A statistical physics approach for the analysis of machine learning algorithms on real data

A statistical physics approach for the analysis of machine learning algorithms on real data

Dörthe Malzahn[1,2] and Manfred Opper[3]

[1] Institute for Mathematical Stochastics, University of Karlsruhe, D-76218 Karlsruhe, Germany
[2] Department of Genetic Epidemiology, University of Göttingen, Humboldtallee 32, D-37073 Göttingen, Germany
[3] School of Electronics and Computer Science, University of Southampton, S017 1BJ, UK
E-mail: dmalzah@gwdg.de and mo@ecs.soton.ac.uk

Abstract. We combine the replica approach of statistical physics with a variational technique to make it applicable for the analysis of machine learning algorithms on real data. The method is applied to Gaussian process models and their relative, the support vector machine. We discuss the quality of our theoretical results in comparison to experiments. As a key result, we apply our theory on real world benchmark data and show its potential for practical applications by deriving approximate expressions for data averaged performance measures which hold for general data distributions and allow us to optimize the performance of the learning algorithm.

Keywords: cavity and replica method, disordered systems (theory), analysis of algorithms

J. Stat. Mech. (2005) P11001

Contents

1. Introduction

In recent years there has been a great interest in applying methods of statistical physics within the area of complex information processing systems [1, 2]. Examples range from natural and artificial models of learning to error correcting codes and cryptosystems.

A common feature of the underlying statistical physics models is the quenched random interactions between degrees of freedom. The randomness is introduced by the random data that have to be processed by these systems. Hence, in order to understand the typical, or average, case performance of such systems, techniques from the physics of disordered materials, especially the replica method [3], have been applied with great success. One usually assumes simple and highly symmetrical probability distributions of the data leading to infinite range interactions which can be treated exactly in a mean field approach, when the data dimensionality becomes infinite.

In this way a large number of now almost classical textbook results for the generalization ability of neural networks and other systems that can learn a rule from examples have been obtained [1, 2]. While a great variety of learning scenarios were studied, the toy model architectures and simple data distributions used did not intend to describe real world learning experiments (where probability densities are usually unknown) in a quantitative way. Hence, the outcome of this research was often considered as too theoretical and not practically applicable by machine learning researchers outside the statistical physics community.

It is the goal of this paper to show that with only slightly more effort it is possible to extend the well known statistical physics approaches in such a way that they can be applied to state of the art learning machines and tested on real data sets. Our strategy will be twofold. First we combine the replica approach with a variational approximation [4] in order to be able to treat arbitrary data distributions, only assuming independence. For simple distributions we will demonstrate that the predicted learning curves come out fairly accurate even in very low dimensional cases. Second, we will show that the formalism is able to predict relations for certain performance measures of learning which hold independently of the data distribution and can thus be tested on real data [5]. We expect that such results can be useful for real world machine learning applications. The formalism will be developed for Gaussian process models, a Bayesian learning model and their relatives, the support vector machines [6, 7, 10, 11]. These can be viewed as generalizations of single-layer neural networks which are nevertheless complex and powerful enough to be of almost universal applicability in real world scenarios.

1.1. Bayesian learning

We consider a typical supervised learning scenario where the goal is to learn an unknown rule, modelled by a real valued function $f(x)$ which maps inputs $x \in R^d$ to outputs $y \in R$. The learner has access to a set D of m example data pairs (x_i, y_i), $i = 1, \ldots, m$. Of course, given only this piece of information, the task is usually ill posed, because there are typically very many functions f which would perform well on the training data. Hence, in a practical learning scenario one tries to produce a predictor $\hat{f}_D(x)$ which balances a goodness of fit on the training data as measured by a training energy

$$E[f; D] = \sum_{i=1}^{m} h(f(x_i), y_i) \tag{1}$$

together with a measure of prior knowledge about the complexity of the unknown function f. Many such methods can be derived from a probabilistic, *Bayesian* approach in which

a so-called posterior probability distribution [1, 2] over possible functions is derived by multiplying a likelihood term proportional to $\mathrm{e}^{-E[f;D]}$ with a *prior distribution* $\mu[f]$ and normalizing, i.e.,

$$\mu_m[f] = \mu[f]\,\mathrm{e}^{-E[f;D]}/Z_m. \tag{2}$$

The posterior probability distribution $\mu_m[f]$ assigns different weights to functions f of being responsible for the observed training examples D. Using proper choices of the prior $\mu[f]$ one can penalize functions that would fit the data well but are too complex to generalize properly on novel unseen inputs x. Equation (2) is of the form of a Gibbs equilibrium distribution in statistical physics. In the course of learning, when more and more data are observed, typically $\mu_m[f]$ will become increasingly concentrated around the *posterior mean*

