Can Gaussian Process Regression Be Made Robust Against Model Mismatch?

Peter Sollich

Department of Mathematics, King's College London
Strand, London WC2R 2LS, U.K.
peter.sollich@kcl.ac.uk

Abstract. Learning curves for Gaussian process (GP) regression can be strongly affected by a mismatch between the 'student' model and the 'teacher' (true data generation process), exhibiting e.g. multiple overfitting maxima and logarithmically slow learning. I investigate whether GPs can be made robust against such effects by adapting student model hyperparameters to maximize the evidence (data likelihood). An approximation for the average evidence is derived and used to predict the optimal hyperparameter values and the resulting generalization error. For large input space dimension, where the approximation becomes exact, Bayes-optimal performance is obtained at the evidence maximum, but the actual hyperparameters (e.g. the noise level) do not necessarily reflect the properties of the teacher. Also, the theoretically achievable evidence maximum cannot always be reached with the chosen set of hyperparameters, and maximizing the evidence in such cases can actually make generalization performance worse rather than better. In lower-dimensional learning scenarios, the theory predicts—in excellent qualitative and good quantitative accord with simulations—that evidence maximization eliminates logarithmically slow learning and recovers the optimal scaling of the decrease of generalization error with training set size.

1 Introduction

Gaussian processes (GPs) are by now a popular alternative to feedforward networks for regression, see e.g. [1–11]. They make prior assumptions about the problem to be learned very transparent, and—even though they are non-parametric models—inference is straightforward. Much work has been done to understand the learning behaviour of GPs as encoded in the learning curve, i.e. the average generalization performance for a given number of training examples [5, 7–10, 12, 13]. This has mostly focused on the case where the 'student' model exactly matches the true 'teacher' generating the data. In practice, such a match is unlikely. In [11] I showed that much richer behaviour then results, with learning curves that can exhibit multiple overfitting maxima, or decay logarithmically slowly if the teacher is less smooth than the student assumes. An intriguing open question was whether these adverse effects of model mismatch can be avoided by adapting the student model during learning. This is the issue I address in

the present paper, focusing on the adaptation of model (hyper-)parameters by maximization of the data likelihood or evidence.

In its simplest form, the regression problem is this: We are trying to learn a function θ_* which maps inputs x (real-valued vectors) to (real-valued scalar) outputs $\theta_*(x)$. A set of training data D consists of n input-output pairs (x^l, y^l); the training outputs y^l may differ from the 'clean' teacher outputs $\theta_*(x^l)$ due to corruption by noise. Given a test input x, we are then asked to come up with a prediction $\hat{\theta}(x)$, plus error bar, for the corresponding output $\theta(x)$. In a Bayesian setting, one does this by specifying a prior $P(\theta)$ over hypothesis functions and a likelihood $P(D|\theta)$ with which each θ could have generated the training data,; from these the posterior distribution $P(\theta|D) \propto P(D|\theta)P(\theta)$ can be deduced. For a GP, the prior is defined directly over input-output functions θ. Any θ is uniquely determined by its output values $\theta(x)$ for all x from the input domain, and for a GP, these are assumed to have a joint Gaussian distribution (hence the name). The means are usually set to zero so that the distribution is fully specified by the *covariance function* $\langle \theta(x)\theta(x') \rangle = C(x, x')$. The latter transparently encodes prior assumptions about the function to be learned. Smoothness, for example, is controlled by the behaviour of $C(x, x')$ for $x' \to x$: The Ornstein-Uhlenbeck (OU) covariance function $C(x, x') = a \exp(-|x - x'|/l)$ produces very rough (non-differentiable) functions, while functions sampled from the radial basis function (RBF) prior with $C(x, x') = a \exp[-|x - x'|^2/(2l^2)]$ are infinitely often differentiable. Here l is a length scale parameter, corresponding directly to the distance in input space over which significant variation in the function values is expected, while a determines the prior variance.

A summary of inference with GPs is as follows (for details see e.g. [14, 15]). The student assumes that outputs y are generated from the 'clean' values of a hypothesis function $\theta(x)$ by adding Gaussian noise of x-independent variance σ^2. The joint distribution of a set of training outputs $\{y^l\}$ and the function values $\theta(x)$ is then also Gaussian, with covariances given (under the student model) by

$$\langle y^l y^m \rangle = C(x^l, x^m) + \sigma^2 \delta_{lm} = (\mathbf{K})_{lm}, \qquad \langle y^l \theta(x) \rangle = C(x^l, x) = (\mathbf{k}(x))_l \quad (1)$$

