Learning curves for Gaussian processes models: Fluctuations and universality

Dörthe Malzahn and Manfred Opper

Department of Computer Science and Applied Mathematics
Aston University, Birmingham B4 7ET, United Kingdom
{malzahnd,opperm}@aston.ac.uk

Abstract. Based on a statistical mechanics approach, we develop a method for approximately computing average case learning curves and their sample fluctuations for Gaussian process regression models. We give examples for the Wiener process and show that universal relations (that are independent of the input distribution) between error measures can be derived.

1 Introduction

Gaussian process (GP) models have gained considerable interest in the Neural Computation Community (see e.g.[1–4]) in recent years. However, being non-parametric models by construction their theoretical understanding is less well developed compared to simpler parametric models like neural networks. In this paper we present new results for approximate computation of learning curves by further developing our framework from [5] which was based on a statistical mechanics approach. In contrast to most previous applications of statistical mechanics to learning theory the method is *not* restricted to the so called "thermodynamic" limit which would require a high dimensional input space.

Our approach has the advantage that it is rather general and may be applied to different likelihoods and allows for a systematic computation of corrections.

In this contribution we will rederive our approximation in an new way based on a general variational method. We will show that we can compute other interesting quantities like the sample fluctuations of the generalization error. Nevertheless, one may criticise this and similar approaches of statistical physics as being not relevant for practical situations, because the analysis requires the knowledge of the input distribution which is usually not available. However, we will show (so far for a toy example) that our approximation predicts universal relations (that are *independent* of the input distribution) between different error measures. We expect that similar relations may be obtained for more practical situations.

2 Regression with Gaussian processes

Regression with Gaussian processes is based on a statistical model [2] where observations $y(x) \in R$ at input points $x \in R^D$ are assumed to be corrupted values of an unknown function $f(x)$. For independent Gaussian noise with variance σ^2, the likelihood for a set of m example data $D = (y(x_1), \ldots, y(x_m)))$ (conditioned on the function f) is given by

$$P(D|f) = \frac{\exp\left(-\sum_{i=1}^{m} \frac{(y_i - f(x_i))^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{\frac{m}{2}}} \tag{1}$$

where $y_i \doteq y(x_i)$. To estimate the function $f(x)$, one supplies the *a priori* information that f is a realization of a Gaussian process (random field) with zero mean and covariance $C(x, x') = E[f(x)f(x')]$, where E denotes the expectation over the Gaussian process prior. Predictions $\hat{f}(x)$ for the unknown function f are computed as the posterior expectation of $f(x)$, i.e. by

$$\hat{f}(x|D) = E\{f(x)|D\} = \frac{Ef(x)P(D|f)}{Z_m} \tag{2}$$

where the partition function Z_m normalises the posterior.

In the sequel, we call the true data generating function f^* in order to distinguish it from the functions over which we integrate in the expectations. We will compute approximations for the learning curve, i.e. the generalization (mean square) error averaged over independent draws of example data, i.e. $\varepsilon_g = [(f^*(x) - \hat{f}(x|D))^2]_{(x,D)}$ as a function of m, the sample size. We will use brackets $[\ldots]$ to denote averages over data sets where we assume that the inputs x_i are drawn *independently* at random from a density $p(x)$. The index at the bracket denotes the quantities that are averaged over. For example, $[\ldots]_{(x,D)}$ denotes both an average over example data D and a *test* input drawn from the same density. We will also approximate the sample fluctuations of the generalization error defined by $\Delta\varepsilon_g = \sqrt{[[(f^*(x) - \hat{f}(x|D))^2]_x^2]_D - \varepsilon_g^2}$.