$$\langle f(x) \rangle = \int \mathrm{d}\mu[f]\, f(x)\,\mathrm{e}^{-E[f;D]}/Z_m \tag{3}$$

which provides a natural definition of a concrete predictor. We will set $\hat{f}_D(x) = \langle f(x) \rangle$ for continuous outputs (regression problems) and $\hat{f}_D(x) = \mathrm{sgn}(\langle f(x) \rangle)$ for binary classification problems with $y = \pm 1$.

Parametric models express f through a set of parameters w, which may, e.g., be the coefficients of a polynomial function or the weights of a neural network. The distribution $\mu[f]$ is then induced by defining a prior probability distribution over the parameters w first. The drawback of parametric models is their fixed complexity which has to be chosen in advance. They have only a finite number of adjustable parameters and may thus not be able to represent functions f of arbitrary complexity. In recent years however, there has been a growing interest in so-called *non-parametric models* which are defined by assigning an *a priori* statistical weight $\mu[f]$ directly over the space of all, say smooth functions [12]–[15].

1.2. Gaussian process models

The simplest definition of a prior distribution over functions is given by a Gaussian statistics. Assuming a zero mean, Gaussian distributions are fully specified by the correlation *kernel* $K(x, x') \doteq \int \mathrm{d}\mu[f]\,\{f(x)f(x')\}$ which must be supplied by the user of the method. The kernel encodes *a priori* assumptions about the typical variability of model functions f with the input x. Such *Gaussian process* (GP) models represent a flexible and widely applicable concept [12]–[15] in the field of machine learning.

One can get an insight into the implicit statistical assumptions made by a GP prior with the help of the *eigenvalue equation* of the kernel operator

$$\int \mathrm{d}x'\, \eta(x')\, K(x, x')\phi_k(x') = \lambda_k \phi_k(x). \tag{4}$$

$\eta(x)$ is an appropriate non-negative measure (usually necessary, when the input space is not compact) and λ_k and $\phi_k(x)$ are eigenvalues and eigenfunctions respectively. The latter are orthonormal with respect to η, i.e.,

$$\int \mathrm{d}x\, \eta(x)\, \phi_k(x)\phi_l(x) = \delta_{kl}. \tag{5}$$

Finally, the kernel itself is expanded as

$$K(x, x') = \sum_k \phi_k(x)\phi_k(x')\lambda_k. \tag{6}$$

Using eigenvalues and eigenfunctions one can generate *random functions* from the prior distribution $\mu[f]$ using the expansion

$$f(x) = \sum_k w_k\sqrt{\lambda_k}\phi_k(x) \tag{7}$$

where the weights w_k are independent zero-mean, unit-variance Gaussian random variables. This is also known as the *Karhunen–Loeve expansion* [16]. Equation (7) models functions $f(x)$ as *generalized linear models* (similar to single-layer neural networks) which are linear in the usually infinitely many parameters w_k, but which can represent highly complex functions by the (nonlinear) features $\phi_k(x)$.

It is also possible to derive GP models from Bayesian feed-forward neural networks in the limit of infinitely many hidden units [13]. Finally, GPs are related to another important class of (non-probabilistic) kernel based algorithms, the so-called *support vector machines* (SVMs) [6, 8, 9]. They can be derived from the Gibbs distribution of a GP model by taking an appropriate zero-temperature limit.

We will next turn to the posterior mean prediction (3). One can express (see appendix B) this quantity as a linear combination of kernel functions centred at the training inputs

$$\langle f(x)\rangle = \sum_{j=1}^m \beta_j K(x, x_j), \tag{8}$$

where the coefficient β_j (which is found to be independent of the point x) gives the influence of data example j on the prediction. *Polynomial kernels* of the form $K(x, x') = (xx'/d)^r$ lead to predictors (8) of bounded complexity which are polynomials in x of degree r. On the other hand *transcendental kernels*, such as the *radial-basis-function* (RBF) kernel $K(x, x') = \exp(-||x - x'||/l^2)$, allow us to fit functions of arbitrary complexity when the number of example data grows. Data x_i with a large distance from x will influence predictions less than close inputs.

In the following chapters, we will study the typical learning performance of GPs under the assumption that training data D are generated at random, where all input–output pairs (x_i, y_i) are generated independently from the same distribution $p(y, x) = p(y|x)p(x)$.