Here I have defined an $n \times n$ matrix $\mathbf{K}$ and an x-dependent n-component vector $\mathbf{k}(x)$. The posterior distribution $P(\theta|D)$ is obtained by conditioning on the $\{y^l\}$; it is again Gaussian and has mean and variance

$$\langle \theta(x) \rangle_{\theta|D} \equiv \hat{\theta}(x) = \mathbf{k}(x)^{\mathrm{T}} \mathbf{K}^{-1} \mathbf{y} \quad (2)$$

$$\langle [\theta(x) - \hat{\theta}(x)]^2 \rangle_{\theta|D} = C(x, x) - \mathbf{k}(x)^{\mathrm{T}} \mathbf{K}^{-1} \mathbf{k}(x) \quad (3)$$

From the student's point of view, this solves the inference problem: the best prediction for $\theta(x)$ on the basis of the data D is $\hat{\theta}(x)$, with a (squared) error bar given by (3).

The squared deviation between the prediction and the teacher is $[\hat{\theta}(x) - \theta_*(x)]^2$; the average generalization error (which, as a function of n, defines the learning curve) is obtained by averaging this over the posterior distribution of teachers, all datasets, and the test input x:

$$\epsilon = \langle \langle \langle [\hat{\theta}(x) - \theta_*(x)]^2 \rangle_{\theta_*|D} \rangle_D \rangle_x \quad (4)$$

Of course the student does not know the true posterior of the teacher; to estimate ϵ, she must assume that it is identical to the student posterior, giving

$$\hat{\epsilon} = \langle\langle[\hat{\theta}(x) - \theta(x)]^2\rangle_{\theta|D}\rangle_D\rangle_x \tag{5}$$

This generalization error estimate $\hat{\epsilon}$ coincides with the true error ϵ if the student model matches the true teacher model and then gives the Bayes error, i.e. the best achievable average generalization performance for the given teacher.

The *evidence* or data likelihood is $P(D) = \int d\theta P(D|\theta)P(\theta)$, i.e. the average of the likelihood $P(D|\theta) = \prod_{l=1}^{n}(2\pi\sigma^2)^{-1/2}\exp[-(y^l - \theta(x^l))^2/(2\sigma^2)]$ over the prior. Since the prior over the $\theta(x^l)$ is a zero mean Gaussian with covariance matrix $C(x^l, x^m)$, the integral can be done analytically and one finds

$$E \equiv \frac{1}{n}\ln P(D) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2n}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2n}\ln|\mathbf{K}| \tag{6}$$

The (normalized log-) evidence E depends, through $\mathbf{K}$, on all student model hyperparameters, i.e. σ^2 and any parameters specifying the covariance function. I will analyse the model selection algorithm which chooses these parameters, for each data set D, by maximizing E. For one particular hyperparameter the maximum can in fact be found analytically: if we write $C(x, x') = a\tilde{C}(x, x')$ and $\sigma^2 = a\tilde{\sigma}^2$, then the second term in (6) scales as $1/a$ and the third one gives the a-dependent contribution $(1/2)\ln a$; maximizing over a gives $a = n^{-1}\mathbf{y}^{\mathrm{T}}\tilde{\mathbf{K}}^{-1}\mathbf{y}$ and

$$\max_a E = -\frac{1}{2}\ln(2\pi/n) - \frac{1}{2} - \frac{1}{2}\ln(\mathbf{y}^{\mathrm{T}}\tilde{\mathbf{K}}^{-1}\mathbf{y}) - \frac{1}{2n}\ln|\tilde{\mathbf{K}}|$$

Note that the value of a does not affect the student's predictions (2), but only scales the error bars (3).

2 Calculating the evidence

A theoretical analysis of the average generalization performance obtained by maximizing the evidence for each data set D is difficult because the optimal hyperparameter values fluctuate with D. However, a good approximation—at least for not too small n—can be obtained by neglecting these fluctuations, and considering the hyperparameter values that maximize the average $\bar{E}$ of the evidence over all data sets D of given size n produced by the teacher. To perform the average, I assume in what follows that the teacher is also a GP, but with a possibly different covariance function $C_*(x, x')$ and noise level σ_*^2. For fixed training inputs, the average of $y^l y^m$ is then $(\mathbf{K}_*)_{lm} = C_*(x^l, x^m) + \sigma_*^2\delta_{lm}$, and inserting into (6) gives

$$\bar{E} = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2n}\langle\mathrm{tr}\,\mathbf{K}_*\mathbf{K}^{-1}\rangle - \frac{1}{2n}\langle\ln|\sigma^{-2}\mathbf{K}|\rangle \tag{7}$$

where the remaining averages are over the distribution of all possible sets of training inputs. To tackle these, it is convenient to decompose (using Mercer's

theorem) the covariance function into its eigenfunctions $\phi_i(x)$ and eigenvalues Λ_i, defined w.r.t. the input distribution so that $\langle C(x,x')\phi_i(x')\rangle_{x'} = \Lambda_i\phi_i(x)$ with the corresponding normalization $\langle\phi_i(x)\phi_j(x)\rangle_x = \delta_{ij}$. Then

$$C(x,x') = \sum_{i=1}^{\infty} \Lambda_i\phi_i(x)\phi_i(x'), \quad \text{and similarly } C_*(x,x') = \sum_{i=1}^{\infty} \Lambda_i^*\phi_i(x)\phi_i(x') \quad (8)$$

