Submitted to *Machine Learning* July 1998. Revised April 1999.

Upper and lower bounds on the learning curve for Gaussian processes

Christopher K. I. Williams*
Division of Informatics
The University of Edinburgh
5 Forrest Hill
Edinburgh EH1 2QL, Scotland, UK
email: ckiw@dai.ed.ac.uk

Francesco Vivarelli†
Neural Computing Research Group
Department of Computer Science and Applied Mathematics
Aston University
Birmingham B4 7ET, UK
email: vivarelf@aston.ac.uk

April 24, 1999

Abstract

In this paper we introduce and illustrate non-trivial upper and lower bounds on the learning curves for one-dimensional Gaussian Processes. The analysis is carried out emphasising the effects induced on the bounds by the smoothness of the random process described by the Modified Bessel and the Squared Exponential covariance functions. We present an explanation of the early, linearly-decreasing behaviour of the learning curves and the bounds as well as a study of the asymptotic behaviour of the curves. The effects of the noise level and the lengthscale on the tightness of the bounds are also discussed.

Keywords: Gaussian processes, Bayesian inference, Generalisation error, Bounds

*To whom correspondence should be addressed.
†Current address: Centro Ricerche Ambientali Montecatini, via Ciro Menotti, 48, 48023 Marina di Ravenna, Italy, email: fvivarelli@cramont.it.

1 Introduction

A fundamental problem for systems learning from examples is to estimate the amount of training samples needed to guarantee satisfactory generalisation capabilities on new data. This is of theoretical interest but also of vital practical importance; for example, algorithms which learn from data should not be used in safety-critical systems until a reasonable understanding of their generalisation capabilities has been obtained. In recent years several authors have carried out analysis on this issue and the results presented depend on the theoretical formalisation of the learning problem.

Approaches to the analysis of generalisation include those based on asymptotic expansions around optimal parameter values (e.g. AIC (Akaike, 1974), NIC (Murata et al., 1994)); the Probably Approximately Correct (PAC) framework (Valiant, 1984); uniform convergence approaches (e.g. Vapnik, 1995); and Bayesian methods.

The PAC and uniform convergence methods are concerned with frequentist-style confidence intervals derived from randomness introduced with respect to the distribution of inputs and noise on the target function. A central concern in these results is to identify the flexibility of the hypothesis class $\mathcal{F}$ to which approximating functions belong, for example, through the Vapnik-Chervonenkis dimension of $\mathcal{F}$. Note that these bounds are independent of the input and noise densities, assuming only that the training and test samples are drawn from the same distribution.

The problem of understanding the generalisation capability of systems can also be addressed in a Bayesian framework, where the fundamental assumption concerns the kinds of function our system is required to model. In other words, from a Bayesian perspective we need to put *priors* over target functions. In this context learning curves and their bounds can be analysed by an average over the probability distribution of the functions. In this paper we use Gaussian priors over functions which have the advantage of being more general than simple linear regression priors, but they are more analytically tractable than priors over functions obtained from neural networks.

Neal (1996) has shown that for fixed hyperparameters, a large class of neural network models will converge to Gaussian process priors over functions in the limit of an infinite number of hidden units. The hyperparameters of the Bayesian neural network define the parameters of the corresponding Gaussian Process (GP). Williams (1997) calculated the covariance functions of GPs corresponding to neural networks with certain weight priors and transfer functions.

The investigation of GP predictors is motivated by the results of Rasmussen (1996), who compared the performances obtained by GPs to those obtained by Bayesian neural networks on a range of tasks. He concluded that GPs were at least as good as neural networks. Although the present study deals with regression problems, GPs have also been applied to classification problems (e.g. Barber and Williams, 1997).

In this paper we are mainly concerned with the analysis of upper and lower bounds on the learning curve of GPs. A plot of the expected generalisation error against the number of training samples n is known as a learning curve. There are many results available concerning leaning curves under different theoretical scenarios. However, many of these are concerned with the asymptotic behaviour of these curves, which is not usually of great practical importance as it is unlikely that we will have enough data to reach the asymptotic regime. Our main goal is to explain some of the early behaviour of learning curves for Gaussian processes.

The structure of the paper is as follows. GPs for regression problems are introduced in Section 2. As will be shown, the whole theory of GPs is based on the choice of the prior covariance function $C_p(\mathbf{x}, \mathbf{x}')$: in Section 3 we present the covariance functions we have been using in this study. In Section 4 the learning curve of a GP is introduced. We present some properties of the learning curve of GPs as well as some problems may arise in evaluating it. Upper and lower bounds on the learning curve of a GP in a non-asymptotic regime are presented in Section 5. These bounds have been derived from two different approaches: one makes use of main properties of the generalisation error, whereas the other is derived from an eigenfunction decomposition of the covariance function. The asymptotic behaviour of the upper bounds is also discussed.

A set of experiments have been run in order to assess the upper and lower bounds of the learning curve. In Section 6 we present the results obtained and investigate the link between tightness of the bounds and the smoothness of the stochastic process modelled by a GP. A summary of the results and some open questions are presented in the last Section.

2 Gaussian Processes

A collection of random variables $\{Y(\mathbf{x}) | \mathbf{x} \in \mathcal{X}\}$ indexed by a set $\mathcal{X}$ defines a stochastic process. In general the domain $\mathcal{X}$ might be $\mathbb{R}^d$ for some dimension d although it could be even more general. A joint distribution characterising the statistics of the random variables gives a complete description of the stochastic process.

A GP is a stochastic process whose joint distribution is Gaussian; it is fully defined by giving a Gaussian prior distribution for every finite subset of variables.

In the following we concentrate to the regression problem assuming that the value of the target function $t(\mathbf{x})$ is generated from an underlying function $y(\mathbf{x})$ corrupted by Gaussian noise with mean 0 and variance σ_ν^2. Given a collection of n training data $\mathcal{D}_n = \left\{\left(\mathbf{x}^i, t^i\right), i = 1 \ldots n\right\}$ (where each t^i is the observed output value at the input point $\mathbf{x}^i$), we would like to determine the posterior probability distribution $p(y|\mathbf{x}, \mathcal{D}_n)$.

In order to set up a statistical model of the stochastic process, the set of n random variables $\mathbf{y} = \left(y^1, y^2, \ldots y^n\right)^{\mathrm{T}}$ modelling the function values at $\mathbf{x}^1, \mathbf{x}^2, \ldots \mathbf{x}^n$ respectively, is introduced. Similarly $\mathbf{t}$ is the collection of target values $\mathbf{t} = \left(t^1, \ldots t^n\right)^{\mathrm{T}}$. Let X^n denote the set of training inputs $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$. We also denote with $\tilde{\mathbf{y}}$ the vector whose components are $\mathbf{y}$ and the test value y at the point $\mathbf{x}$. The distribution $p\left(\tilde{\mathbf{y}}|\mathbf{x}, \mathcal{D}_n\right)$ can be inferred using Bayes' theorem. In order to do so, we need to specify a prior over functions as well as evaluate the likelihood of the model and the evidence for the data.

A choice for a prior distribution of the stochastic vector $\tilde{\mathbf{y}}$ is a Gaussian prior distribution:

$$p\left(\tilde{\mathbf{y}}|\mathbf{x}, X^n\right) \propto \exp\left[-\frac{1}{2}\tilde{\mathbf{y}}^{\mathrm{T}}\Sigma^{-1}\tilde{\mathbf{y}}\right].$$

This is a prior as it describes the distribution of the true underlying values without any reference to the target values $\mathbf{t}$. The covariance matrix Σ can be partitioned as

$$\Sigma = \begin{pmatrix} K_p & \mathbf{k}\left(\mathbf{x}\right) \\ \mathbf{k}^{\mathrm{T}}\left(\mathbf{x}\right) & C_p\left(\mathbf{x}, \mathbf{x}\right) \end{pmatrix}.$$

The element $\left(K_p\right)_{ij}$ is the covariance between the i-th and the j-th training points, i.e. $\left(K_p\right)_{ij} = \mathcal{E}\left[\left(y\left(\mathbf{x}^i\right) - \mu\left(\mathbf{x}^i\right)\right)\left(y\left(\mathbf{x}^j\right) - \mu\left(\mathbf{x}^j\right)\right)\right]$. The components of the vector $\mathbf{k}\left(\mathbf{x}\right)$ are the covariances of the test point with all the training data $\left(\mathbf{k}_i\left(\mathbf{x}\right) = C_p\left(\mathbf{x}, \mathbf{x}^i\right)\right)$; $C_p\left(\mathbf{x}, \mathbf{x}\right)$ is the covariance of the test point with itself.

A GP is fully specified by its mean $\mathcal{E}\left[y\left(\mathbf{x}\right)\right] = \mu\left(\mathbf{x}\right)$ and covariance function $C_p\left(\mathbf{x}, \mathbf{x}'\right) = \mathcal{E}\left[\left(y\left(\mathbf{x}\right) - \mu\left(\mathbf{x}\right)\right)\left(y\left(\mathbf{x}'\right) - \mu\left(\mathbf{x}'\right)\right)\right]$. Below we set $\mu\left(\mathbf{x}\right) = 0$; this is a valid assumption provided that any known offset or trend in the data has been removed. We can also deal with $\mu(\mathbf{x}) \neq 0$, but this introduces some extra notational complexity. A discussion about the possible choices of the covariance function $C_p(\mathbf{x}, \mathbf{x}')$ is given in Section 3. For the moment we note that the covariance function is assumed to depend upon the input variables $(\mathbf{x}, \mathbf{x}')$. Thus the correlation between function values depends upon the spatial position of the input vectors; usually this will be chosen so that the closer the input vectors, the higher the correlation of the function values.

The likelihood relates the underlying values of the function to the target data. Assuming a Gaussian noise corrupting the data, we can write the likelihood as

$$p(\mathbf{t}|\mathbf{y}) \propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{t})^{\mathrm{T}}\Omega^{-1}(\mathbf{y}-\mathbf{t})\right]$$

where $\Omega = \sigma_\nu^2 \mathbb{I}$. The likelihood refers to the stochastic variables representing the data; so $\mathbf{t}, \mathbf{y} \in \mathbb{R}^n$ and Ω is an $n \times n$ matrix.