The paper is organized as follows: section 2 introduces the theoretical framework followed by a presentation and first discussion of general results (section 3). Subsequently, we apply our theory to the analysis of the average learning performance of two learning algorithms which can be derived from GP models and have raised considerable interest in recent years due to their excellent performance on benchmark data: Gaussian process regression (section 4) and support vector classification (section 5). In a first step, we discuss the quality of our theoretical results in comparison to experiments using artificial data which provide well controlled model situations. In a second step, we apply our theory on real world benchmark data and show its potential for practical applications by deriving approximate expressions for data averaged performance measures which hold for general data distributions and allow us to optimize the performance of the learning algorithm. We close with a discussion (section 6).

2. The formalism

In the following, we denote data averages by square brackets $[\ldots]_D$. Using the replica approach [3], we compute the data averaged free energy $F = -[\ln Z_m]_D = -\lim_{n\to 0} \partial \ln[(Z_m)^n]_D / \partial n$ which serves as a generating functional for useful data averaged observables. The replicated and averaged partition function is given by

$$[(Z_m)^n]_D = \int \prod_{a=1}^{n} \mathrm{d}\mu[f_a] \left[\exp\left(-\sum_{a=1}^{n} E[f_a; D] \right) \right]_D. \tag{9}$$

Equation (9) becomes analytically tractable for simple and artificial data distributions in the limit of high input dimensionality. In this paper, we introduce an approximation of (9) which is of a different nature. It contains the data density as free parameter and can adapt to practically relevant situations.

2.1. The grand-canonical ensemble

To facilitate subsequent analytical calculations, we use a *grand-canonical* formulation where the number of examples m is only fixed *on average* by the fugacity ζ or chemical potential $\nu = \ln\zeta$, respectively. An elementary calculation which uses the independence of the data and the general form of the training energy $E[f; D] = \sum_{i=1}^{m} h(f; (x_i, y_i))$ yields the grand-canonical partition function for the n times replicated system

$$\Xi_n(\zeta) \doteq \sum_{m=0}^{\infty} \frac{\zeta^m}{m!} [(Z_m)^n]_D = \int \prod_{a=1}^{n} \mathrm{d}\mu[f_a] \, \mathrm{e}^{-H} \tag{10}$$

in terms of a Hamiltonian $H = [\mathcal{H}(\{f_a\}, x)]_x$ which is the average of a *purely local* Hamiltonian density

$$\mathcal{H}(\{f_a\}, x) = -\zeta \left[\exp\left\{ -\sum_{a=1}^{n} h(f_a(x), y) \right\} \right]_{y|x}. \tag{11}$$

The expectations $[\ldots]_x$, $[\ldots]_{y|x}$ are taken with respect to the input density $p(x)$ and the conditional output density $p(y|x)$, respectively. To set the value of the fugacity ζ, we consider the back transform of equation (10)

$$[Z_m^n]_D = \frac{m!}{2\pi\mathrm{i}} \oint \frac{\Xi_n(t)}{t^{m+1}} \, \mathrm{d}t. \tag{12}$$

Substituting $t = \zeta \mathrm{e}^{\mathrm{i}\phi}$ with $\phi \in [0, 2\pi)$, the Cauchy integral (12) can be approximated by its saddle point value for sufficiently large m

$$\ln[Z_m^n]_D \approx \mathcal{G}(\zeta_s) - \tfrac{1}{2}\ln(2\pi\mathcal{G}''(\zeta_s)) \tag{13}$$

with

$$\mathcal{G}(\zeta) = \ln\Xi_n(\zeta) + \ln m! - m\ln\zeta. \tag{14}$$

The value of the chemical potential $\nu \doteq \ln\zeta_s$ is fixed by the saddle point condition $\phi_s = 0$ and $\mathcal{G}'(\zeta_s) = 0$. We find for $n \to 0$

$$\zeta_s = m, \tag{15}$$

i.e., equation (13) relates the grand-canonical free energy to the original canonical free energy at a given training set size m.