For simplicity I assume here that the student and teacher covariance functions have the *same* eigenfunctions (but different eigenvalues). This is not as restrictive as it may seem; several examples are given below.

Introducing the diagonal eigenvalue matrix $(\mathbf{\Lambda})_{ij} = \Lambda_i\delta_{ij}$ and the 'design matrix' $(\mathbf{\Phi})_{li} = \phi_i(x^l)$, one now has $\mathbf{K} = \sigma^2 + \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^{\mathrm{T}}$, and similarly for $\mathbf{K}_*$. In the second term of (7) we need $\mathrm{tr}\,\mathbf{K}_*\mathbf{K}^{-1}$; the Woodbury formula gives the required inverse as $\mathbf{K}^{-1} = \sigma^{-2}[\mathbf{I} - \sigma^{-2}\mathbf{\Phi}\mathcal{G}\mathbf{\Phi}^{\mathrm{T}}]$, where $\mathcal{G} = (\mathbf{\Lambda}^{-1} + \sigma^{-2}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}$. A little algebra then yields

$$\mathrm{tr}\,\mathbf{K}_*\mathbf{K}^{-1} = -\sigma_*^2\sigma^{-2}\mathrm{tr}\,(\mathbf{I} - \mathbf{\Lambda}^{-1}\mathcal{G}) + \mathrm{tr}\,\mathbf{\Lambda}_*\mathbf{\Lambda}^{-1}(\mathbf{I} - \mathbf{\Lambda}^{-1}\mathcal{G}) + n\sigma_*^2\sigma^{-2} \quad (9)$$

and the training inputs appear only via the matrix $\mathcal{G}$. A similar reduction is possible for the third term of (7). The eigenvalues of the matrix $\sigma^{-2}\mathbf{K} = \mathbf{I} + \sigma^{-2}\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^{\mathrm{T}}$ are easily seen to be the same as the nontrivial ($\neq 1$) ones of $\mathbf{I} + \sigma^{-2}\mathbf{\Lambda}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$, so that $\ln|\sigma^{-2}\mathbf{K}| = \ln|\mathbf{I}+\sigma^{-2}\mathbf{\Lambda}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}|$. If we generalize the definition of $\mathcal{G}$ to $\mathcal{G} = (\mathbf{\Lambda}^{-1} + v\mathbf{I} + \sigma^{-2}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}$ and also define $T(v) = \ln|(\mathbf{\Lambda}^{-1} + v)\mathcal{G}|$, then $T(\infty) = 0$ and so

$$\ln|\sigma^{-2}\mathbf{K}| = \ln|\mathbf{I} + \sigma^{-2}\mathbf{\Lambda}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}| = T(0) - T(\infty) = \int_0^{\infty} dv\ [\mathrm{tr}\,(\mathbf{\Lambda}^{-1} + v)^{-1} - \mathrm{tr}\,\mathcal{G}] \quad (10)$$

Eqs. (9,10) show that all the averages required in (7) are of the form $\langle\mathrm{tr}\,\mathbf{M}\mathcal{G}\rangle$ with some matrix $\mathbf{M}$. We derived an accurate approximation for such averages in [5, 10, 11], with the result $\langle\mathrm{tr}\,\mathbf{M}\mathcal{G}\rangle = \mathrm{tr}\,\mathbf{M}\mathbf{G}$ where

$$\mathbf{G}^{-1} = \mathbf{\Lambda}^{-1} + \left(v + \frac{n}{\sigma^2 + g(n,v)}\right)\mathbf{I} \quad (11)$$

and the function $g(n,v)$ is determined by the self-consistency equation $g = \mathrm{tr}\,\mathbf{G}$. Using this approximation and (9,10) in (7) gives, after a few rearrangements,

$$\bar{E} = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{\sigma_*^2 + \mathrm{tr}\,\mathbf{\Lambda}_*\mathbf{\Lambda}^{-1}\mathbf{G}}{\sigma^2 + g} - \frac{1}{2n}\int_0^{\infty} dv\ [g(0,v) - g(n,v)] \quad (12)$$

where in the second term $\mathbf{G}$ and g are evaluated at $v = 0$. This approximation for the average (normalized log-) evidence is the main result of this paper. The true $\bar{E}$ is known to achieve its maximum value when the student and teacher model are exactly matched, since the deviation from this maximum is essentially the KL-divergence between the student and teacher distributions over data sets. Remarkably, this property is preserved by the approximation (12): a rather lengthy calculation shows that it has a stationary point w.r.t. variation of $\mathbf{\Lambda}$ and σ^2 (which numerically always turns out to be maximum) at $\mathbf{\Lambda} = \mathbf{\Lambda}_*$ and $\sigma^2 = \sigma_*^2$.