Given the prior distribution over the values of the function $p\left(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{x}^1, \ldots \mathbf{x}^n\right)$, Bayes' rule specifies the distribution $p(\tilde{\mathbf{y}}|\mathbf{x}, \mathcal{D}_n)$ in terms of the likelihood of the model $p(\mathbf{t}|\mathbf{y})$ and the evidence of the data $p(\mathcal{D}_n)$ as

$$p(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{t}, X^n) = \frac{p(\mathbf{t}|\mathbf{y})\, p(\tilde{\mathbf{y}}|\mathbf{x}, X^n)}{p(\mathbf{t}|X^n)}.$$

Given such assumptions, it is a standard result (e.g. Whittle, 1963) to derive the analytic form of the predictive distribution marginalising over $\mathbf{y}$. The predictive distribution turns out to be $y(\mathbf{x}) \sim \mathcal{N}\left(\hat{y}(\mathbf{x}), \sigma^2_{\hat{y},n}(\mathbf{x})\right)$ where the mean and the variance of the Gaussian function are

$$\hat{y}(\mathbf{x}) = \mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathbf{t} \quad (1)$$

$$\sigma^2_{\hat{y},n}(\mathbf{x}) = C_p(\mathbf{x},\mathbf{x}) - \mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathbf{k}(\mathbf{x}) . \quad (2)$$

The most probable value $\hat{y}(\mathbf{x})$ is regarded as the prediction of the GP on the test point $\mathbf{x}$; K is the covariance matrix of the targets $\mathbf{t}$: $K = K_p + \sigma_\nu^2 \mathbb{I}$. The estimate of the variance $\sigma^2_{\hat{y},n}(\mathbf{x})$ of the posterior distribution is considered as the error bar of $\hat{y}(\mathbf{x})$. In the following, we always omit the subscript $\hat{y}$ in $\sigma^2_{\hat{y},n}$, taking it as understood. Since the estimate 1 is a linear combination of the training targets, GPs are regarded as linear smoother (Hastie and Tibshirani, 1990).

3 Covariance functions

The choice of the covariance function is a crucial one. The properties of two GPs, which differ only in the choice of the covariance function, can be remarkably diverse. This is due to the rôle of the covariance function which has to incorporate in the statistical model the prior belief about the underlying function. In other words the covariance function is the analytical expression of the prior knowledge about the function being modelled. A misspecified covariance function affects the model inference as it has influence on the evaluation of Equations 1 and 2.

Formally every function which produces a symmetric, positive semi-definite covariance matrix K for any set of the input space $\mathcal{X}$ can be chosen as covariance function. From an applicative point of view we are interested only in functions which contain information about the structure of the underlying process being modelled.

The choice of the covariance function is linked to the *a priori* knowledge about the smoothness of the function $y(\mathbf{x})$ through the connection between the differentiability of the covariance function and the mean-square differentiability of the process. The relation between smoothness of a process and its covariance function is given by the following theorem (see e.g. Adler, 1981): if $\partial^2 C_p(\mathbf{x}, \mathbf{x}') / \partial x_i \partial x'_i$ exists and is finite at $(\mathbf{x}, \mathbf{x})$, then the stochastic process $y(\mathbf{x})$ is mean square differentiable in the i-th Cartesian direction at $\mathbf{x}$. This theorem is relevant as it links the differentiability properties of the covariance function with the smoothness of the random process and justifies the choice of a covariance function depending upon the prior belief about the degree of smoothness of $y(\mathbf{x})$.

In this work we are mainly concerned with stationary covariance functions. A stationary covariance function is translation invariant (i.e. $C_p(\mathbf{x}, \mathbf{x}') = C_p(\mathbf{x} - \mathbf{x}')$) and depends only upon the distance between two data points. In the following, the covariance functions we have been using are presented. In order to simplify the notation, we consider the case $\mathcal{X} = \mathbb{R}$.

The stationary covariance function *squared exponential* (SE) is defined as

$$C_p(x - x') = \exp\left[-\frac{(x - x')^2}{2\lambda^2}\right] \tag{3}$$

where λ is the lengthscale of the process. The parameter λ defines the characteristic length of the process, estimating the distance in the input space in which the function $y(x)$ is expected to vary significantly. A large value of λ indicates that the function is almost constant over the input space, whereas a small value of the lengthscale designates a function which varies rapidly. The graph of this covariance function is shown by the continuous line in Figure 1. As the SE function has infinitely many derivatives it gives rise to smooth random processes ($y(x)$ posses mean-square differentiability up to order ∞).

It is possible to tune the differentiability of a process, introducing the modified Bessel covariance function of order k (MB_k). It is defined as

$$C_p\left(x-x'\right)=\kappa_\nu\left(\frac{|x-x'|}{\lambda}\right)^\nu \mathcal{K}_\nu\left(\frac{|x-x'|}{\lambda}\right)=\kappa_\nu\sum_{i=0}^{k-1}a_i\left(\frac{|x-x'|}{\lambda}\right)^i\exp\left[-\frac{|x-x'|}{\lambda}\right], \tag{4}$$

where $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of order ν (see e.g. Equation 8.468 in Gradshteyn and Ryzhik, 1993), with $\nu=k-1/2$ for integral k. Below we set the constant κ_ν such that $C_p(0)=1$. The factors a_k are constants depending on the order ν of the Bessel function. Matérn (1980) shows that the functions MB_k define a proper covariance. Stein (1989) also noted that the process with covariance function MB_k is $k-1$ times mean-square differentiable.

In this study we deal with modified Bessel covariance function of orders

$k = 1, 2, 3$; their explicit analytic form is

$$\begin{aligned} k = 1,\ C_p\left(x - x'\right) &= \exp\left[-\frac{|x - x'|}{\lambda}\right] \\ k = 2,\ C_p\left(x - x'\right) &= \exp\left[-\frac{|x - x'|}{\lambda}\right]\left(1 + \frac{|x - x'|}{\lambda}\right) \\ k = 3,\ C_p\left(x - x'\right) &= \exp\left[-\frac{|x - x'|}{\lambda}\right]\left(1 + \frac{|x - x'|}{\lambda} + \frac{1}{3}\left(\frac{|x - x'|}{\lambda}\right)^2\right). \end{aligned}$$

We note that MB_1 corresponds to the Ornstein-Uhlenbeck covariance function which describes a process which is not mean square differentiable.

If $k \to \infty$, the MB_k behaves like the SE covariance function; this can be easily shown by considering the power spectra of MB_k and SE which are

$$S_k\left(\omega\right) \propto \frac{\lambda}{\left(1 + \omega^2\lambda^2\right)^k} \text{ and } S_{se}\left(\omega\right) \propto \lambda \exp\left[-\frac{\omega^2\lambda^2}{2}\right].$$

Since

$$\lim_{k \to \infty}\left(1 + \frac{\omega^2\lambda^2}{2k}\right)^{-k} = \exp\left[-\frac{\omega^2\lambda^2}{2}\right],$$

the MB_k behaves like SE for large k, provided that λ is rescaled accordingly.

Modified Bessel covariance functions are also interesting because they describe Markov processes of order k. Ihara (1991) defines $Y\left(x\right)$ to be a *strict sense* Markov process of order k if it is $k-1$ times mean-square differentiable at every $x \in \mathbb{R}$ and if $P(Y\left(t+s\right) \leq y | Y\left(u\right), u \leq t) = P(Y\left(t+s\right) \leq y | Y\left(t\right), Y'\left(t\right), \ldots Y^{k-1}\left(t\right))$ [1]. Ihara also states that a Gaussian process is a

[1]Note that the definition of a Markov process in discrete and continuous time is rather different. In discrete time, a Markov process of order k depends only on the previous k times, but in continuous time the dependence is on the derivatives at the last time. However, function values at previous times clearly allow approximate computation of

Markov process of order k in the strict sense if and only if it is an autoregressive model of order k (AR(k)) with a power spectrum (in the Fourier domain) of the form

$$S(\omega) \propto \prod_{j=1}^{k} \frac{1}{|i\omega + \alpha_j|^2}.$$

As the power spectrum of MB_k has the same form of the power spectrum of an AR(k) model, the stochastic process whose covariance function is MB_k is a strict sense k-ple Markov process. This characteristic of the MB_k covariance functions is important as it ultimately affects the evaluation of the generalisation error (as we shall see in Section 6).

Figure 2 shows the graphs of four (discretised) random functions generated using the MB_k covariance functions (with $k = 1, 2, 3$) and the SE function. We note how the smoothness of the random function specified is dependent of the choice of the covariance function. In particular, the roughest function is generated by the Ornstein-Uhlenbeck covariance function (Figure 2(a)) whereas the smoothest one is produced by the SE (Figure 2(d)). An intermediate level of regularity characterises the functions of Figures 2(b) and 2(c), corresponding to MB_2 and MB_3 respectively. Note that the number of zero-level upcrossings in $[0, 1]$ (denoted N_u) is only weakly dependent on the order of the process. For MB_2 and MB_3 $E[N_u] = \left(\frac{-C_p''(0)}{\pi^2 C_p(0)}\right)^{1/2}$ is $(\pi\lambda)^{-1}$

derivatives (e.g. via finite differences) and thus one would expect that in the continuous-time situation the previous k process values will contain most of the information needed for prediction at the next time. Note that for the Ornstein-Uhlenbeck process $Y(t+s)$ depends only on the previous observation $Y(t)$.

and $(\sqrt{3}\pi\lambda)^{-1}$ respectively (see Papoulis (1991) eqn 16-7 for details). For the SE process $E[N_u] = (2\pi\lambda)^{-1}$. As the Ornstein-Uhlenbeck process is non-differentiable, the formula given for $E[N_u]$ cannot be applied in this case.

4 Learning curve for Gaussian processes

A learning curve of a model is a function which relates the generalisation error to the amount of training data; it is independent of the test points as well as the locations of the training data and depends only upon the amount of data in the training set. The learning curve for a GP is evaluated from the estimation of the generalisation error averaged over the distribution of the training and test data.

For regression problems, a measure of the generalisation capabilities of a GP is the squared difference $E^g_{\mathcal{D}_n}(\mathbf{x}, t)$ between the target value on a test point $\mathbf{x}$ and the prediction made by using Equation 1:

$$E^g_{\mathcal{D}_n}(\mathbf{x}, t) = \left(t - \mathbf{k}^{\mathrm{T}}(\mathbf{x})\, K^{-1}\mathbf{t}\right)^2 .$$

The Bayesian generalisation error at a point $\mathbf{x}$ is defined as the expectation of $E^g_{\mathcal{D}_n}(\mathbf{x}, t)$ over the actual distribution of the stochastic process t: $E^g_{\mathcal{D}_n}(\mathbf{x}) = \mathcal{E}_t\left[E^g_{\mathcal{D}_n}(\mathbf{x}, t)\right]$. Under the assumption that the data set is actually generated from a GP, it is possible to read Equation 2 as the Bayesian generalisation error at $\mathbf{x}$ given training data $\mathcal{D}_n$. To see this, let us consider the $(n+1)$-dimensional distribution of the target values at $\mathbf{x}^1, \mathbf{x}^2, \ldots \mathbf{x}^n$ and

$\mathbf{x}$. This is a zero-mean multivariate Gaussian. The prediction at the test point $\mathbf{x}$ is $\hat{y}(\mathbf{x}) = \mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathbf{t}$, where $K = K_p + \sigma_\nu^2 \mathbb{I}$. Hence the expected generalisation error at $\mathbf{x}$ is given by

$$\begin{aligned}
E_{\mathcal{D}_n}^{g}(\mathbf{x}) &= \mathcal{E}\left[\left(t - \mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathbf{t}\right)^2\right] \\
&= \mathcal{E}\left[t^2\right] - 2\mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathcal{E}\left[t\mathbf{t}\right] + \mathcal{E}\left[\mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathbf{t}\mathbf{t}^{\mathrm{T}}K^{-1}\mathbf{k}(\mathbf{x})\right] \\
&= C_p(\mathbf{0}) + \sigma_\nu^2 - 2\mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathbf{k}(\mathbf{x}) + \mathrm{Tr}\left[K^{-1}\mathbf{k}(\mathbf{x})\mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathcal{E}\left[\mathbf{t}\mathbf{t}^{\mathrm{T}}\right]\right] \\
&= C_p(\mathbf{0}) + \sigma_\nu^2 - \mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathbf{k}(\mathbf{x}) \qquad (5)
\end{aligned}$$

where we have used $\mathcal{E}[t\mathbf{t}] = \mathbf{k}(\mathbf{x})$ and $\mathcal{E}\left[\mathbf{t}\mathbf{t}^{\mathrm{T}}\right] = K$. Equation 5 is identical to $\sigma_n^2(\mathbf{x})$ as given in Equation 2 with the addition of the noise variance σ_ν^2 (since we are dealing with noisy data). The variance of $(t - \mathbf{k}^{\mathrm{T}}(\mathbf{x}) K^{-1}\mathbf{t})^2$ can also be calculated (Vivarelli, 1998).