2.2. The variational approach

Equation (10) is in general analytically intractable. We resort therefore to a variational approximation which replaces H by a suitable trial replica Hamiltonian H_0 such that it minimizes the variational bound [17]

$$-\ln \Xi_n(\zeta) \le -\ln \int \prod_{a=1}^n \mathrm{d}\mu[f_a]\, \mathrm{e}^{-H_0} + \langle H - H_0 \rangle_0 \qquad (16)$$

where brackets $\langle \cdots \rangle_0$ denote an average with respect to the measure $p^0(f_1, \ldots, f_n) \propto \prod_{a=1}^n \mu[f_a]\mathrm{e}^{-H_0}$. For Gaussian measures $\mu[f_a]$, a local trial Hamiltonian $H_0 = [\mathcal{H}_0(\{f_a\}, x)]_x$ of the form

$$\mathcal{H}_0(\{f_a\}, x) = \tfrac{1}{2} \sum_{a,b}^n \hat{Q}_{ab}(x) f_a(x) f_b(x) + \sum_{a=1}^n \hat{R}_a(x) f_a(x) \qquad (17)$$

is an appropriate choice. $\hat{R}_a(x)$ and $\hat{Q}_{ab}(x) = \hat{Q}_{ba}(x)$ are the variational parameters. The resulting Gaussian approximation is expected to become asymptotically exact for training energies $h(f, y)$ that are smooth functions of f, when the Gibbs distribution (2) becomes increasingly concentrated around its mean for large m. An important feature of equation (17) is the explicit dependence of the variational parameters on the input variable x. We will see later that the variationally optimal functions $\hat{Q}_{ab}(x)$ and $\hat{R}_a(x)$ are given by specific averages $[\cdots]_{y|x}$ over the conditional output density $p(y|x)$.

The variational free energy (16) can be expressed as a function of the local moments

$$R_a(x) \doteq \langle f_a(x) \rangle_0 \qquad Q_{ab}(x, x) \doteq \langle f_a(x) f_b(x) \rangle_0 \qquad (18)$$

which replace the simpler order parameters of previous replica calculations for learning problems [1] by *order parameter fields*. (Note that $Q_{ab}(x, x)$ is a special case of the general two-point function $Q_{ab}(x, x')$.) Straightforward variation of the variational free energy (16) yields $\delta\langle \mathcal{H} - \mathcal{H}_0 \rangle_0 / \delta p = 0$ for the optimal set $p = \hat{R}_a(x), \hat{Q}_{ab}(x)$ of variational parameters where the variation δ acts on the Gaussian measure $\langle \cdots \rangle_0$. The latter is fully characterized by its moments (18) and we obtain the variational equations

$$\frac{\mathrm{d}\langle \mathcal{H} \rangle_0}{\mathrm{d}R_a(x)} = \hat{R}_a(x); \qquad \frac{\mathrm{d}\langle \mathcal{H} \rangle_0}{\mathrm{d}Q_{aa}(x, x)} = \frac{1}{2}\hat{Q}_{aa}(x) \qquad \text{and} \qquad \frac{\mathrm{d}\langle \mathcal{H} \rangle_0}{\mathrm{d}Q_{ab}(x, x)} = \hat{Q}_{ab}(x) \qquad (19)$$

for $a \ne b$. We solve equation (19) under the assumption of replica symmetry, i.e., we set $\hat{R}_a(x) = \hat{R}(x)$ and $\hat{Q}_{aa}(x) = \hat{Q}_0(x)$ for all a as well as $\hat{Q}_{ab}(x) = \hat{Q}(x)$ for all $a \ne b$. The order parameter fields (18) inherit this symmetry. In the limit $n \to 0$, we obtain a simple physical interpretation of the order parameter fields as approximate values of specific data averages: $R(x) \approx [\langle f(x) \rangle]_D$ is the mean prediction of the trained model at a test input x whereas $Q_0(x, x') \approx [\langle f(x) f(x') \rangle]_D$ and $Q(x, x') \approx [\langle f(x) \rangle \langle f(x') \rangle]_D$, i.e.,

$$G(x, x') \doteq Q_0(x, x') - Q(x, x') \qquad (20)$$

$$V(x, x') \doteq Q(x, x') - R(x)R(x') \qquad (21)$$

are the average posterior correlation function and the covariance of the trained model under the data average.