3 Examples

If the eigenvalue spectra $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_*$ are known, the approximation (12) for the average evidence can easily be evaluated numerically and maximized over the hyperparameters. As in the case of the unaveraged evidence (6), the maximization over the overall amplitude factor a can be carried out analytically. The resulting generalization performance can then be predicted using the results of [11].

As a first example scenario, consider inputs x which are binary vectors[1] with d components $x_a \in \{-1, 1\}$, and assume that the input distribution is uniform. I consider covariance functions for student and teacher that depend on the product $x \cdot x'$ only; this includes the standard choices (e.g. OU and RBF) which are functions of the Euclidean distance $|x-x'|$, since $|x-x'|^2 = 2d-2x\cdot x'$. All these covariance functions have the same eigenfunctions [17], so our above assumption is satisfied. The eigenfunctions are indexed by subsets ρ of $\{1, 2 \dots d\}$ and given explicitly by $\phi_\rho(x) = \prod_{a\in\rho} x_a$. The corresponding eigenvalues depend only on the size $s = |\rho|$ of the subsets and are therefore $\binom{d}{s}$-fold degenerate; letting $e = (1, 1 \dots 1)$ be the 'all ones' input vector, they can be written as $\Lambda_s = \langle C(x, e)\phi_\rho(x)\rangle_x$. From this the eigenvalues can easily be found numerically for any d, but here I focus on the limit of large d where all results can be obtained in closed form. If we write $C(x, x') = f(x \cdot x'/d)$, the eigenvalues become, for $d \to \infty$, $\Lambda_s = d^{-s} f^{(s)}(0)$ where $f^{(s)}(z) \equiv (d/dz)^s f(z)$. The contribution to $C(x, x) = f(1)$ from the s-th eigenvalue block is then $\lambda_s \equiv \binom{d}{s}\Lambda_s \to f^{(s)}(0)/s!$, consistent with $f(1) = \sum_{s=0}^{\infty} f^{(s)}(0)/s!$. Because of their scaling with d, the Λ_s become infinitely separated for $d \to \infty$. For training sets of size $n = \mathcal{O}(d^L)$, one then sees in (11) that eigenvalues with $s > L$ contribute as if $n = 0$, since $\Lambda_s \gg n/(\sigma^2 + g)$; these correspond to components of the teacher that have effectively not yet been learned [11]. On the other hand, eigenvalues with $s < L$ are completely suppressed and have been learnt perfectly. A hierarchical learning process thus results, where different scalings of n with d—as defined by L—correspond to different 'learning stages'. Formally, one can analyse the stages separately by letting $d \to \infty$ at a constant ratio $\alpha = n/\binom{d}{L}$ of the number of examples to the number of parameters to be learned at stage L; note that $\binom{d}{L} = \mathcal{O}(d^L)$ for large d. A replica calculation along the lines of Ref. [16] shows that the approximation (12) for the average evidence actually becomes *exact* in this limit. Fluctuations in E across different data sets also tend to zero so that considering $\bar{E}$ rather than E introduces no error.

Intriguingly, the resulting exact expression for the evidence at stage L turns out to depend only on two functions of the student hyperparameters. Setting $f_L = \sum_{s\geq L} \lambda_s$ (so that $f_0 = f(1)$), they are $f_{L+1} + \sigma^2$ and λ_L. The learning curve analysis in [11] showed that these correspond, respectively, to the student's

[1] This assumption simplifies the determination of the eigenfunctions and eigenvalues. For large d, one expects distributions with continuously varying x and the same first- and second-order statistics to give similar results [16]. A case where this can be shown explicitly is that of a uniform distribution over input vectors x of fixed length, which gives spherical harmonics as eigenfunctions.

assumed values for the effective level of noise and for the signal to be learnt in the current stage. Independently of the number of training examples α, the evidence as calculated above can be shown to be maximal when these two parameters match the true values for the teacher, and it follows from the results of [11] that the resulting generalization error is then optimal, i.e. equal to the Bayes error. This implies in particular that overfitting maxima cannot occur.