The covariance matrix pertinent for these calculations is the true prior; if a GP predictor with a different (incorrect) covariance function is used, the expression for the generalisation error becomes

$$C(0) - 2\mathbf{k}_c^{\mathrm{T}}(\mathbf{x}) K_i^{-1}\mathbf{k}_i(\mathbf{x}) + \mathbf{k}_i^{\mathrm{T}}(\mathbf{x}) K_i^{-1}K_c K_i^{-1}\mathbf{k}_i(\mathbf{x}),$$

where the indices c and i denote the correct and incorrect covariance functions respectively. It can be shown (Vivarelli, 1998) that this is always larger than Equation 5.

Another property of the generalisation error can be derived from the following observation: adding more data points never increases the size of the error bars on prediction ($\sigma_{n+1}^2(\mathbf{x}) \leq \sigma_n^2(\mathbf{x})$). This can be proved using standard results on the conditioning of a multivariate Gaussian (see

Vivarelli, 1998). It can also be understood by the information theoretic argument that conditioning on additional variables never increases the entropy of a random variable. Considering $t(\mathbf{x})$ to be the random variable, we observe that its distribution is Gaussian, with variance independent of $\mathbf{t}$ (although the mean does depend on $\mathbf{t}$). The entropy of a Gaussian is $\frac{1}{2}\log\left(2\pi e\sigma^2(\mathbf{x})\right)$. As log is monotonic, the assertion is proved. This argument is an extension of that in (Qazaz et al., 1997), where the inequality was derived for generalized linear regression.

Since $\sigma_n^2(\mathbf{x}) = E^g_{\mathcal{D}_n}(\mathbf{x})$, a similar inequality applies also to the Bayesian generalisation errors and hence

$$E^g_{\mathcal{D}_{n+1}}(\mathbf{x}) \le E^g_{\mathcal{D}_n}(\mathbf{x}). \tag{6}$$

This remark will be applied in Section 5 for evaluating upper bounds on the learning curve.

Equation 5 calculates the generalisation error at a point $\mathbf{x}$. Averaging $E^g_{\mathcal{D}_n}(\mathbf{x})$ over the density distribution of the test points $p(\mathbf{x})$, the expected generalisation error $E^g_{\mathcal{D}_n}$ is

$$E^g_{\mathcal{D}_n} = \int \left(C_p(\mathbf{0}) + \sigma_\nu^2 - \mathbf{k}^{\mathrm{T}}(\mathbf{x})\, K^{-1}\mathbf{k}(\mathbf{x})\right) p(\mathbf{x})\, d\mathbf{x}. \tag{7}$$

For particular choices of $p(\mathbf{x})$ and $C_p(\mathbf{x})$ the computation of this expression can be reduced to a $n \times n$ matrix computation as $\mathcal{E}_{\mathbf{x}}\left[\mathbf{k}^{\mathrm{T}}(\mathbf{x})\, K^{-1}\mathbf{k}(\mathbf{x})\right] = \mathrm{Tr}\left[K^{-1}\mathcal{E}_{\mathbf{x}}\left[\mathbf{k}(\mathbf{x})\,\mathbf{k}^{\mathrm{T}}(\mathbf{x})\right]\right]$. We also note that Equation 7 is independent of the test point $\mathbf{x}$ but still depends upon the choice of the training data $\mathcal{D}_n$. In order to obtain a proper learning curve for GP, $E^g_{\mathcal{D}_n}$ needs to be averaged[2] over the possible choices of the training data $\mathcal{D}_n$. However, it is very difficult to obtain the analytical form of E^g for a GP as a function of n. Because of the presence of the $\mathbf{k}^{\mathrm{T}}(\mathbf{x})\, K^{-1}\mathbf{k}(\mathbf{x})$ term in Equation 5, the matrix K and vector $\mathbf{k}(\mathbf{x})$ depend on the location of the training points: the calculations of the averages with respect to the data points seems very hard. This motivates looking for upper and lower *bounds* on the learning curve for GP.

5 Bounds on the learning curve

For the noiseless case, a lower bound on the generalisation error after n observations is due to Michelli and Wahba (1981). Let $\eta_1, \eta_2, \ldots$ be the ordered eigenvalues of the covariance function on some domain of the input space $\mathcal{X}$. They showed that $E^g(n) \geq \sum_{k=n+1}^{\infty} \eta_k$. Plaskota (1996) gives a bound on the learning curve for the noisy case; since the bound uses observations consisting of projections of the random function onto the first N eigenfunctions, it is not expected that it will be tight for observations which consist of function evaluations.

Other results that we are aware of pertain to asymptotic properties of $E^g(n)$. Ritter (1996) has shown that for an optimal sampling of the input space, the asymptotics of the generalisation error is $O\left(n^{-(2s+1)/(2s+2)}\right)$ for

[2] Hansen (1993) showed that for linear regression models it is possible to average over the distribution of the training sets.

a random process which obeys to the Sacks-Ylvisaker[3] conditions of order s (see Ritter et al., 1995 for more details on Sacks-Ylvisaker conditions). In general, the Sacks-Ylvisaker order of the MB_k covariance function is $s = k - 1$. For example an MB_1 process has $s = 0$ and hence the generalisation error shows a $n^{-1/2}$ asymptotic decay. In the case that $\mathcal{X} \subset \mathbb{R}$, the asymptotically optimal design of the input space is the uniform grid.

Silverman (1985) proved a similar result for random designs. Haussler and Opper (1997) have developed general (asymptotic) bounds for the expected log-likelihood of a test point after seeing n training points.

In the following we introduce upper and lower bounds on the learning curve of a GP in a non-asymptotic regime. An upper bound is particularly useful in practice as it provides an (over)estimate of the number of examples needed to give a certain level of performance. A lower bound is similarly important because it contributes to fix the limit which can not be outperformed by the model.

The bounds presented are derived from two different approaches. The first approach makes use of the particular form assumed by the generalisation error at $\mathbf{x}$ ($E^g_{\mathcal{D}_n}(\mathbf{x}) = \sigma^2_n(\mathbf{x})$). As the error bar generated by one data point is greater than that generated by n data points, the former can be considered as an upper bound of the latter. Since this observation holds for the variance due to each one the data points, the envelope of the surfaces generated by

[3] Loosely speaking, a stochastic process possessing s mean-square derivatives but not $s + 1$ is said to satisfy the Sacks-Ylvisaker conditions of order s.

the variances due to each data point is also an upper bound of $\sigma_n^2(\mathbf{x})$. In particular as $\sigma_n^2(\mathbf{x}) = E_{\mathcal{D}_n}^g(\mathbf{x})$ (cf. Equation 5), the envelope is an upper bound of the generalisation error of the GP. Following this argument, we can assert that an upper bound on $E_{\mathcal{D}_n}^g(\mathbf{x})$ is the one generated by every GP trained with a subset of $\mathcal{D}_n$. The larger the subset of $\mathcal{D}_n$ the tighter the bound.

The two upper bounds we present differ in the number of training points considered in the evaluation of the covariance: the derivation of the one-point upper bound $E_1^u(n)$ and the two-point upper bound $E_2^u(n)$ are presented in Section 5.1 and Section 5.2 respectively. Section 5.3 reports the asymptotic expansion of $E_1^u(n)$ in terms of λ and σ_ν^2.

The second approach is based on the expansion of the stochastic process in terms of the eigenfunctions of the covariance function. Within this framework, Opper proposed bounds on the training and generalisation error (Opper and Vivarelli, 1999) in terms of the eigenvalues of $C_p(\mathbf{x}, \mathbf{x}')$; the lower bound $E^l(n)$ obtained is presented in Section 5.4.

In order to have tractable analytical expressions, all the bounds have been derived by introducing three assumptions:

i The input space $\mathcal{X}$ is restricted to the interval $[0, 1]$;

ii The probability density distribution of the input points is uniform: $p(x) = 1, x \in [0, 1]$;

iii The prior covariance function $C_p(x, x')$ is stationary.

5.1 The one-point upper bound $E_1^u(n)$

For the derivation of the one-point upper bound, let us consider the error bar generated by one data point x^i. Since $C(0) = C_p(x^i, x^i) + \sigma_\nu^2 = K$, Equation 2 becomes

$$\sigma_1^2(x) = C(0) - \frac{C_p^2(x - x^i)}{C(0)}.$$

For x far away from the training point x^i, $\sigma_1^2(x) \sim C(0)$: the confidence on the prediction for a test point lying far apart from the data point x^i is quite low as the error bar is large. The closer x to x^i, the smaller the error bar on $\hat{y}(x)$. When $x = x^i$, $\sigma_1^2(x) = \sigma_\nu^2(1 + r)$ where $r = C_p(0)/C(0)$. Irrespective of the value of $C_p(0)$, r varies from 0 to 1. As normally $C_p(0) \gg \sigma_\nu^2$, $r \sim 1$ and thus $\sigma_1^2(x) \sim 2\sigma_\nu^2$. So far we have not used any hypothesis concerning the dimension of the variable x, thus this observation holds regardless the dimension of the input space.

The effect of just one data point helps in introducing the first upper bound. The interval $[0, 1]$ is split up in n subintervals $\left[a^i, b^i\right], i = 1 \ldots n$ (where $a^i = \left(x^i + x^{i-1}\right)/2$ and $b^i = \left(x^{i+1} + x^i\right)/2$) centred around the i-th data point x^i, with $a^1 = 0$ and $b^n = 1$.