3 The Partition Function

As typical of statistical mechanics approaches, we base our analysis on the averaged "free energy" $[-\ln Z_m]_D$ where the partition function Z_m (see Eq. (2)) is $Z_m = EP(D|f)$. $[\ln Z_m]_D$ serves as a generating function for suitable posterior averages. The computation of $[\ln Z_m]_D$ is based on the replica trick $[\ln Z_m]_D = \lim_{n\to 0} \frac{\partial \ln [Z_m^n]_D}{\partial n}$, where we compute $[Z^n]_D$ for integer n and perform the continuation at the end. We have

$$Z_n(m) \doteq [Z_m^n]_D = E_n \left[\frac{\exp\left(-\sum_{a=1}^{n} \frac{(f_a(x) - y)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}^{n}}\right]_x^m , \tag{3}$$

where E_n denotes the expectation over the GP measure for the n-times replicated GPs $f_a(x)$, $a = 1, \ldots, n$.

For further analytical treatment, it is convenient to introduce the "grand canonical" free energy

$$\Xi_n(\mu) = \sum_{m=0}^{\infty} \frac{e^{\mu m}}{m!} Z_n(m) = E_n \exp[-H_n] \tag{4}$$

where the energy H_n is a functional of $\{f_a\}$

$$H_n = -e^{\mu} \left[\frac{\exp\left(-\sum_{a=1}^{n} \frac{(f_a(x)-y)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}^{\,n}} \right]_x . \tag{5}$$

This represents a "poissonized" version of our model where the number of examples is fluctuating. For sufficiently large m, the relative fluctuations are small and both models will give the same answer, provided the "chemical potential" μ and the desired m are related by $m = \frac{\partial \ln \Xi_n(\mu)}{\partial \mu}$. Using a Laplace argument for the sum in Eq. (4), we have $\ln Z_n(m) \approx \ln \Xi_n(\mu) + m(\ln m - 1) - m\mu$. Note that as a result of the data average, the model defined by H_n is no longer Gaussian and we cannot compute $\ln \Xi_n(\mu)$ exactly. We will therefore resort to a variational approximation.

4 Variational approximation

Our goal is to approximate H_n by a simpler quadratic Hamiltonian of the form $H_n^0 = \frac{1}{2}\sum_{a,b=1}^{n} \eta_{ab}[(f_a(x) - y)(f_b(x) - y)]_{(x,y)}$, where η_{ab} are parameters to be optimised. Assuming η_{ab} to be fixed for the moment, we can expand the free energy in a power series in the deviations $H - H_n^0$

$$-\ln \Xi_n(\mu) = -\ln E_n \exp[-H_n^0] + \langle H - H_n^0\rangle_0 - \frac{1}{2}\left(\langle (H-H_n^0)^2\rangle_0 - \langle H - H_n^0\rangle_0^2\right) \pm \ldots . \tag{6}$$

The brackets $\langle \ldots \rangle_0$ denote averages with respect to the effective Gaussian measure induced by the replicated prior and $e^{-H_n^0}$. As is well known [6], the first two terms in Eq. (6) are an upper bound to $-\ln \Xi_n(\mu)$. We will optimise H_n^0, by choosing the matrix η_{ab} such that this upper bound is minimised. Thereafter, a replica symmetric continuation to real n is achieved by restricting the variations to the form $\eta_{ab} = \eta$ for $a \neq b$ and $\eta_{aa} = \eta_0$. Note however, that after this continuation we can no longer establish a bound on $-\ln \Xi_n(\mu)$. To compute the generalization error and other quantities we will use the effective Gaussian measure induced by H_n^0. The variational equations on η_0 and η can be expressed as functionals of the local generalization error

$$\varepsilon_g(x) = \lim_{n\to 0} \langle (f_1(x) - f^*(x))(f_2(x) - f^*(x))\rangle_0 \tag{7}$$

and the local posterior variance

$$v_p(x) = \lim_{n\to 0} \langle (f_1(x) - f^*(x))^2\rangle_0 - \varepsilon_g(x). \tag{8}$$

By neglecting variations of these quantities with x we arrive at the following set of equations

$$[\hat{C}(x,x)]_x + \sigma^2 = \frac{m}{(\eta_0 - \eta)} \tag{9}$$

$$[\tilde{E}^2(f(x) - y)]_{(x,y)} - \eta[\hat{C}^2(x,x')]_{(x,x')} = -\frac{m\eta}{(\eta_0 - \eta)^2} \tag{10}$$

that determine the values of the variational parameters η_0 and η. Eqs. (9,10) require the mean $\tilde{E}(f(x) - y)$ and the covariance $\hat{C}(x,x') = \hat{E}(f(x)f(x'))$ with respect to the Gaussian measures $\tilde{E} \propto E\exp(-\frac{(\eta_0-\eta)}{2}[(f(x)-y)^2]_{x,y})$ and $\hat{E} \propto E\exp(-\frac{(\eta_0-\eta)}{2}[f^2(x)]_x)$.