In order to be able to solve the variational equations (19), we finally have to express the order parameter fields by the variational parameters. Using standard properties of Gaussian measures, we obtain from equations (17) and (18) a second set of order parameter equations

$$
\begin{aligned}
R(x) &= -[G(x,x')\hat{R}(x')]_{x'} \\
V(x,x') &= -[G(x,x'')\hat{Q}(x'')G(x'',x')]_{x''}.
\end{aligned}
\tag{22}
$$

To compute (Gaussian) averages $\langle\cdots\rangle_0$ for single replicas $f_a(x)$ we set $f_a(x) = R(x) + \tilde{f}(x)$. The effective Gaussian measure for the zero-mean random field $\tilde{f}(x)$ is found to be

$$
\mu_m^0[\tilde{f}] = \frac{\mu[\tilde{f}]\exp\left(-\frac{1}{2}[\tilde{f}^2(x')\Delta\hat{Q}(x')]_{x'}\right)}{\int \mathrm{d}\mu[\tilde{f}]\exp\left(-\frac{1}{2}[\tilde{f}^2(x')\Delta\hat{Q}(x')]_{x'}\right)}
\tag{23}
$$

where $\Delta\hat{Q}(x) = \hat{Q}_0(x) - \hat{Q}(x)$. Using the fact that the prior $\mu[\tilde{f}]$ is a Gaussian measure with covariance $K(x,x')$, the approximate posterior covariance (20) is found to be the Green's function of the operator equation

$$
\int \left(K^{-1} + U\right)(x,z)G(z,x')\,\mathrm{d}z = \delta(x-x').
\tag{24}
$$

K^{-1} is the inverse of the kernel integral operator K and $U(x,x') \doteq U(x)\delta(x-x')$ is a diagonal operator with

$$
U(x) = p(x)\Delta\hat{Q}(x).
\tag{25}
$$

Here $p(x)$ denotes the input density. As an example take the one-dimensional Wiener process (Brownian motion) where $K^{-1}(x,x') = -(\partial^2/\partial x^2)\delta(x-x')$. For this case, equation (24) becomes a one-dimensional Schrödinger equation with $U(x)$ playing the role of the potential.

3. General results

In this section, we discuss the order parameter equations and their large sample size, i.e., $m \to \infty$, behaviour. Further, we explain the derivation of general expressions for data averaged error measures and their sample fluctuations. Error measures on training and test data play an important role for the assessment of the generalization ability of a learning algorithm. Applications of the theory will be given later in sections 4 and 5. Finally, we discuss an interpretation of our results in terms of the cavity method.

3.1. The limit of large sample size

Explicit computations with the Gaussian measure (23) require the Green's function (24), which will usually be available (e.g. in terms of eigenvectors and eigenvalues of the sum of the non-commuting operators K^{-1} and U) only for simple potentials, i.e., only for simple input densities $p(x)$. Nevertheless, simplifications will occur in the limit of large training data sets $m \to \infty$. In this limit, we find that the Green's function (24)

$$
G = \left(K^{-1} + U\right)^{-1}
\tag{26}
$$

is *dominated* by the 'potential' U. To see this, we use the basic definition (25), the variational equations (19), the definition of the Hamiltonian (11) and finally equation (15) and find that $U(x)$ inherits an explicit linear dependence on the variable $\zeta_s = m$.

Motivated by similar treatments of such a 'quasi-classical' limit in quantum mechanics, we can derive a general asymptotic result for the diagonal element $G(x, x)$ by neglecting the non-commutativity of the operators U and K^{-1} [18]. Using an integral representation of the inverse of an operator and expanding with respect to a suitable orthogonal basis leads to the approximation

$$G(x, x) \approx \int_0^\infty \mathrm{d}\beta \ \mathrm{e}^{-\beta U(x)} \langle x | \mathrm{e}^{-\beta K^{-1}} | x \rangle = \sum_k \frac{|\langle x | \phi_k \rangle|^2}{\lambda_k^{-1} + U(x)} \tag{27}$$

where $\langle x | \phi_k \rangle = \phi_k(x)$ and λ_k are eigenfunctions and eigenvalues of the kernel K defined in equation (4). Note that the approximation (27) is exact if U is constant in x.

Applying the inverse operator G^{-1} to equation (22) and rearranging terms yields

$$\begin{aligned}
R(x) &= -\frac{\hat{R}(x)}{\Delta\hat{Q}(x)} - \frac{(K^{-1}R)(x)}{p(x)\Delta\hat{Q}(x)} \\
V(x, x) &= -G(x, x)\frac{\hat{Q}(x)}{\Delta\hat{Q}(x)} - \frac{(K^{-1}V)(x, x)}{p(x)\Delta\hat{Q}(x)}.
\end{aligned} \tag{28}$$

Asymptotic, 'quasi-classical' results for the order parameters are obtained by neglecting the second terms on the right-hand side of equation (28) in the large data limit $m \to \infty$.