Let us consider the i-th training point and the error bar $\sigma_1^2(x)$ generated by x^i. When $x \in \left[a^i, b^i\right]$, $E_{\mathcal{D}_n}^g(x) \leq \sigma_1^2(x)$; this relation is illustrated in Figure 3, where the envelope of the surfaces of the errors due to each datapoint (denoted by $E_{\mathcal{D}_1}^g(\mathbf{x})$) is an upper bound of the overall generalisation error. Since we are dealing with positive functions, an upper bound of the expected generalisation error on the interval $\left[a^i, b^i\right]$ can be written as

$$\int_{a^i}^{b^i} E^g_{\mathcal{D}_n}(x)\, p(x)\, dx \leq \int_{a^i}^{b^i} \sigma_1^2(x)\, p(x)\, dx \tag{8}$$

where $p(x)$ is the distribution of the test points. Summing up the contributions coming from each training datapoint in both sides of Equation 8 and setting $p(x) = 1$, we obtain

$$E^g_{\mathcal{D}_n} = \sum_{i=1}^{n} \int_{a^i}^{b^i} E^g_{\mathcal{D}_n}(x)\, dx \leq \sum_{i=1}^{n} \int_{a^i}^{b^i} \sigma_1^2(x)\, dx \tag{9}$$

The interval where the contribution of the variance due to x^i contributes to Equation 8 is also shown in Figure 3.

Under the assumption of the stationarity of the covariance function, integrals such as those in the right hand side of Equation 9 depend only upon differences of adjacent training points (i.e. $x^i - x^{i-1}$ and $x^{i+1} - x^i$). The right hand side of Equation 9 can be rewritten as

$$\begin{aligned} E^g_{\mathcal{D}_n} &\leq \sum_{i=1}^{n} \int_{a^i}^{b^i} \sigma_1^2(x)\, dx = C(0) \sum_{i=1}^{n} \left(b^i - a^i\right) \qquad (10) \\ &\qquad - \frac{1}{C(0)} \sum_{i=1}^{n} \left[\int_{a^i}^{x^i} C_p^2\left(x^i - x\right) dx + \int_{x^i}^{b^i} C_p^2\left(x - x^i\right) dx \right] \\ &= C(0) - \frac{1}{C(0)} \left[I\left(x^1\right) + 2 \sum_{i=2}^{n} I\left(\frac{x^i - x^{i-1}}{2}\right) + I\left(1 - x^n\right) \right] \qquad (11) \end{aligned}$$

where

$$I(\tau) = \int_0^{\tau} C_p^2(\xi) d\xi. \tag{12}$$

Equation 11 can be derived changing the variables in the two integrals of Equation 10 as $\xi = x^i - x$ and $\xi = x - x^i$, respectively. Equation 11 is an upper bound on $E^g_{1,\mathcal{D}_n}$ and still depends upon the choice of the training data $\mathcal{D}_n$ through the interval of integration. We note that the arguments of the integrals $I(\cdot)$ in Equation 11 are the differences between adjacent training points. Denoting those differences with $\omega^i = x^{i+1} - x^i$, we can model their probability density distribution by using the theory of order statistics (David, 1970). Given an uniform distribution of n training data over the interval $[0,1]$, the density distribution of the differences between adjacent points is $p(\omega) = n(1-\omega)^{n-1}$. Since this is true for all the differences ω^i we can omit the superscript i and thus the expectation of the integrals in Equation 11 over $p(\omega)$ is

$$\mathcal{E}_\omega \left[I\left(\omega^0\right) + 2\sum_{i=2}^{n} I\left(\frac{\omega^i}{2}\right) + I\left(\omega^n\right) \right] = 2(n-1)\mathcal{E}_\omega\left[I\left(\omega/2\right)\right] + 2\mathcal{E}_\omega\left[I\left(\omega\right)\right], \tag{13}$$

where $\omega^0 = x^1$ and $\omega^n = 1 - x^n$. Both the integrals $\mathcal{E}_\omega\left[I\left(\omega/2\right)\right]$ and $\mathcal{E}_\omega\left[I\left(\omega\right)\right]$ can be calculated following a similar procedure. Let us consider $\mathcal{E}_\omega\left[I\left(\omega\right)\right]$:

$$\begin{aligned}
\mathcal{E}_\omega\left[I\left(\omega\right)\right] &= \int_0^1 I\left(\omega\right) n\left(1-\omega\right)^{n-1} d\omega \\
&= -\left[I\left(\omega\right)\left(1-\omega\right)^n\right]_0^1 + \int_0^1 C_p^2\left(\omega\right)\left(1-\omega\right)^n d\omega \\
&= \int_0^1 C_p^2\left(\omega\right)\left(1-\omega\right)^n d\omega,
\end{aligned}$$

where the second line has been obtained integrating by parts. The last line follows from the fact that $\left[I\left(\omega\right)\left(1-\omega\right)^n\right]_0^1 = 0$.

We are now able to write an upper bound on the learning curve as

$$E^g(n) \leq E_1^u(n) \doteq C(0) - \frac{1}{C(0)}\left[(n-1)\int_0^1 C_p^2\left(\frac{\omega}{2}\right)(1-\omega)^n\,d\omega + 2\int_0^1 C_p^2(\omega)(1-\omega)^n\,d\omega\right]. \tag{14}$$

The calculations of the integrals in the above expression are straightforward though they involve the evaluation of hyper-geometric functions (because of the term $(1-\omega)^n$). As the evaluation of such functions is computationally intensive, we found preferable to evaluate Equation 14 numerically.

5.2 The two-points upper bound $E_2^u(n)$

The second bound we introduce is the natural extension of the previous idea, using two data points rather than one. By construction, we expect that it will be tighter than the one introduced in Section 5.1.

Let us consider two adjacent data points x^i and x^{i+1} of the interval $[0,1]$, with $x^i < x^{i+1}$. By the same argument presented in the previous section, the following inequality holds:

$$\int_{x^i}^{x^{i+1}} E_{\mathcal{D}_n}^g(x)\,p(x)\,dx \leq \int_{x^i}^{x^{i+1}} \sigma_2^2(x)\,p(x)\,dx \tag{15}$$

where $\sigma_2^2(x)$ is the variance on the prediction $\hat{y}(x)$ generated by the data points x^i and x^{i+1}. Similarly to Equation 9, summing up the contributions of both sides of Equation 15 we get an upper bound on the generalisation error:

$$E_{\mathcal{D}_n}^g = \sum_{i=0}^{n}\int_{x^i}^{x^{i+1}} E_{\mathcal{D}_n}^g(x)\,dx \leq \sum_{i=0}^{n}\int_{x^i}^{x^{i+1}} \sigma_2^2(x)\,dx, \tag{16}$$

where we have defined $x^0 = 0$ and $x^{n+1} = 1$.

After some calculations (see Appendix A) we obtain

$$E^g(n) \leq E_2^u(n) \doteq C(0) - 2(n-1)\,\mathcal{E}_\omega\left[\frac{(I_1(\omega) - I_2(\omega))}{\Delta(\omega)}\right] - \frac{2}{C(0)}\mathcal{E}_\omega\left[I(\omega)\right], \tag{17}$$

where

$$I_1(\tau) = C(0)\int_0^\tau C_p^2(\xi)\,d\xi, \qquad I_2(\tau) = C_p(\tau)\int_0^\tau C_p(\xi)\,C_p(\tau-\xi)\,d\xi$$

and $\Delta(\omega) = C^2(0) - C_p^2(\omega)$. The calculation of the integrals with respect to ω in $E_2^u(n)$ is complicated by the determinant $\Delta(\omega)$ in the denominator and by the distribution $n(1-\omega)^{n-1}$, so we preferred to evaluate them numerically as we did for $E_1^u(n)$.

5.3 Asymptotics of the upper bounds

From Equation 14, an expansion of $E_1^u(n)$ in terms of λ and σ_ν^2 in the limit of a large amount of training data can be obtained. The expansion depends upon the covariance function we are dealing with. Expanding the covariance function around 0, the asymptotic form of $E_1^u(n)$ for MB_1 is

$$E_1^u(n) \sim C(0)\left[1 - r^2 + \frac{r^2}{n\lambda}\right] + O\left(n^{-2}\right) \tag{18}$$

whereas for the functions MB_2, MB_3 and SE it is

$$E_1^u(n) \sim C(0)\left[1 - r^2 + \frac{r^2}{n^2\lambda^2}\right] + O\left(n^{-3}\right) \tag{19}$$

where $r = C_p(0)/C(0)$ (Vivarelli, 1998).

The asymptotic value of $E_1^u(n)$ depends neither on the lengthscale of the process nor on the order of the covariance function MB_k for $k \geq 1$ but is a function of the ratio r:

$$\lim_{n\to\infty} E_1^u(n) = C(0)\left(1 - r^2\right) = \sigma_\nu^2 (1 + r). \tag{20}$$

As we pointed out in Section 5.1, this is the minimum generalisation error achievable by a GP when it is trained with just one datapoint. The $n \to \infty$ scenario corresponds to the situation in which every test point is close to a datapoints. As mentioned at the beginning of this Section, the asymptotics of the learning curve for the MB_k and SE covariance functions are $O\left(n^{(2k-1)/2k}\right)$ and $O\left(n^{-1}\log n\right)$ respectively. Although the expansions of $E_1^u(n)$ decay asymptotically faster than the learning curves, they reach an asymptotic plateau $\sigma_\nu^2(1+r) \geq \sigma_\nu^2$. We also note that the asymptotic values $\overline{E}_1^g(n)$ get closer to the true noise level when $r \ll 1$, i.e. for the unrealistic case $\sigma_\nu^2 \gg C_p(0)$.

The smoothness of the process enters into the asymptotics through a factor $O\left(r^2/(\lambda n)\right)$ for MB_1 and $O\left(r^2/\left(\lambda^2 n^2\right)\right)$ for MB_2, MB_3 and SE. This factor affects the rate of approach to the asymptotic value $\sigma_\nu^2(1+r)$ of $E_1^u(n)$. We notice that larger lengthscales and noise levels increase the rate of decay of $E_1^u(n)$ to the asymptotic plateau.

The asymptotic form of $E_2^u(n)$ for the MB_1, MB_2, MB_3 and SE covariance functions is (Vivarelli, 1998)

$$E_2^u(n) \sim C(0)\left(1 - \frac{2r^2}{1+r}\right) + \frac{a}{n+1} + O\left(n^{-2}\right), \tag{21}$$

where the value of a depends upon the choice of the covariance function and $r = C_p(0)/C(0)$. Similarly to the expansion of $E_1^u(n)$, the decay rate of $E_2^u(n)$ is faster than the asymptotic decay of the actual learning curves but it reaches an asymptotic plateau of

$$\lim_{n\to\infty} E_2^u(n) = C(0)\left(1 - \frac{2r^2}{1+r}\right) = \sigma_\nu^2\left(1 + \frac{r}{1+r}\right). \tag{22}$$

It is straightforward to verify that the asymptotic plateau of $E_2^u(n)$ is lower than the one of $E_1^u(n)$ and that it corresponds to the error bar estimated by a GP with two observations located at the test point.

