]^{-1}$ where P is the diagonal matrix containing $p(\mathbf{x})$ at the grid points. This is directly analogous to the grid estimator we used for uniform input density, except for the replacement of the data point density ρ by nP. Numerically the grid estimate appears more robust, in particular for small ℓ where a larger number of eigenfunctions contribute to the EK.

A new issue in the case of non-uniform input densities is that both the EK $h(\mathbf{x}_*, \mathbf{x})$ and the combination $h(\mathbf{x}_*, \mathbf{x})p(\mathbf{x})$ are meaningful. The former should approximate the weights $\mathbf{h}(\mathbf{x}_*)$ one obtains for predicting at $\mathbf{x}$ for a given training sample of n discrete points. Indeed, the integral on the right-hand side of equation (11) is approximated by the discrete sum $(1/n)\sum_i h(\mathbf{x}_*, \mathbf{x}_i)y_i$ so that $h(\mathbf{x}_*, \mathbf{x}_i)$ approximates (n times) the weight given to training output y_i. Figure 3 shows that this correspondence holds rather well.

The fact that the combination of the EK with the input density, $h(\mathbf{x}_*, \mathbf{x})p(\mathbf{x})$, could be relevant is suggested by comparing the EK prediction (4) for the uniform case with its non-uniform analogue (11). This shows that $h(\mathbf{x}_*, \mathbf{x})p(\mathbf{x})$ plays the role of an effective smoothing kernel applied to the target function $\eta(\mathbf{x})$. We plot this combination in figure 4 and compare it to the result of the hard cutoff approximation (13), for sample size $n = 100$ and noise level $\sigma^2 = 0.1$. On the left, where the covariance function lengthscale is $\ell = 0.5$, the approximation is reasonable; as in the case of uniform input density, making the cutoff hard produces more ringing. In the plot on the right, where $\ell = 0.2$, this effect is rather more pronounced. This is consistent with the fact that the hard cutoff

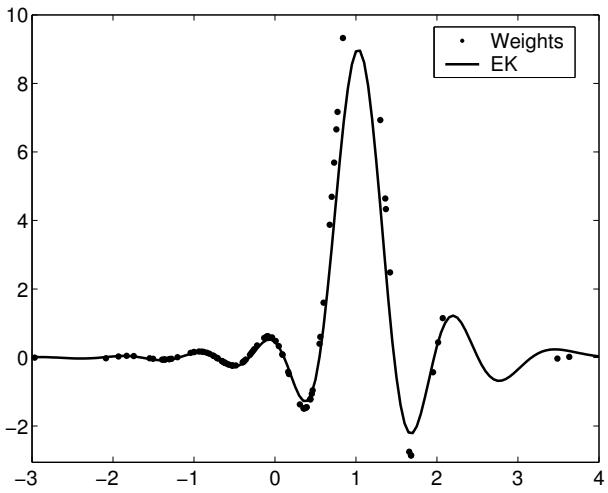

Fig. 3. The EK for Gaussian input density $p(x)$ and an SE covariance function with $\ell = 0.5$, at noise level $\sigma^2 = 0.1$ and for $n = 100$ data points. Solid line: numerically calculated EK $h(x_*, x)$ for $x_* = 1$. Markers: weights $\mathbf{h}(x_*)$ for prediction at the same point, for a random sample of $n = 100$ training inputs. The weights are multiplied by a factor of n to show the correspondence with the EK.

approximation should improve as $n\lambda_0/\sigma^2$ becomes large: this ratio is ≈ 390 for the situation on the left but ≈ 181 on the right.

Finally, one suspects that the EK for non-uniform input densities should reduce to the one for uniform densities if it is sufficiently peaked. More precisely, if the combination $h(\mathbf{x}_*, \mathbf{x})p(\mathbf{x})$ is concentrated in a region $\mathbf{x} \approx \mathbf{x}_*$ that is sufficiently small for $p(\mathbf{x})$ to be regarded as constant there, then it should coincide with the corresponding EK $h(\mathbf{x}_* - \mathbf{x})$ for a *uniform* input density of $\rho = np(\mathbf{x}_*)$. If ρ is large enough, one should be able to approximate further by using the asymptotic form of the EK from Proposition 1.