A first implication of the above analysis is that even though evidence maximization can ensure optimal generalization performance, the resulting hyperparameter values are not meaningful as estimates of the underlying 'true' values of the teacher. Consider e.g. the case where the student assumes an OU covariance function, i.e. $C(x, x') = \exp[-|x - x'|/(ld^{1/2})]$ and therefore $f(z) = \exp[-\sqrt{2 - 2z}/l]$, but the teacher has an RBF covariance function, for which $C_*(x, x') = \exp[-|x - x'|^2/(2l^2 d)]$ and $f_*(z) = \exp[-(1 - z)/l^2]$. The length scales have been scaled by $d^{1/2}$ here to get sensible behaviour for $d \to \infty$. Then one has, for example,

$$\lambda_0 = e^{-\sqrt{2}/l}, \quad \lambda_1 = \lambda_0/(\sqrt{2}l), \qquad \lambda_0^* = e^{-1/l_*^2}, \quad \lambda_1^* = \lambda_0^*/l_*^2$$

For a given teacher length scale l_*, the optimal value of the student length scale l determined from the criterion $\lambda_L = \lambda_L^*$ will therefore generally differ from l_*, and actually depend on the learning stage L. Similarly, the optimal student noise level will not be identical to the true teacher noise level. At stage $L = 1$, for example, the optimal choice of length scale implies $\lambda_1 = \lambda_1^*$; but then $f_2 = 1 - \lambda_0 - \lambda_1$ will differ from f_2^* and the optimality condition $f_2 + \sigma^2 = f_2^* + \sigma_*^2$ tells us that $\sigma^2 \neq \sigma_*^2$.

A second interesting feature is that, since the λ_L and f_L depend in a complicated way on the hyperparameters, the optimality conditions $\lambda_L = \lambda_L^*$ and $f_{L+1} + \sigma^2 = f_{L+1}^* + \sigma_*^2$ may have more than one solution, or none at all, depending on the situation. An example is shown in Fig. 1. For a noise free ($\sigma_*^2 = 0$) RBF teacher with $l_* = 0.55$, one has $\lambda_1^* = 0.121$ and at learning stage $L = 1$ there are two very different optimal assignments for the student length scale, $l = 0.639$ and $l = 4.15$ (marked by arrows in Fig. 1) which achieve $\lambda_1(l) = \lambda_1^*$. The corresponding optimal noise levels are also very different at $\sigma^2 = 0.0730$ and $\sigma^2 = 0.674$, respectively. At stage $L = 2$, on the other hand, $\lambda_2^* = 0.2004$ and there is no value of the student length scale l for which $\lambda_2(l) = \lambda_2^*$. One finds that the evidence is in this case maximized by choosing l as large as possible. With l large, all λ_i for $i > 0$ are very small, and the student's assumed effective noise-to-signal ratio $(f_3 + \sigma^2)/\lambda_2$ becomes large. The results of [11] imply that the generalization error will decay extremely slowly in this case, and in fact not at all in the strict limit $l \to \infty$. Here we therefore have a case where strongly suboptimal performance results from evidence maximization, for the reason that the 'ideal' evidence maximum cannot be reached by tuning the chosen hyperparameters. In fact evidence maximization performs worse than learning with *any* fixed set of hyperparameters! Including a tunable overall amplitude factor a for the student's covariance function and noise level would, for the example values used above, solve this problem, and in fact produce a one-parameter family of optimal

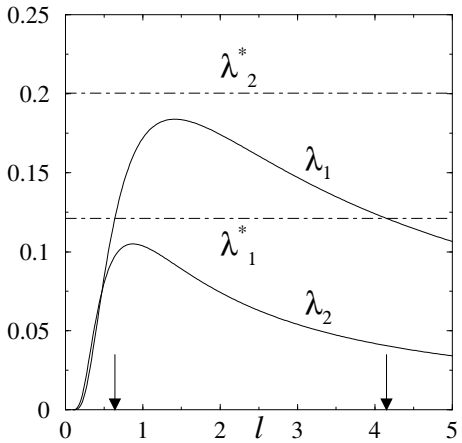

Fig. 1. Illustration of choice of optimal length scale in a scenario with large input space dimension d, for an OU student learning an RBF teacher with length scale $l_* = 0.55$. Evidence maximization gives the optimality criterion $\lambda_L = \lambda_L^*$ for learning stage L. At stage $L = 1$, this has two solutions for the student length scale l, marked by the arrows, while at stage $L = 2$ no solutions exist.

assignments of a, l and σ^2. One might expect this to be the generic situation but even here there are counter-examples: the optimality conditions demand equality of the student's effective noise-to-signal ratio, $\kappa_L = (f_{L+1} + \sigma^2)/\lambda_L$ with that of the teacher. But κ_L is independent of the amplitude factor a and $\geq f_{L+1}/\lambda_L$, and the latter ratio may be bounded above zero, e.g. $f_3/\lambda_2 \geq 3$ for any l for an OU student. For sufficiently low κ_L^* there is then no choice of l for which $\kappa_L = \kappa_L^*$.