3.2. Data average of test errors and their sample fluctuations

Besides the order parameters, one is interested in the computation of a variety of other quantities which describe the average case performance of the GP learning approach. For example, test errors play an important role for the evaluation of the generalization capability of the trained model $\hat{f}_D$. They are computed on *test* data (x, y) which were not contained in the training set D

$$\varepsilon_L(D) = \left[L(\hat{f}_D(x); x, y) \right]_{x,y} \tag{29}$$

where L denotes an arbitrary *loss function* which is designed to measure the accuracy of the prediction $\hat{f}_D$. Our replica variational approach allows us to compute the approximate *data average* of test errors $[\varepsilon_L(D)]_D$ as a function of the order parameters in a straightforward way due to the independence of test and training data. To illustrate the general argument, we specialize to the square error loss generalization error, i.e., $L(\hat{f}_D(x); x, y) = (\hat{f}_D(x) - y)^2$. We get

$$[\varepsilon(D)]_D \doteq \left[\left[(\hat{f}_D(x) - y)^2 \right]_{x,y} \right]_D \approx \left[\lim_{n \to 0} \left\langle \prod_{a=1}^2 (f_a(x) - y) \right\rangle_0 \right]_{x,y} \tag{30}$$

$$= \left[(R(x) - y)^2 + V(x, x) \right]_{x,y} \tag{31}$$

where we used the definition of $\hat{f}_D$ as the mean of the posterior Gibbs distribution (3). We replaced the expectation over the replicated, data averaged canonical ensemble by

an *approximate* expectation over the Gaussian variational distribution from the grand-canonical ensemble.

Sample fluctuations $\Delta\varepsilon_L \doteq \sqrt{[\varepsilon_L(D)^2]_D - [\varepsilon_L(D)]_D^2}$ can be calculated easily in a similar manner to equation (30). For example, using the abbreviation $\mathcal{L}_a(x,y) = f_a(x) - y$, we obtain for square error loss

$$[\varepsilon(D)^2]_D \approx \left[\lim_{n \to 0} \left\langle \prod_{a=1}^{2} \mathcal{L}_a(x,y) \prod_{b=3}^{4} \mathcal{L}_b(x',y') \right\rangle_0 \right]_{x,x',y,y'}. \tag{32}$$

This yields

$$(\Delta\varepsilon)^2 = \left[4(R(x) - y)V(x,x')(R(x') - y') + 2V^2(x,x') \right]_{x,x',y,y'}. \tag{33}$$

General loss functions L can be represented by formal Taylor expansions in the variable $\hat{f}_D(x)$. An expectation of powers of this quantity is obtained by a simple generalization of equation (30). We finally get

$$[\varepsilon_L(D)]_D \approx \left[\int \mathrm{D}z\, L\left(R(x) + z\sqrt{V(x,x)}; x, y \right) \right]_{x,y} \tag{34}$$

with the Gaussian measure $\mathrm{D}z = (2\pi)^{-(1/2)}\, \mathrm{e}^{-(1/2)z^2}\mathrm{d}z$.

3.3. Data average of empirical errors and their sample fluctuations

Empirical errors evaluate the trained model $\hat{f}_D$ on the *training* data set D

$$\varepsilon_{t,L}(D) \doteq \frac{1}{m} \sum_{i=1}^{m} L(\hat{f}_D(x_i); x_i, y_i) \tag{35}$$

which complicates the computation of the data average $[\varepsilon_{t,L}(D)]_D$. These are most conveniently calculated by a linear response method where one introduces suitable external fields in the Hamiltonian which are conjugate to the observables of interest.

As an example, we introduce the modified *canonical* partition function in replica space

$$Z_m^{(n)}(\lambda_1, \lambda_2) = \int \prod_a \mathrm{d}\mu[f_a]$$

$$\times \prod_{i=1}^{m} \exp\left(-\sum_{a=1}^{n} h(f_a(x_i), y_i) + \lambda_1 \prod_{a=1}^{2} \mathcal{L}_a(x_i, y_i) + \lambda_2 \prod_{b=3}^{4} \mathcal{L}_b(x_i, y_i) \right) \tag{36}$$

where $\mathcal{L}_a(x_i, y_i) = f_a(x_i) - y_i$. From equation (36) one can derive the square *training* error

$$[\varepsilon_t(D)]_D \doteq \left[\frac{1}{m} \sum_{i=1}^{m} (\hat{f}_D(x_i) - y_i)^2 \right]_D = \frac{1}{m} \lim_{n \to 0} \left. \frac{\partial \ln[Z_m^{(n)}(\lambda_1, \lambda_2)]_D}{\partial \lambda_1} \right|_{\lambda_1, \lambda_2 = 0} \tag{37}$$