We test this intuition in figure 5, for a covariance function with lengthscale $\ell = 0.2$. For prediction at $x_* = 0, 1$ and 2 we observe that the EK calculated for uniform ρ provides a good approximation to the EK for the actual non-uniform $p(x)$, and the quality of asymptotic form from Proposition 1 is similar to the uniform case. As $|x_*|$ increases, the approximation of uniform density becomes worse. This arises from two trends. On the one hand, the width of the EK itself increases. On the other, the lengthscale over which $p(x)$ varies – which is $\sim 1/x$ for $x > 1$ – decreases. Eventually these two lengths therefore become comparable, and the variation of $p(x)$ can then no longer be neglected.

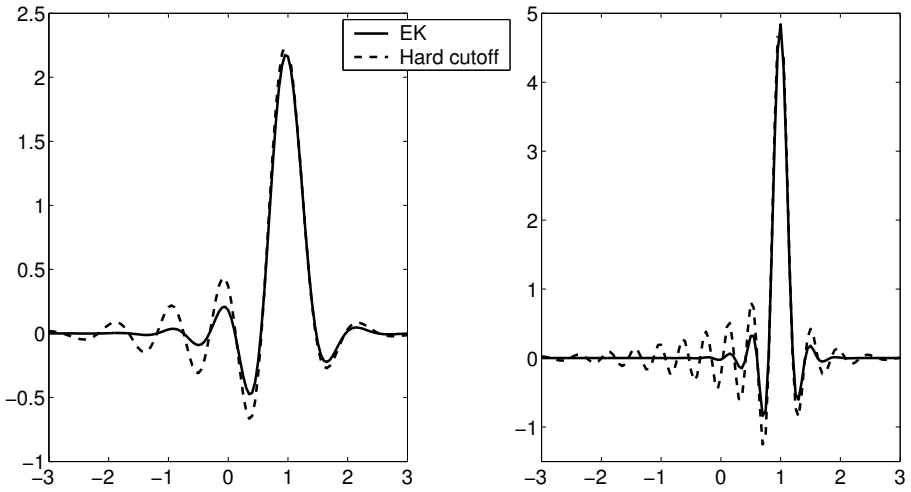

Fig. 4. EK times input density, $h(x_*, x)p(x)$, for Gaussian $p(x)$, $n = 100$, $\sigma^2 = 0.1$, and prediction point $x_* = 1$. The lengthscale of the covariance function is $\ell = 0.5$ on the left, and $\ell = 0.2$ on the right. The dashed lines show the corresponding results for the hard cutoff approximation (13).

4 Understanding GP Learning Using the Equivalent Kernel

We now turn to using EK analysis to get a handle on average case learning curves for Gaussian processes. Here the setup is that a function η is drawn from a Gaussian process, and we obtain ρ noisy observations of η per unit $\mathbf{x}$-space at random $\mathbf{x}$ locations; note that for simplicity we revert to the case of uniform input density in this section. We are concerned with the mean squared error (MSE) between the GP prediction $\overline{f}$ and η. Averaging over the noise process, the $\mathbf{x}$-locations of the training data and the prior over η we obtain the average MSE $\overline{\epsilon}$ as a function of ρ. See e.g. [13] and [14] for an overview of earlier work on GP learning curves.

To understand the asymptotic behaviour of $\overline{\epsilon}$ for large ρ, we now approximate the true GP predictions with the EK predictions from noisy data, given by $f_{\mathrm{EK}}(\mathbf{x}) = \int h(\mathbf{x} - \mathbf{x}')y(\mathbf{x}')d\mathbf{x}'$ in the continuum limit of "smoothed out" input locations. We assume as before that $y =$ target + noise, i.e. $y(\mathbf{x}) = \eta(\mathbf{x}) + \nu(\mathbf{x})$ where $\mathbb{E}[\nu(\mathbf{x})\nu(\mathbf{x}')] = (\sigma_*^2/\rho)\delta(\mathbf{x} - \mathbf{x}')$. Here σ_*^2 denotes the true noise variance, as opposed to the noise variance assumed in the EK; the scaling of σ_*^2 with ρ is explained in footnote 1. For a fixed target η, the MSE is $\epsilon = (\int d\mathbf{x})^{-1} \int [\eta(\mathbf{x}) - f_{\mathrm{EK}}(\mathbf{x})]^2 d\mathbf{x}$. Averaging over the noise process ν and target function η gives in