In the second example scenario, I consider continuous-valued input vectors, uniformly distributed over the unit interval $[0, 1]$; generalization to d dimensions ($x \in [0, 1]^d$) is straightforward. For covariance functions which are stationary, i.e. dependent on x and x' only through $x - x'$, and assuming periodic boundary conditions (see [10] for details), one then again has covariance function-independent eigenfunctions. They are indexed by integers[2] q, with $\phi_q(x) = e^{2\pi i q x}$; the corresponding eigenvalues are $\Lambda_q = \int dx\, C(0, x) e^{-2\pi i q x}$. For the ('periodified' version of the) RBF covariance function $C(x, x') = a \exp[-(x - x')^2/(2l^2)]$, for example, one has $\Lambda_q \propto \exp(-\tilde{q}^2/2)$, where $\tilde{q} = 2\pi l q$. The OU case $C(x, x') = a \exp(-|x - x'|/l)$, on the other hand, gives $\Lambda_q \propto (1 + \tilde{q}^2)^{-1}$, thus $\Lambda_q \propto q^{-2}$ for large q. I also consider below covariance functions which interpolate in smoothness between the OU and RBF limits. E.g. the MB2 (modified Bessel or Matern class) covariance $C(x, x') = e^{-b}(1 + b)$, with $b = |x - x'|/l$, yields functions which are once differentiable [13, 15]; its eigenvalues $\Lambda_q \propto (1 + \tilde{q}^2)^{-2}$ show a

[2] Since $\Lambda_q = \Lambda_{-q}$, one can assume $q \geq 0$ if all Λ_q for $q > 0$ are taken as doubly degenerate.

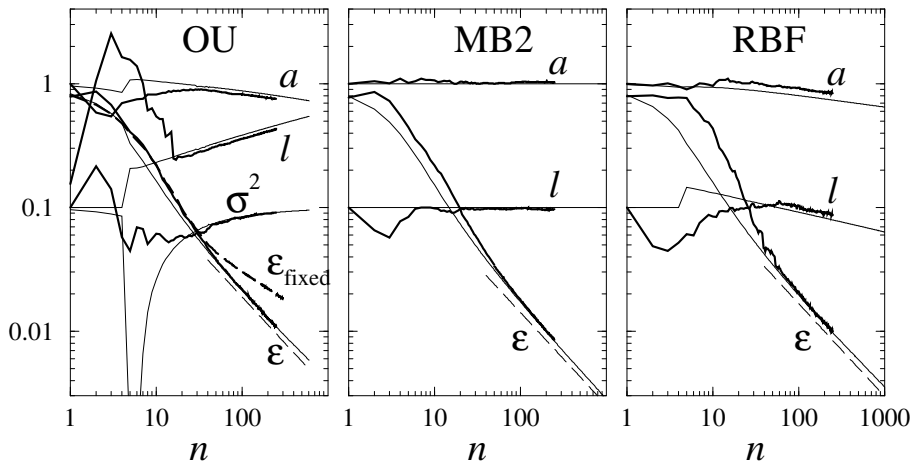

Fig. 2. Evidence maximization for a teacher with MB2 covariance function, $l_* = \sigma_*^2 = 0.1$, and inputs x uniformly distributed over $[0, 1]$. Bold lines: simulation averages over 20 to 50 independent sequences of training examples, thin lines: theory. Left: Hyperparameters a, l, σ^2 and generalization error ϵ vs. number of training examples n for OU student; the more slowly decaying generalization error ϵ_{fixed} for a fixed student model with $l = \sigma^2 = 0.1$, $a = 1$ is also shown. For the MB2 (middle) and RBF (right) students, σ^2 is close to constant at $\sigma^2 = \sigma_*^2$ in both theory and simulation and not shown. Dashed lines indicate the Bayes-optimal scaling of the asymptotic generalization error, $\epsilon \sim n^{-3/4}$, which with evidence maximization is obtained even in the cases with model mismatch (OU and RBF).

faster asymptotic power law decay, $\Lambda_q \propto q^{-4}$, than those of the OU covariance function. Writing the asymptotic behaviour of the eigenvalues generally as $\Lambda_q \propto q^{-r}$, and similarly $\Lambda_q^* \propto q^{-r_*}$, one has $r = 2$ for OU, $r = 4$ for MB2 and, due to the faster-than-power law decay of its eigenvalues, effectively $r = \infty$ for RBF. For the case of a fixed student model [11], the generalization error ϵ then generically decays as a power law with n for large n. If the student assumes a rougher function than the teacher provides ($r < r_*$), the asymptotic power law exponent $\epsilon \propto n^{-(r-1)/r}$ is determined by the student alone. In the converse case, the asymptotic decay is $\epsilon \propto n^{-(r_*-1)/r}$ and can be very slow, actually becoming logarithmic for an RBF student ($r \to \infty$). For $r = r_*$, the fastest decay for given r_* is obtained, as expected from the properties of the Bayes error.