and its sample fluctuations

$$(\Delta\varepsilon_t)^2 = \frac{1}{m^2} \lim_{n \to 0} \left. \frac{\partial^2 \ln[Z_m^{(n)}(\lambda_1, \lambda_2)]_D}{\partial \lambda_1 \partial \lambda_2} \right|_{\lambda_1, \lambda_2 = 0} \tag{38}$$

by derivatives. Using equation (36) within the grand-canonical ensemble (10) yields a modified grand-canonical Hamiltonian which is

$$H(\lambda_1, \lambda_2) = -\zeta_{\text{s}} \left[\exp\left(-\sum_{a=1}^{n} h(f_a(x), y) + \lambda_1 \prod_{a=1}^{2} \mathcal{L}_a(x,y) + \lambda_2 \prod_{b=3}^{4} \mathcal{L}_b(x,y) \right) \right]_{x,y}. \tag{39}$$

Our basic strategy is to perform the derivatives explicitly within the grand-canonical ensemble. This will translate the original canonical expectations into grand-canonical ones. Finally, grand-canonical expectations will be calculated using the approximate Gaussian measure which we have computed from the variational principle. We will explain the method for the single derivative (37). We set $\lambda_1 = \lambda$ and do not show the dependence on λ_2 for simplicity. Using equations (13) and (10), neglecting the fluctuation term $\mathcal{G}''(\zeta_{\text{s}})$, we get

$$\frac{\text{d}\ln[Z_m^{(n)}(\lambda)]_D}{\text{d}\lambda} \approx \frac{\partial(\ln \Xi_n(\zeta_{\text{s}}, \lambda))}{\partial \lambda} + \frac{\partial \mathcal{G}(\zeta_{\text{s}})}{\partial \zeta_{\text{s}}} \frac{\text{d}\zeta_{\text{s}}}{\text{d}\lambda} = -\left\langle \frac{\text{d}H(\lambda)}{\text{d}\lambda} \right\rangle \approx -\left\langle \frac{\text{d}H(\lambda)}{\text{d}\lambda} \right\rangle_0. \tag{40}$$

Note that the derivative with respect to ζ vanishes at the saddle point. For the square loss training error we get

$$[\varepsilon_{\text{t}}(D)]_D \approx \left[\lim_{n \to 0} \left\langle \exp\left(-\sum_{a=1}^{n} h(f_a(x), y) \right) \prod_{b=1}^{2} (f_b(x) - y) \right\rangle_0 \right]_{x,y} \tag{41}$$

$$= \left[\frac{(R(x) - y)^2 + V(x,x)}{(1 + G(x,x)/\sigma^2)^2} \right]_{x,y}. \tag{42}$$

Note that the average $[\cdots]_{x,y}$ on the right-hand side of equation (41) is now over *test* points. It is instructive to compare equations (30) and (41). Measures which are calculated on training data are optimistically biased towards smaller errors due to the correlation between the estimator $\hat{f}_D$ and the training data points. This is reflected by the factor $\exp(-\sum_{a=1}^{n} h(f_a(x), y))$, which modifies the replica measure and makes equation (41) explicitly dependent on the training energy h. When comparing the results (31) and (42), we see that the averaged local *training* error is given by the local *test* error $\varepsilon(x, y) = (R(x) - y)^2 + V(x,x)$, weighted by an additional factor which depends on the average posterior correlation $G(x,x) \geq 0$. The latter can be understood as a Bayesian measure for the uncertainty of the model prediction.

The analysis of the sample fluctuations for empirical errors is more complicated. We have to include the fluctuation terms around the ζ saddle point and to pay attention to the implicit λ-dependence of the saddle point $\zeta_{\text{s}}(\lambda_1, \lambda_2)$. Details of this calculation are given in appendix A. The result for the sample fluctuations of the training error yields $(\Delta\varepsilon_t)^2 = \lim_{n \to 0} (\Delta\varepsilon_{t,n})^2$, where

$$(\Delta\varepsilon_{t,n})^2 \approx \frac{1}{m} \left\langle \left[\exp\left(-\sum_{a=1}^{n} h(f_a(x), y) \right) \prod_{b=1}^{4} \mathcal{L}_b(x,y) \right]_{x,y} \right\rangle_0 - [\varepsilon_{\text{t}}(D)]_D^2 \left(1 - \frac{2}{m} \right)$$

$$+ \left(1 - \frac{3}{m} \right) \left\langle \left[\exp\left(-\sum_{a=1}^{n} h(f_a(x), y) \right) \prod_{c=1}^{2} \mathcal{L}_c(x,y) \right]_{x,y} \right.$$

$$\times \left. \left[\exp\left(-\sum_{a=1}^{n} h(f_a(x), y) \right) \prod_{d=3}^{4} \mathcal{L}_d(x,y) \right]_{x,y} \right\rangle_0. \tag{43}$$