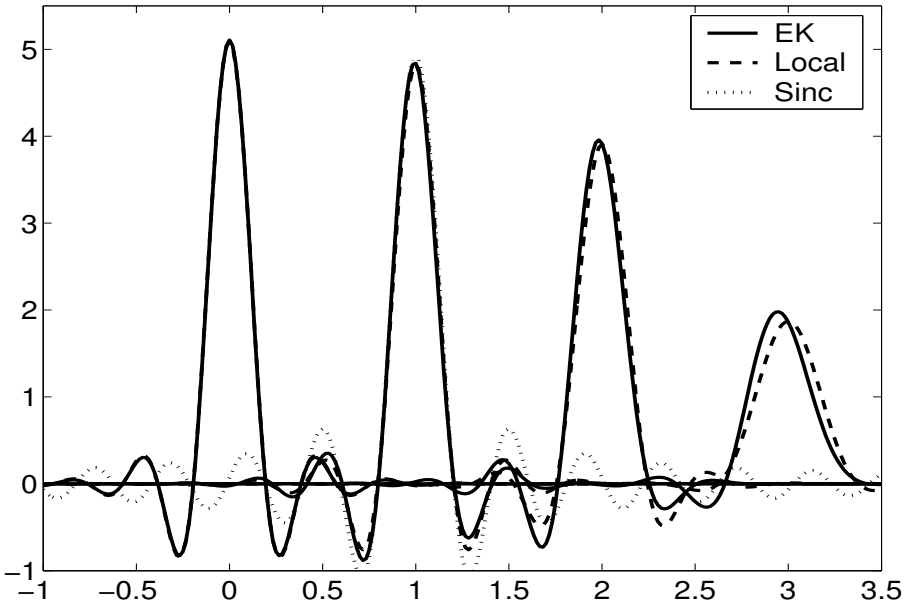

Fig. 5. EK times input density, $h(x_*, x)p(x)$ for Gaussian $p(x)$, $\ell = 0.2$, $n = 100$, $\sigma^2 = 0.1$, and $x_* = 0, 1, 2, 3$. Shown are the numerically calculated EK (solid lines), the EK calculated for a constant data point density $\rho = np(x_*)$ equal to the local density at x_* (dashed), and the approximation from Proposition 1 evaluated for the same ρ (dotted, shown for $x_* = 1$ only).

Fourier space

$$\begin{aligned}\overline{\epsilon} &= \int \left\{ S_\eta(\mathbf{s})[1 - \tilde{h}(\mathbf{s})]^2 + (\sigma_*^2/\rho)\tilde{h}^2(\mathbf{s}) \right\} d\mathbf{s} \\ &= \frac{\sigma^2}{\rho} \int \frac{(\sigma^2/\rho)S_\eta(\mathbf{s})/S^2(\mathbf{s}) + \sigma_*^2/\sigma^2}{[1 + \sigma^2/(\rho S(\mathbf{s}))]^2}\, d\mathbf{s} \end{aligned} \tag{14}$$

where $S_\eta(\mathbf{s})$ is the power spectrum of the prior over target functions. In the case $S(\mathbf{s}) = S_\eta(\mathbf{s})$ and $\sigma^2 = \sigma_*^2$ where the kernel is exactly matched to the structure of the target, equation (14) gives the Bayes error $\overline{\epsilon}_{\mathrm{B}}$ and simplifies to

$$\overline{\epsilon}_{\mathrm{B}} = (\sigma^2/\rho) \int [1 + \sigma^2/(\rho S(\mathbf{s}))]^{-1} d\mathbf{s} \tag{15}$$

(see also [6, eq. 14-16]). Interestingly, this is just the analogue (for a continuous power spectrum of the kernel rather than a discrete set of eigenvalues) of the lower bound of [14] on the MSE of standard GP prediction from finite datasets. In experiments this bound provides a good approximation to the actual average MSE for large dataset size n [13]. This supports our approach of using the EK to understand the learning behaviour of GP regression.