5 Results for learning curves and fluctuations

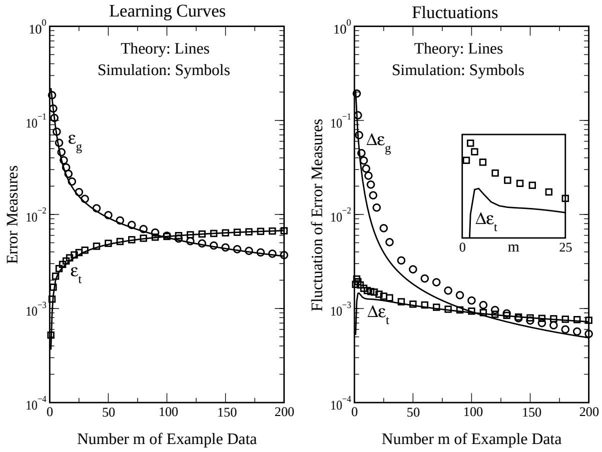

Fig. 1. Learning curves (*left*) and their Fluctuations (*right*) for the periodic Wiener Process. Our theory is represented by *lines* whereas *symbols* give simulation results

We compare our analytical results for the generalization error ε_g, the training error $\varepsilon_t = [\sum_{i=1}^m (\hat{f}(x_i|D) - y_i(x))^2]_D/m$ and for their sample fluctuations $\Delta\varepsilon_g$, $\Delta\varepsilon_t$ with simulations of GP regression. For simplicity, we have chosen the Wiener process $C(x,x') = min(x,x')$ as a toy model. For Fig. 1, the target function f^* is a fixed but random realisation from the prior distribution and the data noise is Gaussian with variance $\sigma^2 = 0.01$. The left panel of Fig. 1 shows learning curves while their fluctuations are displayed in the right panel. Symbols represent simulation results and our theory is given by lines. The training error ε_t converges to the noise level σ^2. As one can see from the pictures our theory is very accurate when m is sufficiently large. It also predicts the initial increase of $\Delta\varepsilon_t$ for small values of m (see *inset* of Fig. 1, right panel).

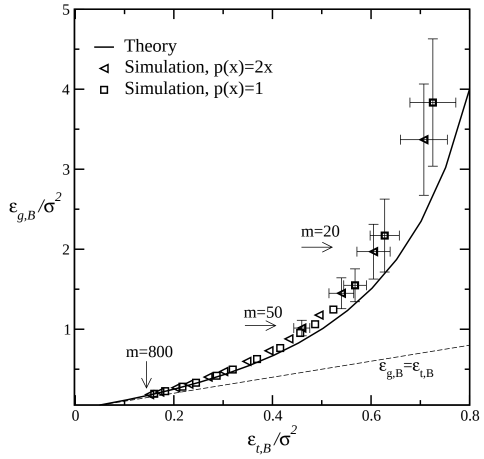

Fig. 2. Bayes errors for the periodic Wiener Process, $\sigma^2 = 0.01$. Theory (*bold line*) versus Simulations (*Symbols*). Simulations have been performed for two input distributions $x \in [0,1]$ and $p(x) = 1$ (*squares*) or $p(x) = 2x$ (*triangles*). The number m of example data is indicated by *arrows*. As m increases, the generalization error $\varepsilon_{g,B}$ and the error bars decrease. The *dashed line* illustrates the trivial limes $\varepsilon_{g,B} \approx \varepsilon_{t,B}$ for $m \to \infty$