5.4 The lower bound $E^l(n)$

Opper (Opper and Vivarelli, 1999) proposed a bound on the learning curve and on the training error based on the decomposition of the stochastic process $y(\mathbf{x})$ in terms of the eigenfunctions of the covariance $C_p(\mathbf{x}, \mathbf{x}')$.

Denoting with $\varphi_k(\mathbf{x}), k = 1 \ldots \infty$ a complete set of functions satisfying the integral equation

$$\int C_p(\mathbf{x}, \mathbf{x}')\,\varphi_k(\mathbf{x})\,p(\mathbf{x})\,d\mathbf{x} = \eta_k \varphi_k(\mathbf{x}),$$

the Bayesian generalisation error $E^g(\mathbf{x}, \mathcal{D}_n) = \mathcal{E}_y\left[(y(\mathbf{x}) - \hat{y}(\mathbf{x}))^2\right]$ (where $y(\mathbf{x})$ is the true underlying stochastic function and $\hat{y}(\mathbf{x})$ is the GP prediction) can be written in terms of the eigenvalues of $C_p(\mathbf{x}, \mathbf{x}')$. In particular, after an average over the distribution of the input data, $E^g(\mathcal{D}_n)$ can be written as $E^g(\mathcal{D}_n) = \sigma_\nu^2 \mathrm{Tr}\left[\Lambda\left(\sigma_\nu^2 \mathbb{I} + \Lambda V\right)^{-1}\right]$, where Λ is the infinite dimension diagonal matrix of the eigenvalues and V is a matrix depending on the training data, i.e. $V_{kl} = \sum_{i=1}^{n} \varphi_k\left(\mathbf{x}^i\right) \varphi_l\left(\mathbf{x}^i\right)$.

By using Jensen's inequality, it is possible to show that a lower bound of the learning curve and an upper bound of the training error is (Opper and Vivarelli, 1999)

$$E_y^l\left(n\right) \doteq \sigma_\nu^2 \sum_{k=1}^{\infty} \frac{\eta_k}{\left(\sigma_\nu^2 + n\eta_k\right)}. \tag{23}$$

In this paper we mean to compare this lower bound to the actual learning curve of a GP. As our bounds are on t rather than y, we must add σ_ν^2 to the expression obtained in Equation 23 giving an actual lower bound of

$$E^l\left(n\right) \doteq \sigma_\nu^2 \left(1 + \sum_{k=1}^{\infty} \frac{\eta_k}{\left(\sigma_\nu^2 + n\eta_k\right)}\right). \tag{24}$$

6 Results

As we pointed out in Section 4, the analytic calculation of the learning curve of a GP is infeasible. Since the generalisation error

$$E_{\mathcal{D}_n}^g = \int \left(C_p\left(\mathbf{0}\right) + \sigma_\nu^2 - \mathbf{k}^{\mathrm{T}}\left(\mathbf{x}\right) K^{-1}\mathbf{k}\left(\mathbf{x}\right)\right) p\left(\mathbf{x}\right) d\mathbf{x} \tag{25}$$

is a complicated function of the training data (which are inside the elements of $\mathbf{k}\left(\mathbf{x}\right)$ and K^{-1}), it is problematic to perform an integration over the distribution of the training points. For comparing the learning curve of the GP with the bounds we found, we need to evaluate the expectation of the integral in Equation 25 over the distribution of the data: $E^g\left(n\right) = \mathcal{E}_{\mathcal{D}_n}\left[E_{\mathcal{D}_n}^g\right]$. An estimate of $E^g\left(n\right)$ can be obtained using a Monte Carlo approximation of the expectation. We used 50 generations of training data, sampling uniformly the input space $[0, 1]$. For each generation, the expected generalisation error for a GP has been evaluated using up to 1000 datapoints. Using the 50 generations of training data, we can obtain an estimate of the learning curve $E^g(n)$ and its 95% confidence interval.

Since this study is focused on the behaviour of bounds on learning curve on GP, we assume the true values of the parameters of the GP are known. So we chose the value of the constant κ_ν for the covariance functions MB_1, MB_2 and MB_3 (see Equation 4) such that $C_p(0) = 1$ and we allowed the lengthscale λ and the noise level σ_ν^2 to assume several values ($\lambda = 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1$ and $\sigma_\nu^2 = 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1$).

To begin with, we study how the smoothness of a process affects the behaviour of the learning curve. The empirical learning curves of Figure 4 have been obtained for processes whose covariance functions are MB_1, MB_2 and MB_3, with $\lambda = 0.01$ and $\sigma_\nu^2 = 0.1$. We can notice that all the learning curves exhibit an initial linear decrease. This can be explained considering that without any training data, the generalisation error is the maximum allowable by the model ($C(0) = C_p(0) + \sigma_\nu^2$). The introduction of a training point x^1 creates a hole on the error surface: the volume of the hole is proportional to the value of the lengthscale and depends on the covariance function. The addition of a new data point x^2 will have the effect of generating a new hole in the surface. With such a few data points it is likely that the two data lie down far apart one from the other, giving rise to two distinct holes. Thus the effect that a small dataset exerts to *pull* *down* the error surface is proportional to the amount of training points and explains the initial linear trend.

Concerning the asymptotic behaviour of the learning curves, we have verified that they agree with the theoretical analysis carried out by Ritter (1996). In particular, a log-log plot of the learning curves with a MB_k covariance function shows an asymptotic behaviour as $O\left(n^{-(2k-1)/2k}\right)$. A similar remark applies to the SE covariance function, with an asymptotic decay rate of $O\left(n^{-1}\log n\right)$ (Opper, 1997). We have also noted that the smoother the process described by the covariance function the smaller the the amount of training data needed to reach the asymptotic regime.

The behaviour of the learning curves is affected also by the value of the lengthscale of the process and by the noise level and this is illustrated in Figure 7. The learning curves shown in Figure 5(a) have been obtained for the MB_1 covariance function setting the noise level $\sigma_\nu^2 = 0.1$ and varying the values of the parameters $\lambda = 10^{-2}, 10^{-1}$. Intuitively, Figure 5(a) suggests that decreasing the lengthscale stretches the early behaviour of the learning curve and the approach to the asymptotic plateau lasts longer; this is due to the effect induced by different values of the lengthscale which stretch or compress the input space. We have verified that rescaling the amount of data n by the ratio of the two lengthscales, the two curves of Figure 5(a) lay on top of each other.

The variation of the noise level shifts the learning curves from the prior value $C_p(0)$ by an offset equal to the noise level itself (cf. Equation 5);

in order to see any significant effect of the noise on the learning curve, Figure 5(b) shows a log-log graph of $E^g(n) - \sigma_\nu^2$ obtained for a stochastic process with MB_3 covariance function, setting $\lambda = 0.1$ and noise variance $\sigma_\nu^2 = 10^{-4}, 10^{-1}$. We can notice two main effects. The noise variance affects the actual values of the generalisation error since the learning curve obtained with high noise level is always above the one obtained with a low noise level. A second effect concerns the amount of data necessary to reach the asymptotic regime. The learning curve characterised by an high noise level needs fewer datapoints to attain to the asymptotic regime.

Stochastic processes with different covariance functions and different values of lengthscales and noise variance behave in a similar way.

In the following we discuss the results in two main subsections: results about the bounds $E_1^u(n)$ and $E_2^u(n)$ are presented in Section 6.1, whereas the lower bound of Section 5.4 is shown in Section 6.2. As the results we obtained for these experiments show common characteristics, we show the bounds of the learning curve obtained by setting $\lambda = 0.01$ and $\sigma_\nu^2 = 0.1$.

6.1 The upper bounds $E_1^u(n)$ and $E_2^u(n)$

Each graph in Figure 6 shows the empirical learning curve with its confidence interval and the two upper bounds $E_1^u(n)$ and $E_2^u(n)$. The curves are shown for the MB_1, MB_2, MB_3 and the SE covariance functions.

For a limited amount of training data it is possible to notice that the upper error bar associated to $\mathcal{E}_{\mathcal{D}_n}[E^g(n)]$ lies above the actual upper bounds.

This effect is due to the variability of the generalisation error for small data sets and suggests that the bounds are quite tight for small n. The effect disappears for large n, when the estimate of the generalisation error is less sensitive to the composition of the training set.

As expected, the two-point upper bound $E_2^u(n)$ is tighter than the one-point upper bound $E_1^u(n)$.

We note that the tightness of the upper bound depends upon the covariance function, being tighter for rougher processes (such as MB_1) and getting worse for smoother processes. This can be explained by recalling that covariance functions such as the MB_k correspond to Markov processes of order k (cf. Section 3). Although the Markov process is actually hidden by the presence of the noise, $E^g(n)$ is still more dependent on training data lying close to the test point x than on more distant points. Since the bounds $E_1^u(n)$ and $E_2^u(n)$ have been calculated by using only local information (namely the closest datapoint to the test point, or the closest datapoints to the left and right, respectively), it is natural that the more the variance at x depends on local data points, the tighter the bounds become.

For instance, let us consider MB_1, the covariance function of a first order Markov process. For the noise-free process, knowledge of data-points lying beyond the the left and right neighbours of x does not reduce the generalisation error at x[4]. Although in the noisy case more distant data-points

[4] This is because the process values at the training points and test point form a Markov chain, and knowledge of the process values to the left and right of the test point "blocks"

reduce the generalisation error (because of the term σ_ν^2 in the covariance matrix K), it is likely that local information is still the most important.

The bounds on the learning curves computed for MB_2 and MB_3 confirm this remark, as they are looser than for MB_1. For the SE covariance function, this effect still holds and is actually enlarged.

In Section 5.3 we have shown that the asymptotic behaviour of the bound $E_1^u(n)$ depends on the covariance function, being $O\left(n^{-1}\right)$ for MB_1 and $O\left(n^{-2}\right)$ for MB_2 and MB_3. Log-log plots of the upper bounds confirm the analysis carried out in Section 5.3, where we showed that $E_1^u(n)$ and $E_2^u(n)$ approach asymptotic plateaux. In particular, $E_1^u(n)$ tends to $\sigma_\nu^2(1+r)$ as $O\left(n^{-1}\right)$ for MB_1 and $O\left(n^{-2}\right)$ for MB_2 and MB_3, whereas $E_2^u(n)$ tends to $\sigma_\nu^2(1+r/(1+r))$ as $O\left(n^{-1}\right)$.