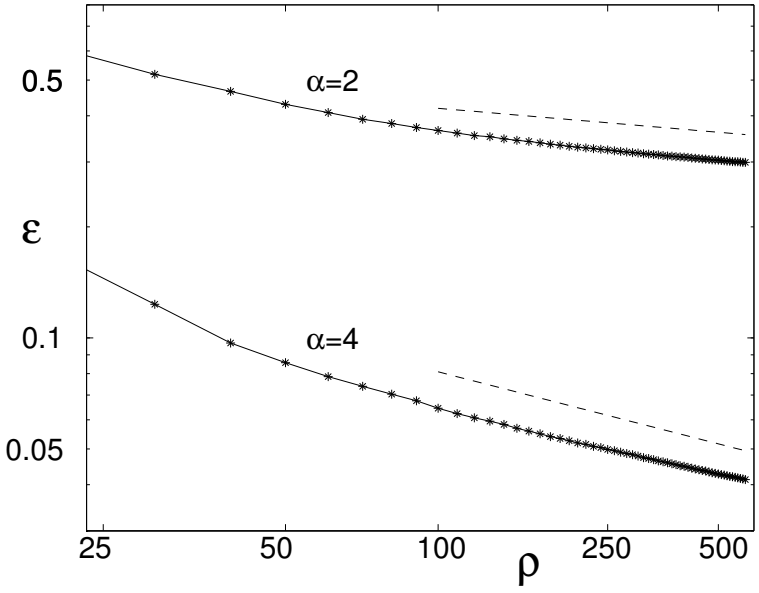

Fig. 6. Log-log plot of $\overline{\epsilon}$ against $\log(\rho)$ for the OU and Matern-class processes ($\alpha = 2,\ 4$ respectively). The dashed lines have gradients of $-1/2$ and $-3/2$ which are the predicted rates.

Treating the denominator in the expression (3) for $\overline{\epsilon}_{\mathrm{B}}$ again as a hard cutoff at $s = s_c$, which is justified for large ρ, one obtains for an SE target and learner $\overline{\epsilon} = \overline{\epsilon}_{\mathrm{B}} \approx \sigma^2 s_c/\rho \propto (\log(\rho))^{D/2}/\rho$. Note that the Bayes error $\overline{\epsilon}_{\mathrm{B}}$ also indicates the mean-squared size of the errorbar of our predictions, i.e. the predictive variance, whether or not kernel and noise level match the target. This can be seen from the terms quadratic in f in the functional $J_\rho[f]$, equation (3).

To get analogous predictions of the MSE for the mismatched case, one can write equation (14) as

$$\overline{\epsilon} = \frac{\sigma_*^2}{\rho} \int \frac{[1 + \sigma^2/(\rho S(\mathbf{s}))] - \sigma^2/(\rho S(\mathbf{s}))}{[1 + \sigma^2/(\rho S(\mathbf{s}))]^2}\, d\mathbf{s} + \int \frac{S_\eta(\mathbf{s})}{[S(\mathbf{s})\rho/\sigma^2 + 1]^2}\, d\mathbf{s}.$$

The first integral is smaller than $(\sigma_*^2/\sigma^2)\overline{\epsilon}_{\mathrm{B}}$ and can be neglected as long as $\overline{\epsilon} \gg \overline{\epsilon}_{\mathrm{B}}$. In the second integral we can again make the cutoff approximation—though now with s having to be *above* s_c – to get the scaling $\overline{\epsilon} \propto \int_{s_c}^{\infty} s^{D-1} S_\eta(s)\, ds$. For target functions with a power-law decay $S_\eta(s) \propto s^{-\alpha}$ of the power spectrum at large s this predicts $\overline{\epsilon} \propto s_c^{D-\alpha} \propto (\log(\rho))^{(D-\alpha)/2}$. So we generically get slow logarithmic learning, consistent with the observations in [15]. For $D = 1$ and an OU target ($\alpha = 2$) we obtain $\overline{\epsilon} \sim (\log(\rho))^{-1/2}$, and for the Matern-class covariance function $k(r) = (1 + r/\ell)\exp(-r/\ell)$ (which has power spectrum $\propto (3/\ell^2 + 4\pi^2 s^2)^{-2}$, so $\alpha = 4$) we get $\overline{\epsilon} \sim (\log(\rho))^{-3/2}$. These predictions were