6 Universal relations

Although the explicit computations of our results requires the knowledge of the data distribution, we can establish universal relations (valid in the framework of our approximation) which are independent of this density. We restrict ourselves to the full Bayesian scenario where all quantities are averaged over the prior distribution of true functions f^*. The uncertainty of the prediction at a point x is measured by the posterior variance $\varepsilon_B(x) = E(\hat{f}(x|D) - f(x))^2$. Bayesian generalization errors defined as $\varepsilon_{g,B} = [\varepsilon_B(x)]_{(x,D)}$ for this scenario were computed previously by Peter Sollich [7] under the assumption of a uniform input distribution. Our results for this special case turn out to be *identical* to Sollich's result.

However, extending our framework to arbitrary input densities, we find that the Bayesian generalization error and its empirical estimate $\varepsilon_{t,B} = \frac{1}{m}\sum_{i=1}^{m}[\varepsilon_B(x_i)]_D$ are expressed by a single variational parameter of our model only. This can be eliminated to give the following surprisingly simple relation

$$\bar{\varepsilon}_{g,B} \approx \frac{\bar{\varepsilon}_{t,B}}{1 - \bar{\varepsilon}_{t,B}} \tag{11}$$

where $\bar{\varepsilon}_{t,B} = \varepsilon_{t,B}/\sigma^2$ and $\bar{\varepsilon}_{g,B} = \varepsilon_{g,B}/\sigma^2$. Fig. 2 displays simulation results for Wiener process regression with Gaussian noise of variance $\sigma^2 = 0.01$. We used two different input distributions $p(x) = 1$ (*squares*) and $p(x) = 2x$ (*triangles*), $x \in [0,1]$. The number m of example data is indicated by *arrows*. Eq. (11) is represented by the *bold line* and holds for sufficiently large m.

7 Future work

In the future, we will extend our method in the following directions:

- Obviously, our method is not restricted to a regression model but can also be directly generalized to other likelihoods such as the classification case [4, 8]. A further application to Support Vector Machines should be possible.
- We will establish further universal relations between different error measures for the more realistic case of a fixed (unknown) function $f^*(x)$. It will be interesting if such relations may be useful to construct new methods for model selection, i.e. hyper-parameter estimation.
- By computing the influence of the first neglected term in Eq. (6) which is quadratic in $H_n - H_n^0$, we will estimate the region in which our approximation is valid.

Acknowledgement

This work has been supported by EPSRC grant GR/M81601.

References

1. D. J. C. Mackay, Gaussian Processes, A Replacement for Neural Networks, NIPS tutorial 1997, May be obtained from `http://wol.ra.phy.cam.ac.uk/pub/mackay/`.
2. C. K. I. Williams and C. E. Rasmussen, Gaussian Processes for Regression, in *Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer and M. E. Hasselmo eds., 514-520, MIT Press (1996).
3. C. K. I. Williams, Computing with Infinite Networks, in *Neural Information Processing Systems 9*, M. C. Mozer, M. I. Jordan and T. Petsche, eds., 295-301. MIT Press (1997).
4. D. Barber and C. K. I. Williams, Gaussian Processes for Bayesian Classification via Hybrid Monte Carlo, in *Neural Information Processing Systems 9*, M . C. Mozer, M. I. Jordan and T. Petsche, eds., 340-346. MIT Press (1997).
5. D. Malzahn, M. Opper, Learning curves for Gaussian processes regression: A framework for good approximations, in *Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich and V. Tresp, eds., MIT Press (2001) to appear.
6. R. P. Feynman and A. R. Hibbs, Quantum mechanics and path integrals, Mc Graw-Hill Inc., 1965.
7. P. Sollich, Learning curves for Gaussian processes, in *Neural Information Processing Systems 11*, M. S. Kearns, S. A. Solla and D. A. Cohn, eds. 344 - 350, MIT Press (1999).
8. L. Csató, E. Fokoué, M. Opper, B. Schottky, and O. Winther. Efficient approaches to Gaussian process classification. In *Advances in Neural Information Processing Systems*, volume 12, 2000.