The predictions for the effect of evidence maximization, based on (12), are shown in Fig. 2 for the case of an MB2 teacher ($r_* = 4$) being learned by a student with OU ($r = 2$), MB2 ($r = 4$) and RBF ($r = \infty$) covariance functions. Simulation results, obtained by averaging over 20 to 50 data sets for each n, are also shown. The most striking feature is that the theory predicts that in *all* three cases the generalization error now decays with the optimal power law scaling $\epsilon \sim n^{-(r_*-1)/r_*} = n^{-3/4}$; the simulations are consistent with this. In particular,

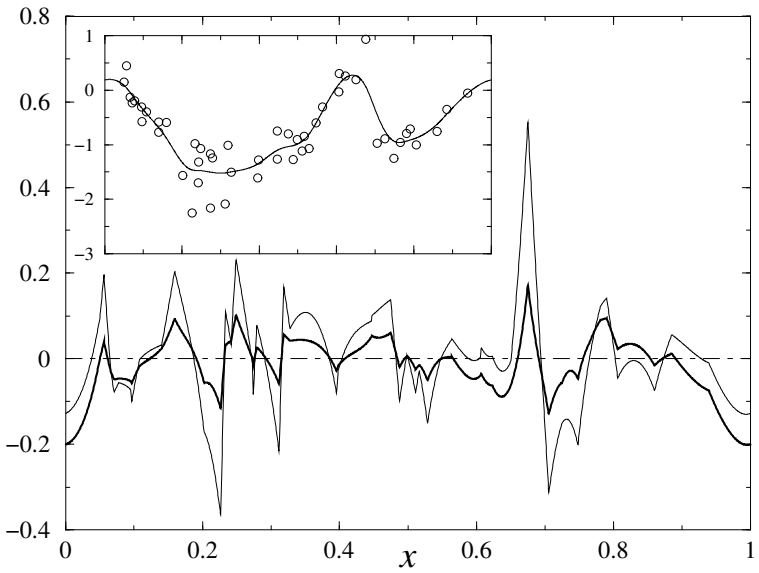

Fig. 3. Effects of evidence maximization for an OU student learning an RBF teacher, for input space dimension $d = 1$. Shown is a sample from one of the runs for $n = 50$ summarized in Fig. 2. Inset: Training data (circles) and Bayes optimal prediction function. Main graph: Difference between the prediction of the OU student and the Bayes-optimal prediction, for hyperparameters set equal to those of the teacher ($\sigma^2 = l = 0.1$, $a = 1$, thin line). Evidence maximization gives a larger l and thus a smoother student prediction that differs less from the Bayes-optimal prediction (bold line).

for the RBF student the logarithmically slow learning has been eliminated. For the case of the MB2 student, the theory predicts that the optimal values of the student hyperparameters are constant and identical to those of the teacher; this is as expected since then the models match exactly. The simulations again agree, though for small n the effects of our approximation of averaging the evidence over data sets and only then maximizing it become apparent.

For the OU student, inspection of the simulation results shows that the evidence maximum can, for some data sets, result in either one of two extreme hyperparameter assignments: $\sigma^2 = 0$, in which case the rough OU covariance function takes all noise on the teacher's underlying smooth target function as genuine signal, or l very large so that the covariance function is essentially constant and the student interprets the data as a constant function plus noise. Instances of the first type reduce the average of the optimal σ^2-values, a trend which the theory correctly predicts, but have a much stronger effect on the average optimal l through the rare occurrence of large values; our theory based on neglecting fluctuations cannot account for this. For larger n, where theory and simulation agree well, the optimal length scale l increases with n. This makes intuitive sense, since it effectively reduces the excessive roughness in the func-

tions from the student's OU prior to produce a better match to the smoother teacher MB2 covariance function. An example of this effect is shown in Fig. 3. For the RBF student, the opposite trend in the variation of the optimal length scale l is seen: as n increases, l must be reduced to prevent the student from over-smoothing features of the rougher teacher.