tested experimentally using a GP learner with SE covariance function ($\ell = 0.1$ and assumed noise level $\sigma^2 = 0.1$) against targets from the OU and Matern-class priors (with $\ell = 0.05$) and with noise level $\sigma_*^2 = 0.01$, averaging over 100 replications for each value of ρ. To demonstrate the predicted power-law dependence of $\overline{\epsilon}$ on $\log(\rho)$, in Figure 6 we make a log-log plot of $\overline{\epsilon}$ against $\log(\rho)$. The dashed lines show the gradients of $-1/2$ and $-3/2$ and we observe good agreement between experimental and theoretical results for large ρ. We note that the predictive variance $\overline{\epsilon}_{\mathrm{B}} \sim 1/\rho$ (up to log-factors, see above) decays much faster than the true MSE $\overline{\epsilon}$, illustrating the possibility of overconfident predictions in mismatched scenarios.

5 Using the Equivalent Kernel in Kernel Regression

Above we have used the EK to understand how standard GP regression works. One could alternatively envisage using the EK to perform kernel regression, on given finite data sets, producing a prediction $\rho^{-1}\sum_i h(\mathbf{x}_* - \mathbf{x}_i)y_i$ at $\mathbf{x}_*$. Intuitively this seems appealing as a cheap alternative to full GP regression, particularly for kernels such as the SE where the EK can be calculated analytically, at least to a good approximation. We now analyse how such an EK predictor would perform compared to standard GP prediction.

Letting $\langle\cdot\rangle$ denote averaging over noise, training input points and the test point and setting $f_\eta(\mathbf{x}_*) = \int h(\mathbf{x}, \mathbf{x}_*)\eta(\mathbf{x})d\mathbf{x}$, the average MSE of the EK predictor is

$$\begin{aligned}
\overline{\epsilon}_{\mathrm{pred}} &= \left\langle \left[\eta(\mathbf{x}) - \frac{1}{\rho}\sum_i h(\mathbf{x}, \mathbf{x}_i) y_i \right]^2 \right\rangle \\
&= \left\langle [\eta(\mathbf{x}) - f_\eta(\mathbf{x})]^2 + \frac{\sigma_*^2}{\rho}\int h^2(\mathbf{x}, \mathbf{x}')d\mathbf{x}' \right\rangle \\
&\quad + \frac{1}{\rho}\left\langle \int h^2(\mathbf{x}, \mathbf{x}')\eta^2(\mathbf{x}')d\mathbf{x}' \right\rangle - \frac{1}{\rho}\left\langle f_\eta^2(\mathbf{x}) \right\rangle \\
&= \frac{\sigma^2}{\rho}\int \frac{(\sigma^2/\rho)S_\eta(\mathbf{s})/S^2(\mathbf{s}) + \sigma_*^2/\sigma^2}{[1 + \sigma^2/(\rho S(\mathbf{s}))]^2}\, d\mathbf{s} + \frac{\langle \eta^2 \rangle}{\rho}\int \frac{d\mathbf{s}}{[1 + \sigma^2/(\rho S(\mathbf{s}))]^2}
\end{aligned}$$

Here we have set $\langle \eta^2 \rangle = (\int d\mathbf{x})^{-1}\int \eta^2(\mathbf{x})\, d\mathbf{x} = \int S_\eta(\mathbf{s})\, d\mathbf{s}$ for the spatial average of the squared target amplitude. Taking the matched case, ($S_\eta(\mathbf{s}) = S(\mathbf{s})$ and $\sigma_*^2 = \sigma^2$) as an example, the first term in $\overline{\epsilon}_{\mathrm{pred}}$ (which is the one we get for the prediction from "smoothed out" training inputs, see equation (14)) is of order $\sigma^2 s_c^D/\rho$, while the second one is $\sim \langle \eta^2 \rangle s_c^D/\rho$. Thus both terms scale in the same way, but the ratio of the second term to the first is the signal-to-noise ratio $\langle \eta^2 \rangle/\sigma^2$, which in practice is often large. The EK predictor will then perform significantly worse than standard GP prediction, by a roughly constant factor, and we have confirmed this prediction numerically. This result is somewhat surprising given the good agreement between the weight function $\mathbf{h}(\mathbf{x}_*)$ and the EK that we saw in figure 2, leading to the conclusion that the detailed structure of the weight function is important for optimal prediction from finite data sets.