The quality of the bounds for processes characterised by different length-scales and different noise levels are comparable to the ones described so far: the tightness of $E_1^u(n)$ and $E_2^u(n)$ still depend on the smoothness of the process. As explained at the beginning of this section, a variation of the lengthscale has the same effect of a rescaling in the number of training data. This can be observed explicitly in the asymptotic analysis of Equations 18 and 19, where the decay rate depends on the factor $n\lambda$.

For a fixed covariance function, we note that the bounds are tighter for lower noise variance; this is due to the fact that the lower the noise level the better the hidden Markov process manifests itself. For smaller noise levels

the influence of more remote observations.

the learning curve becomes closer to the bounds because the generalisation error relies on the local behaviour of the processes around the test data; on the contrary, a larger noise level hides the underlying Markov Process thus loosening the bounds.

6.2 The bound $E^l(n)$

We have also run experiments computing the lower bound we obtained from Equation 24 for processes generated by the covariance priors MB_1, MB_2, MB_3 and SE .

Equation 24 shows that the evaluation of $E^l(n)$ involves the computation of an infinite sum of terms; we truncated the series considering only those terms which add a significant contribution to the sums, i.e. $\eta_k/\sigma_\nu^2 \geq \varepsilon$, where ε is the machine precision. Since each contribution in the series is positive, the quantity computed is still a lower bound of the learning curve.

Figure 7 shows the results of the experiment in which we set $\lambda = 0.01$ and $\sigma_\nu^2 = 0.1$. The graphs of the lower bound lies below the empirical learning curve, being tighter for large amount of data; in particular for the smoothest processes with large amount of data, the 95% confidence intervals lay below the actual lower bound.

For $n \to \infty$, the lower bound tends to the noise level σ_ν^2. As with the empirical learning curve, log-log plots of $E_y^l(n)$ show an asymptotic decay to zero as $O(n^{-(2k-1)/2k})$ and $O\left(n^{-1}\log n\right)$ for the MB_k and the SE covariance functions, respectively.

The graphs of Figure 7 show also that the tightness of the bound depends on the smoothness of the stochastic process; in particular smooth processes are characterised by a tight lower bound on the learning curve $E^g(n)$. This can be explained by observing that $E^l(n)$ is a lower bound on the learning curve and an upper bound of the training error. The values of smooth functions do not have large variation between training points and thus the model can infer better on test data; this reduces the generalisation error pulling it closer to the training error. Since the two errors sandwich the bound of Equation 24, $E^l(n)$ becomes tight for smooth processes.

We can also notice that the tightness of the lower bound depends on the noise level, becoming tight for high the noise level and loose for small noise level. This is consistent with a general characteristic of $E^l(n)$ which is monotonically decreasing function of the noise variance (Opper and Vivarelli, 1999).

7 Discussion

In this paper we have presented non-asymptotic upper and lower bounds for the learning curve of GPs. The theoretical analysis has been carried out for one-dimensional GPs characterised by several covariance functions and has been supported by numerical simulations.

Starting from the observation that increasing the amount of training data never worsens the Bayesian generalisation error, an upper bound on the learning curve can be estimated as the generalisation error of a GP trained with a reduced dataset. This means that for a given training set the envelope of the generalisation errors generated by one and two datapoints is an upper bound of the actual learning curve of the GP. Since the expectation of the generalisation error over the distribution of the training data is not analytically tractable, we introduced the two upper bounds $E_1^u(n)$ and $E_2^u(n)$ which are amenable to average over the distribution of the test and training points. In this study we have evaluated the expected value of the bounds; future directions of research should also deal with the evaluation of the variances.

In order to highlight the behaviour of the bounds with respect to the smoothness of the stochastic process, we investigated the bounds for the modified Bessel covariance function of order k (describing stochastic processes $k-1$ times mean-square differentiable) and the squared exponential function (describing processes mean square-differentiable up to the order ∞).

The experimental results have shown that the learning curves and their bounds are characterised by an early, linearly decreasing behaviour; this is due to the effect exerted by each datapoint in pulling down the surface of the prior generalisation error. We also noticed that the tightness of the bounds depends on the smoothness of the stochastic processes. This is due to the facts that the bounds rely on subsets of the training data (i.e. one or two datapoints) and the modified Bessel covariance functions describe

Markov processes of order k; although in our simulations the Markovian processes were hidden by noise, the learning curves depend mainly on local information and our bounds become tighter for rougher processes.

We also investigated the behaviour of the curves with respect to the variation of the correlation lengthscale of the process and the variance of the noise corrupting the stochastic process. We noticed that the lengthscale stretches the behaviour of the curves effectively rescaling the number of training data. As the noise level has the effect of hiding the underlying Markov process, the upper bounds become tighter for smaller noise variance.

The expansion of the bounds in the limit of large amount of data highlights an asymptotic behaviour depending upon the covariance function; $E_1^u(n)$ approaches the asymptotic plateau as $O\left(n^{-1}\right)$ (for the MB_1 covariance function) and as $O\left(n^{-2}\right)$ for smoother processes; the rate of decay to the plateau of $E_2^u(n)$ is $O\left(n^{-1}\right)$. Numerical simulations supported our analysis.

One limitation of our analysis is the dimension of the input space; the bounds have been made analytically tractable by using order statistics results after splitting up the one dimensional input space of the GP. In higher dimensional spaces the partition of the input space can be replaced by a Voronoi tessellation that depends on the data $\mathcal{D}_n$ but averaging over this distribution appears to be difficult. One can suggest an approximate evaluation of the upper bounds by an integration over a ball whose radius depends upon the number of examples and the volume of the input space in which the bound holds. In any case we expect that the effect due to larger input dimension is to loosen the upper bounds. We note that recent work by (Sollich, 1999) has derived some good approximations to the learning curve, and that his methods apply in more than one dimension[5].

We also ran some experiments by using the lower bound proposed by Opper, based on the knowledge of the eigenvalues of the covariance function of the process. Since the bound $E^l(n)$ is also an upper bound on the training error, we observed that the bound is tighter for smooth processes, when the learning curve becomes closer to the training error. Also the noise can vary the tightness of $E^l(n)$; a low noise level loosens the lower bound. Unlike the upper bounds, the lower bound can be applied also in multivariate problems, as it is easily extended to high dimension input space; however it has been verified (Opper and Vivarelli, 1999) that the bound becomes less tight in input space of higher dimension.

Appendix A: The two-points upper bound $E_2^u(n)$

In this Appendix we derive Equation 17 starting from Equation 16.

We start by calculating $\sigma_2^2(x)$. As the covariance matrix generated by two data points is a 2×2 matrix, it is straightforward to evaluate $\sigma_2^2(x)$. Considering the two training data x^i and x^{i+1}, the covariance matrix of the

[5]The reference to Sollich (1999) was added when the manuscript was revised in April 1999.

GP is

$$K = \begin{pmatrix} C(0) & C_p\left(x^{i+1}-x^i\right) \\ C_p\left(x^{i+1}-x^i\right) & C(0) \end{pmatrix}.$$

From the evaluation of the determinant of K as $\Delta\left(x^{i+1}-x^i\right) = (C(0))^2 - \left(C_p\left(x^{i+1}-x^i\right)\right)^2$ follows that

$$K^{-1} = \frac{1}{\Delta\left(x^{i+1}-x^i\right)} \begin{pmatrix} C(0) & -C_p\left(x^{i+1}-x^i\right) \\ -C_p\left(x^{i+1}-x^i\right) & C(0) \end{pmatrix}.$$

As the covariance vector for the test point x is $\mathbf{k}(\mathbf{x}) = \left(C_p\left(x-x^i\right), C_p\left(x^{i+1}-x\right)\right)^{\mathrm{T}}$, the variance assumes the form

$$\sigma_2^2(x) = C(0) - \frac{C(0)\left(C_p^2\left(x^{i+1}-x\right)+C_p^2\left(x-x^i\right)\right) - 2C_p\left(x^{i+1}-x^i\right)C_p\left(x-x^i\right)C_p\left(x^{i+1}-x\right)}{\Delta\left(x^{i+1}-x^i\right)}.$$

Changing variables in the covariances $C_p\left(x^{i+1}-x^i\right)$ and $C_p\left(x-x^i\right)$ (as $\xi = x^{i+1}-x$ and $\xi = x-x^i$ respectively), it turns out that the upper bound generated by $\sigma_2^2(x)$ in the interval $\left[x^i, x^{i+1}\right]$ (when $i \neq 0, n$), is

$$\int_{x^i}^{x^{i+1}} \sigma_2^2(x)\,dx = C(0)\left(x^{i+1}-x^i\right) - \frac{2\left(I_1\left(x^{i+1}-x^i\right) - I_2\left(x^{i+1}-x^i\right)\right)}{\Delta\left(x^{i+1}-x^i\right)}$$

where

$$I_1(\tau) = C(0)\int_0^\tau C_p^2(\xi)\,d\xi \text{ and } I_2(\tau) = C_p(\tau)\int_0^\tau C_p(\xi)\,C_p(\tau-\xi)\,d\xi.$$

It is noticeable that, similarly to Equation 11, also the integrals $I_1(\cdot)$, $I_2(\cdot)$ and the determinant $\Delta\left(x^{i+1}-x^i\right)$ depend upon the length of the interval of integration $\omega^i = x^{i+1}-x^i$. We evaluate the contributions to the upper bound over the intervals $\left[0, x^1\right]$ and $[x^n, 1]$ by integrating the variance $\sigma_1^2(x)$

generated by x^1 and x^n over $\left[0, x^1\right]$ and $[x^n, 1]$ respectively. Hence the right hand side of Equation 16 can be rewritten as

$$E^g_{\mathcal{D}_n} \leq C(0) - 2\sum_{i=2}^{n-1} \frac{I_1\left(\omega^i\right) - I_2\left(\omega^i\right)}{\Delta\left(\omega^i\right)} - \frac{1}{C(0)}\left(I\left(\omega^1\right) + I\left(\omega^n\right)\right) \tag{26}$$

where $I(\cdot)$ is defined in Equation 12.

Equation 26 is still dependent on the distribution of the training data because it is a function of the distances between adjacent training points ω^i. Similarly to Equation 11, we obtain an upper bound independent of the training data by integrating Equation 13 over the distribution of the differences $p(\omega) = n(1-\omega)^{n-1}$:

$$E^g(n) \leq E^u_2(n) \doteq C(0) - 2(n-1)\,\mathcal{E}_\omega\left[\frac{(I_1(\omega) - I_2(\omega))}{\Delta(\omega)}\right] - \frac{2}{C(0)}\mathcal{E}_\omega\left[I(\omega)\right]. \tag{27}$$

Acknowledgments

This research forms part of the "Validation and Verification of Neural Network Systems" project funded jointly by EPSRC (GR/K 51792) and British Aerospace. We thank Dr. Manfred Opper, and Dr. Andy Wright of BAe for helpful discussions. We also thank the anonymous referees for their comments which have helped improve this paper. F. V. was supported by a studentship from British Aerospace.