4 Conclusion

In summary, the theory presented above shows that evidence maximization goes a long way towards making GP regression robust against model mismatch. The exact results for input spaces of large dimension $d \to \infty$ show that evidence maximization yields the (Bayes-)optimal generalization performance, as long as the true evidence maximum is achievable with the chosen hyperparameters. The optimal hyperparameter values are not, however, meaningful as estimates of the corresponding teacher parameters. The analysis also shows that evidence maximization has its risks, and does not always improve generalization performance: in cases where the ideal evidence maximum cannot be reached by tuning the available hyperparameters, evidence maximization can perform *worse* than learning with any fixed set of hyperparameters.

In the low-dimensional scenarios analysed, the theory predicts correctly that the optimal decay of the generalization error with training set size is obtained even for mismatched models, mainly by appropriate adaptation of the covariance function length scale. Our approximation of optimizing the evidence on average rather than for each specific data set performs worse for small data set sizes here, but predicts simulation results for larger n with surprising quantitative accuracy.

As an issue for further work, it would be interesting to derive the asymptotic decay of the generalization error analytically from (12). One puzzling issue is the increase of the length scale seen for an OU student in Fig. 2. One might argue naively that this increase cannot continue indefinitely because eventually the student covariance function would degenerate into a constant; the length scale should level off for sufficiently large n to prevent this. On the other hand, small deviations from a truly constant covariance function will be amplified by the presence of a large amount of data confirming that the target function is *not* a constant, and this suggests that a true divergence of the optimal length scale with n could occur.

A closer understanding of the effect of increasing d would also be worthwhile. For example, the fact that for $d \to \infty$ continuous ranges of optimal hyperparameter assignments can occur suggests that large fluctuations in the optimal values should be seen if scenarios with large but finite d are considered.

Acknowledgment: 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 author's views.

Bibliography

[1] C K I Williams and C E Rasmussen. Gaussian processes for regression. In D S Touretzky, M C Mozer, and M E Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 514–520, Cambridge, MA, 1996. MIT Press.

[2] C K I Williams. Computing with infinite networks. In M C Mozer, M I Jordan, and T Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 295–301, Cambridge, MA, 1997. MIT Press.

[3] D Barber and C K I Williams. Gaussian processes for Bayesian classification via hybrid Monte Carlo. In M C Mozer, M I Jordan, and T Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 340–346, Cambridge, MA, 1997. MIT Press.

[4] P W Goldberg, C K I Williams, and C M Bishop. Regression with input-dependent noise: A Gaussian process treatment. In M I Jordan, M J Kearns, and S A Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 493–499, Cambridge, MA, 1998. MIT Press.

[5] P Sollich. Learning curves for Gaussian processes. In M S Kearns, S A Solla, and D A Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 344–350, Cambridge, MA, 1999. MIT Press.

[6] Lehel Csató, Ernest Fokoué, Manfred Opper, Bernhard Schottky, and Ole Winther. Efficient approaches to gaussian process classification. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 251–257, Cambridge, MA, 2000. MIT Press.

[7] D Malzahn and M Opper. Learning curves for Gaussian processes regression: A framework for good approximations. In T K Leen, T G Dietterich, and V Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 273–279, Cambridge, MA, 2001. MIT Press.

[8] D Malzahn and M Opper. Learning curves for Gaussian processes models: fluctuations and universality. *Lect. Notes Comp. Sci.*, 2130:271–276, 2001.

[9] D Malzahn and M Opper. A variational approach to learning curves. In T G Dietterich, S Becker, and Z Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 463–469, Cambridge, MA, 2002. MIT Press.

[10] P Sollich and A Halees. Learning curves for Gaussian process regression: approximations and bounds. *Neural Comput.*, 14(6):1393–1428, 2002.

[11] P Sollich. Gaussian process regression with mismatched models. In T G Dietterich, S Becker, and Z Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 519–526, Cambridge, MA, 2002. MIT Press.

[12] C A Michelli and G Wahba. Design problems for optimal surface interpolation. In Z Ziegler, editor, *Approximation theory and applications*, pages 329–348. Academic Press, 1981.

[13] C K I Williams and F Vivarelli. Upper and lower bounds on the learning curve for Gaussian processes. *Mach. Learn.*, 40(1):77–102, 2000.

[14] C K I Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M I Jordan, editor, *Learning and Inference in Graphical Models*, pages 599–621. Kluwer Academic, 1998.

[15] M Seeger. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14(2):69–106, 2004.

[16] M Opper and R Urbanczik. Universal learning curves of Support Vector Machines. *Phys. Rev. Lett.*, 86(19):4410–4413, 2001.

[17] R Dietrich, M Opper, and H Sompolinsky. Statistical mechanics of Support Vector Networks. *Phys. Rev. Lett.*, 82(14):2975–2978, 1999.