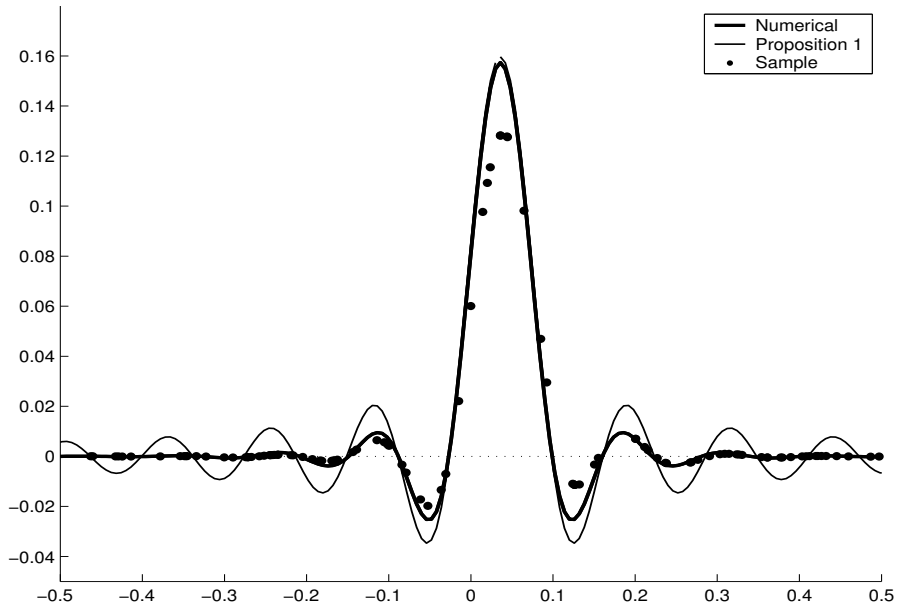

Fig. 7. Plot of the weight function corresponding to $\rho = 100$ training points/unit length (dots), plus the numerically computed equivalent kernel at $x_* = 0.036$ (thick solid line) and the sinc approximation from Proposition 1 (thin line).

One suspects intuitively that the suboptimal performance of the EK smoother must be related to the underlying assumption of uniformly distributed inputs. We therefore show in figure 7 a situation with the same parameters as in figure 2, but now for prediction at $x_* = 0.036$. In the training set for this figure there are two training points very near to this location, at $x = 0.0359$ and $x = 0.0362$. One sees that the effect of this is to depress the weights of these and nearby points significantly, causing deviations from the weights estimated from the EK. This phenomenon is easiest to understand in the limiting case where two training inputs essentially coincide, and there is no output noise. The true GP predictor then halves the weights these points would have if they were far apart: effectively, only one of the two points is used for prediction because the second one contributes no further information about the target function. The EK smoother, on the other hand, produces weights which are not sensitive to the locations of the training points in this way, and effectively overcounts the signal provided by the two nearby inputs. This problem would not be present when inputs are located on a regular grid, and indeed we find numerically that then the EK performs very similarly to the full GP predictor.

If the above picture is correct, then the EK smoother should not only generalize relatively poorly, but also provide a suboptimal fit to the training data.

To check this, we worked out the average training error of the EK smoother,

$$\overline{\epsilon}_{\text{train}} = \frac{1}{n} \sum_i \left\langle \left[y_i - \frac{1}{\rho} \sum_j h(\mathbf{x}, \mathbf{x}_j) y_j \right]^2 \right\rangle$$

with the average taken as before over the noise process, the locations of the training inputs and the prior over the underlying target function η. The averages can be performed as in the calculation of the prediction error, but the number of terms is larger because the case where $i = j$ need to be treated separately. With the abbreviation $H = h(0)/\rho = \rho^{-1} \int [1 + \sigma^2/(\rho S(\mathbf{s}))]^{-1} d\mathbf{s}$ the result can be written as

$$\overline{\epsilon}_{\text{train}} - (\overline{\epsilon}_{\text{pred}} + \sigma_*^2) = H^2(\sigma_*^2 + \langle \eta^2 \rangle) - 2H \left\{ \sigma_*^2 + \int S_\eta(\mathbf{s}) \left[1 - \frac{1}{1 + \sigma^2/(\rho S(\mathbf{s}))} \right] d\mathbf{s} \right\} \tag{16}$$