References

Adler, R. J. (1981). *The Geometry of Random Fields.* John Wiley and Sons, New York.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Barber, D. and Williams, C. K. I. (1997). Gaussian processes for Bayesian classification via hybrid Monte Carlo. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems 9.* MIT Press.

David, H. A. (1970). *Order Statistics.* John Wiley and Sons, New York.

Gradshteyn, E. S. and Ryzhik, I. M. (1993). *Table of Integrals, Series and Products.* Academic Press, New York, fifth edition.

Hansen, L. K. (1993). Stochastic linear learning: Exact test and training error averages. *Neural Networks*, 6:393–396.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Haussler, D. and Opper, M. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451.

Ihara, S. (1993). *Information Theory.* World Scientific Publishing, Singapore.

Matérn, B. (1986). *Spatial Variation.* Springer-Verlag, Berlin, second edition. Lecture Notes in Statistics 36.

Michelli, C. A. and Wahba, G. (1981). Design problems for optimal surface interpolation. In Ziegler, Z., editor, *Approximation Theory and Applications*, pages 329–348. Academic Press.

Murata, N., Yoshizawa, S., and Amari, S. (1994). Network information criterion–determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks*, 5:865–872.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks.* Springer. Lecture Notes in Statistics 118.

Opper, M. (1997). Regression with gaussian processes: average case performance. In Kwok-Yee, M. W., Irwin, K., and Dit-Yan, Y., editors, *Theoretical Aspects of Neural Computation: A Multidisciplinary Perspective.* Springer-Verlag.

Opper, M. and Vivarelli, F. (1999). General bounds on Bayes errors for regression with Gaussian Processes. To appear in: Kearns, M. J., Solla, S. A. and Cohn, D. A., editors, *Advances in Neural Information Processing Systems 11.* MIT Press.

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes* McGraw-Hill, New York. Third Edition.

Plaskota, L. (1996). *Noisy information and computational complexity.* Cambridge University Press, Cambridge.

Qazaz, C. S., Williams, C. K. I., and Bishop, C. M. (1997) An Upper Bound on the Bayesian Error Bars for Generalized Linear Regression. In Ellacott, S. W., Mason, J. C. and Anderson, I. J., editors, *Mathematics of Neural Networks*, 295-299. Kluwer.

Rasmussen, C. E. (1996). *Evaluation of Gaussian Processes and Other Methods for Non-linear Regression.* PhD thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

Ritter, K. (1996). Almost optimal differentiation using noisy data. *Journal of Approximation Theory*, 86(3):293–309.

Ritter, K., Wasilkowski, G. W., and Wozniakowski, H. (1995). Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. *Ann. Appl. Prob.*, 5:518–540.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve filtering. *Journal of the Royal Statistical Society B*, 47(1):1–52.

Sollich, P. (1999) Learning Curves for Gaussian Processes. To appear in: Kearns, M. J., Solla, S. A. and Cohn, D. A., editors, *Advances in Neural Information Processing Systems 11.* MIT Press.

Stein, M. L. (1989). Comment on the paper by Sacks, J. *et al.* Design and Analysis of Computer Experiments. *Statistical Science*, 4(4):432-433.

Valiant, L. G. (1984). A theory of the learnable. *Communication of the Association for Computing Machinery*, 27:1134–1142.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.

Vivarelli, F. (1998). *Studies on generalisation in Gaussian processes and Bayesian neural networks.* PhD thesis, Neural Computing Research Group, Aston University, Birmingham, United Kingdom.

Whittle, P. (1963). *Prediction and regulation by linear least square methods.* English Universities Press.

Williams, C. K. I. (1997). Computing with infinite networks. In Mozer, M. C.and Jordan, M. I. and Petsche, T., editors, *Advances in Neural Information Processing Systems 9.* MIT Press.

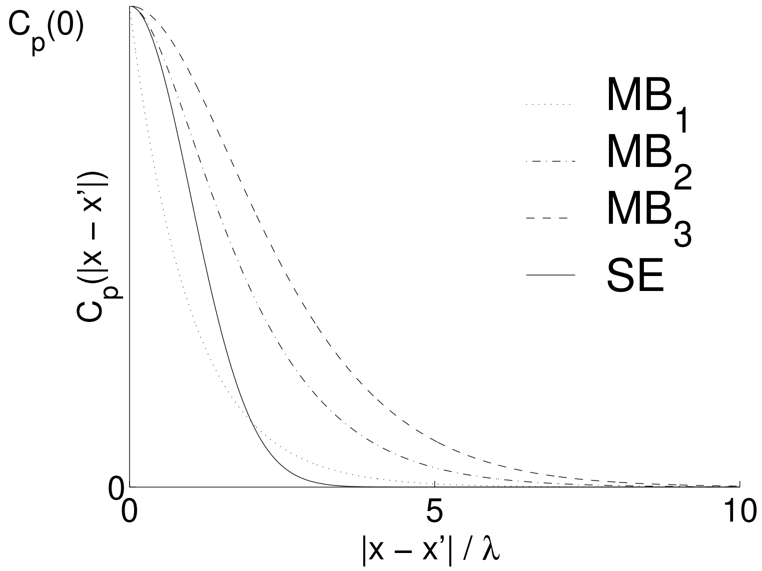


Figure 1: The Figure shows the covariance functions used in this work. The solid line is the SE covariance function; the dotted, dash-dot and dashed lines draw the graph of the MB_k covariance functions with $r = 1, 2$ and 3 respectively. The values of $|x - x'| / \lambda$ are reported on the x-axis.

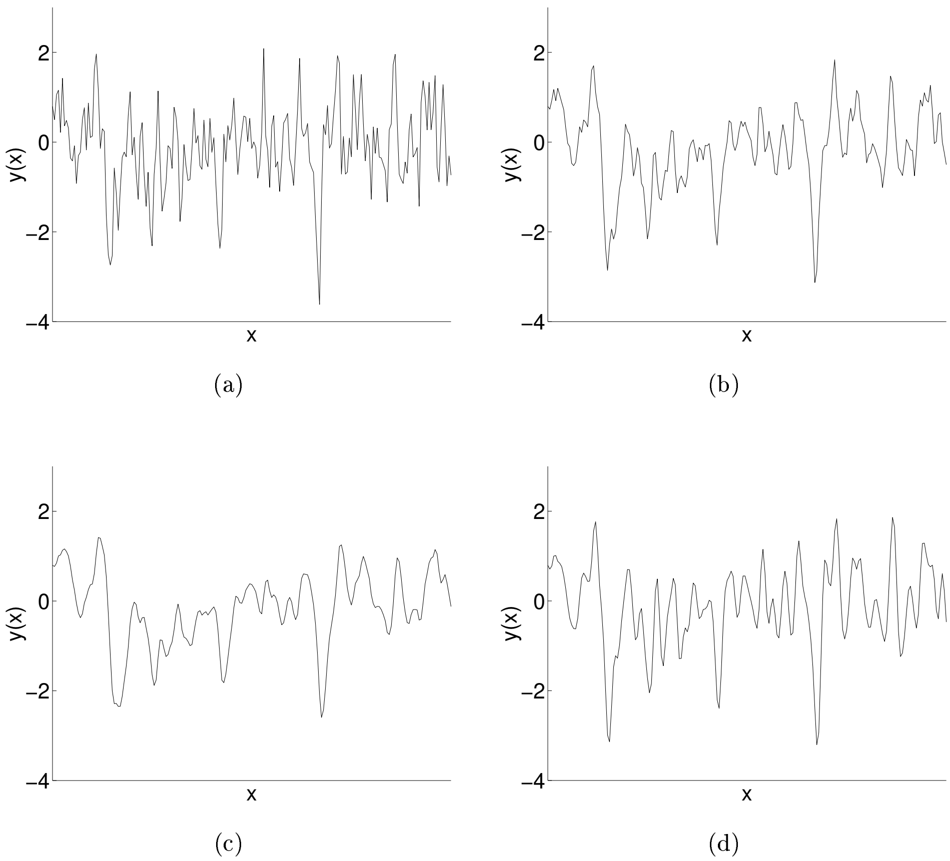


Figure 2: Discretised sample of random functions generated from the MB covariance function of first (2(a)), second (2(b)), third order (2(c)), and the SE function (2(d)) with $\lambda = 0.01$. The order $r-1$ of a process refers to the number of mean square derivatives of the random process. In all cases the x-axis is the interval $[0, 1]$.

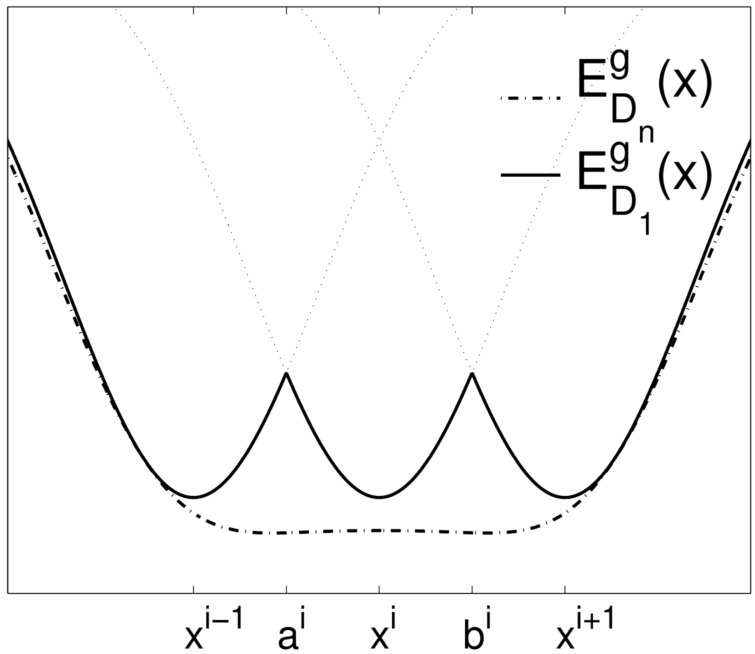


Figure 3: The figure suggests a pictorial argument for the upper bound $E_1^u(n)$. The solid and the dash-dotted lines indicate the bound and the actual generalisation error, respectively. The dotted lines are the generalisation errors evaluated considering training sets composed by each training point singularly, i.e. $\mathcal{D}_1 = \{x^{i-1}\}$, $\mathcal{D}_1 = \{x^i\}$ and $\mathcal{D}_1 = \{x^{i+1}\}$. As explained in the text, $E^g_{\mathcal{D}_n}(x) \leq E^g_{\mathcal{D}_1}(x)$ for all the input points of the input space and thus the latter is regarded as an upper bound of the former. $[a^i, b^i]$ specifies the interval of integration of Equation 8.