The difference on the left-hand side is that between the training error and the noisy prediction error. One expects both of these quantities to tend to σ_*^2 for large datasets; the noisy prediction error $\overline{\epsilon}_{\text{pred}} + \sigma_*^2$ clearly approaches the limit from above, while the training error $\overline{\epsilon}_{\text{train}}$ normally does so from below. For the EK smoother one finds, by estimating the dominant term on the right-hand side of equation (16) for the matched case (and for kernels where the approximation of the integrals in terms of a hard cutoff on s works),

$$\frac{\overline{\epsilon}_{\text{train}} - \sigma^2}{(\overline{\epsilon}_{\text{pred}} + \sigma^2) - \sigma^2} \to \frac{\langle \eta^2 \rangle - \sigma^2}{\langle \eta^2 \rangle + \sigma^2} \qquad \text{for} \qquad \rho \to \infty$$

For small noise, $\sigma^2 \ll \langle \eta^2 \rangle$, the training error of the EK smoother thus actually *decreases* towards its limiting value, in the same manner as the noisy prediction error. This is consistent with our expectation that the EK smoother provides a suboptimal fit to the training data. Only for large noise, $\sigma^2 > \langle \eta^2 \rangle$, do we recover the conventional behaviour whereby the training error approaches its limit value from below. This makes sense: in this regime even the full GP predictor will not ignore the second of two outputs corresponding to nearby input points, because significant output noise needs to be averaged out before the target function is learned.

6 Summary and Conclusion

In summary, we have derived accurate approximations for the equivalent kernel (EK) of GP regression with the widely used squared exponential kernel, and have shown that the same analysis in fact extends to a whole class of kernels. We discussed how our results generalize to cases with non-uniform input densities, and saw for the example of a Gaussian density that the resulting EK can often be approximated in a reasonable manner by taking the data point density to be uniform at its local value.

We have also demonstrated that EKs provide a simple means of understanding the learning behaviour of GP regression, even in cases where the learner's covariance function is not well matched to the structure of the target function. In future work, it will be interesting to explore in more detail the use of the EK in kernel smoothing. This is suboptimal compared to standard GP regression as we saw. However, it does remain feasible even for very large datasets, and may then be competitive with sparse methods for approximating GP regression. From the theoretical point of view, the average error of the EK predictor which we calculated may also provide the basis for useful upper bounds on GP learning curves.

Acknowledgments: This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. We are grateful to Manfred Opper for pointing out the Christoffel-Darboux formula used to derive equation (13), and to two anonymous referees for helpful comments.

Bibliography

[1] B. W. Silverman. *Annals of Statistics*, 12:898–916, 1984.

[2] C. K. I. Williams. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 599–621. Kluwer Academic, 1998.

[3] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[4] P. Sollich and C. K. I. Williams. In *NIPS 17*, 2005, to appear.

[5] F. Girosi, M. Jones, and T. Poggio. *Neural Computation*, 7(2):219–269, 1995.

[6] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, 1991. Third Edition.

[7] C. Thomas-Agnan. *Numerical Algorithms*, 13:21–32, 1996.

[8] T. Poggio, H. Voorhees, and A. Yuille. Tech. Report AI Memo 833, MIT AI Laboratory, 1985.

[9] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[10] M. L. Stein. *Interpolation of Spatial Data*. Springer-Verlag, New York, 1999.

[11] H. Zhu, C. K. I. Williams, R. J. Rohwer and M. Morciniec. In C. M. Bishop, editor, *Neural Networks and Machine Learning*. Springer, 1998.

[12] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. Dover, New York, 1965.

[13] P. Sollich and A. Halees. *Neural Computation*, 14:1393–1428, 2002.

[14] M. Opper and F. Vivarelli. In *NIPS 11*, pages 302–308, 1999.

[15] P. Sollich. In *NIPS 14*, pages 519–526, 2002.