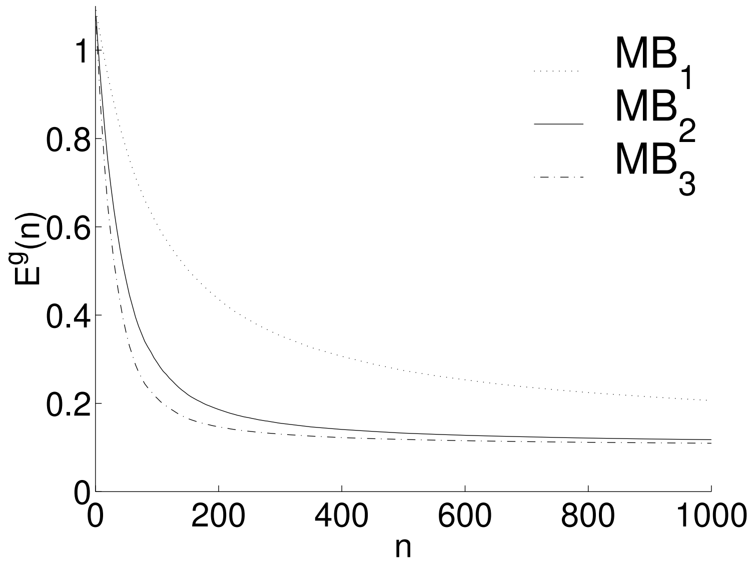
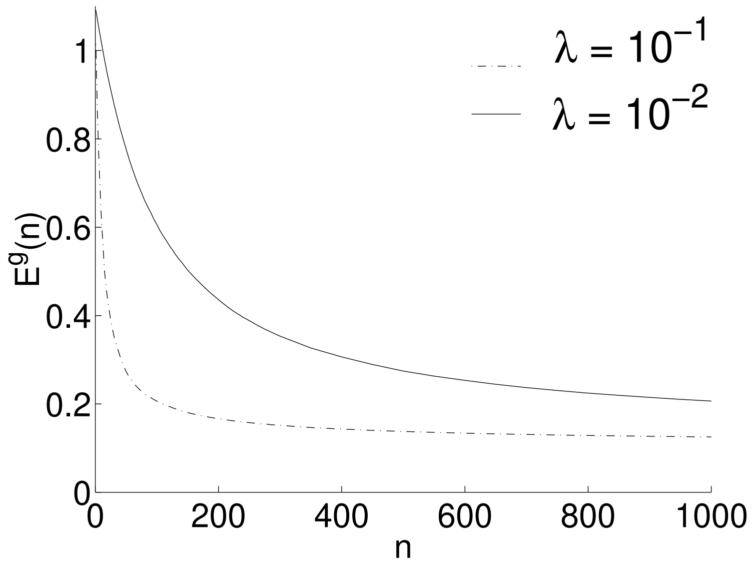
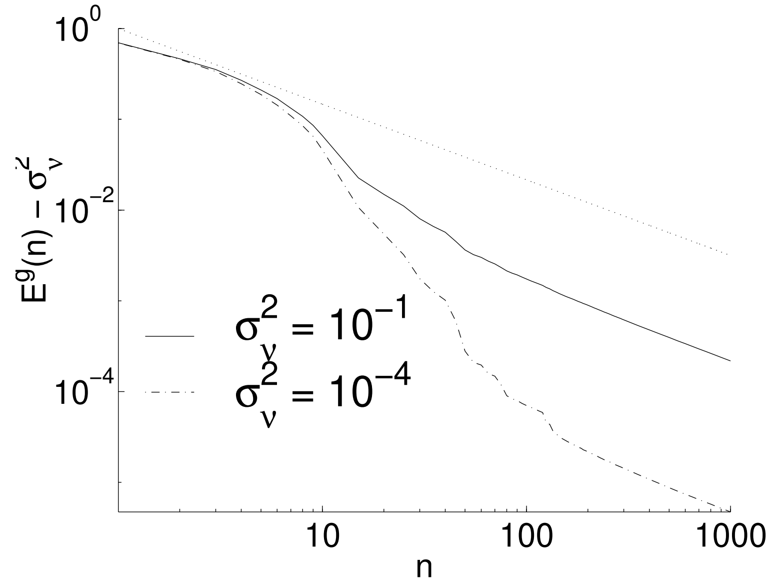


Figure 4: The Figure show the graph of the learning curve computed for the covariance functions MB_1, MB_2 and MB_3 indicated by the dotted, solid and dash-dotted lines, respectively.



(a)



(b)

Figure 5: Figure 5(a) shows the graphs of the learning curves for the MB_1 covariance function obtained for a fixed noise level $\sigma_\nu^2 = 0.1$ and lengthscales $\lambda = 10^{-2}, 10^{-1}$; the lengthscale contributes to stretch the input domain and a similar effect is observed on the learning curves. A log-log plot of the learning curve of a MB_3 stochastic process is shown in Figure 5(b), with $\lambda = 10^{-1}$ and the noise variance is set to 10^{-4} (solid line) and 10^{-1} (dash-dotted line); the dotted line draws the asymptotic behaviour of the learning curve. The curve with a larger noise level attains the asymptotic regime with fewer datapoints than with a lower noise variance.

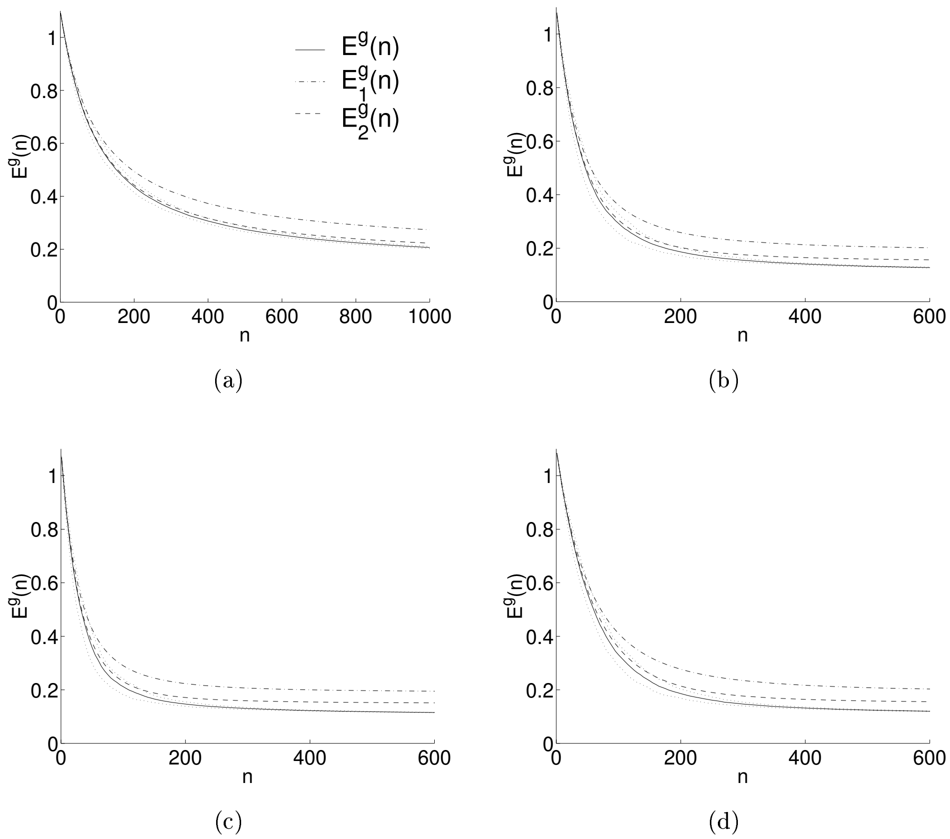


Figure 6: Figures 6(a), 6(b), 6(c) and 6(d) show the graphs of the learning curves and their upper bounds computed for the covariance functions MB_1, MB_2, MB_3 and the SE respectively. In all the graphs, the learning curve is drawn by the solid line and its 95% confidence interval is indicated by the dotted curves. The upper bounds $E_1^u(n)$ and $E_2^u(n)$ are indicated by the dash dotted and the dashed lines, respectively.

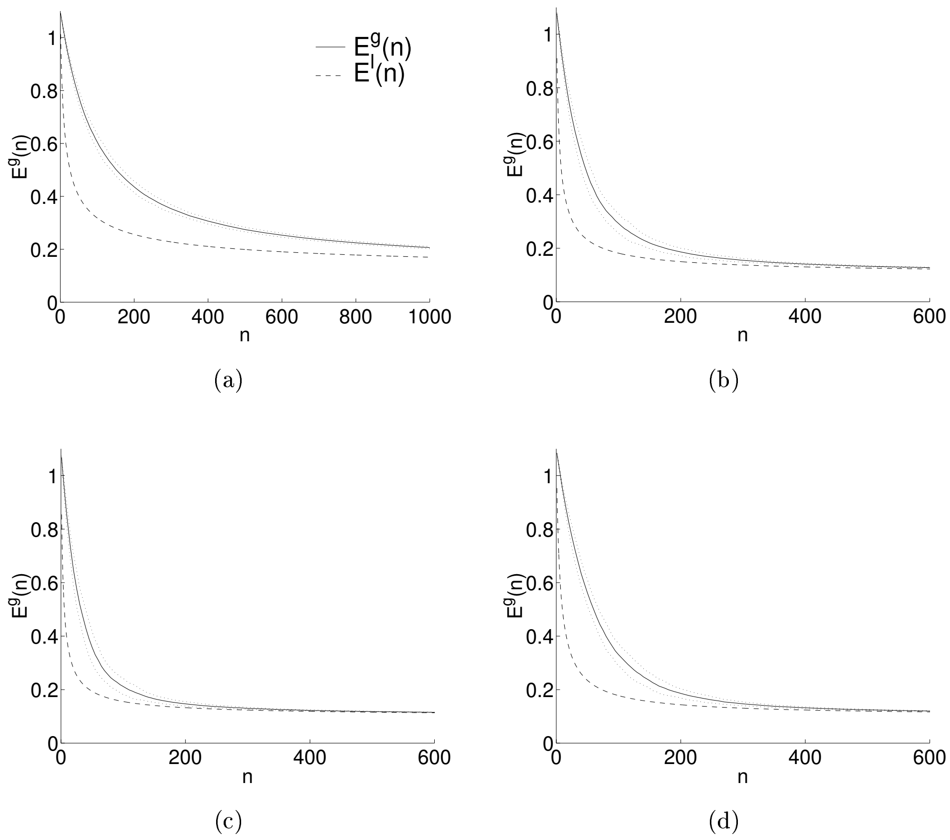


Figure 7: Figures 7(a), 7(b), 7(c) and 7(d) show the graphs of the learning curves and their lower bounds computed for the covariance functions MB_1, MB_2, MB_3 and the SE respectively. In all the graphs, the learning curve is drawn by the solid line and its 95% confidence interval is signed by the dotted curves. The lower bound $E^l(n)$ is indicated by the dashed lines.