The final result is a lengthy expression which we omit for brevity. Note that the back-transform (13) is essential for the computation of the higher order statistics of the original *canonical* system. Naturally, the fluctuations of the grand-canonical system are larger than the fluctuations of the original canonical system. Equation (13) yields the appropriate corrections.

3.4. Cavity interpretation

Using equation (41) we can compute more general data averages involving arbitrary functions of field variables at training points. We obtain

$$\frac{1}{m}\left[\sum_{i=1}^{m} g(\langle F_1(f_i)\rangle, \ldots, \langle F_k(f_i)\rangle, x_i, y_i)\right]_D = \left[\int \mathrm{D}z\; g(\langle F_1(\phi)\rangle_\phi, \ldots, \langle F_k(\phi)\rangle_\phi, x, y)\right]_{x,y} \quad (44)$$

where g and $F_1, \ldots, F_k$ are arbitrary functions, $f_i \doteq f(x_i)$ and $\langle \cdots \rangle_\phi$ denotes an expectation with respect to the distribution

$$P(\phi, h) = \frac{\exp\left[-h(\phi, y) - \frac{(R(x) + z\sqrt{V(x,x)} - \phi)^2}{2G(x,x)}\right]}{\sqrt{2\pi G(x,x)}\; Z(x, y, z)} \quad (45)$$

with a normalizing partition function $Z(x, y, z)$. Equation (44) converts any data averaged empirical estimate into an expression which depends solely on the order parameter functions R, V, and G at test inputs. The latter are linked to *generalization properties* of the learning algorithm. Moreover, we remark that the data density $p(x, y)$ is a free parameter of the theory and relation (44) holds for any density *universally* within the variational approach.

Equations (44) and (34) are similar to expressions obtained within the *cavity* approach to mean field disordered systems introduced by [3] allowing for a simple probabilistic interpretation. Let us discuss first the case where the input x is not contained in the training set. $f(x)$ resembles a magnetic field measured at the *cavity*, which is induced by removing a spin from the site x. The lack of dependences between $f(x)$ and the Hamiltonian makes the cavity field $f(x)$ a simple Gaussian random variable which we can write as the sum $\phi = \psi + \phi_c$. ψ is a zero-mean random variable with variance G which describes the (posterior) 'thermal fluctuations' of $f(x)$. The mean ϕ_c of the cavity field shows Gaussian fluctuations *from data sample to data sample* around the average $R(x)$ with a variance $V(x, x)$. When the 'site' x is included in the data sample, we introduce correlations and the original Gaussian 'thermal' probabilities become reweighted by the Boltzmann factor e^{-h}.

One can use the general results (44) and (34) to derive empirical estimators for test errors by matching the right-hand sides of equations (44) with (34), which could be highly useful in practical applications. We will give examples in the next section.

4. Application to Gaussian process regression

In the following, we apply our theory to the analysis of the average learning performance of two learning algorithms which can be derived from GP models and have raised considerable

interest in recent years due to their excellent performance on benchmark data: Gaussian process regression (section 4) and support vector classification (section 5).

Gaussian process regression models are based on the assumption that the data are generated as $y_i = f(x_i) + \xi_i$, where ξ is Gaussian white noise with variance σ^2. That is, the model has training energy

$$h(f; x, y) = \frac{1}{2\sigma^2}(y - f(x))^2. \tag{46}$$

This model finds widespread applications [12, 14] and has the advantage that the computation of Gibbs averages (3) such as means and correlation functions can be performed analytically in closed form (see appendix B). Nevertheless, the analysis of the data average is non-trivial because averaging leads to a non-Gaussian model in replica space, which makes the application of our variational approximation still necessary. We obtain for the variational equations (19)

$$\Delta \hat{Q}(x) = \frac{m}{\sigma^2 + G(x, x)} \tag{47}$$

$$\hat{Q}(x) = -\varepsilon(x) \frac{\Delta \hat{Q}^2(x)}{m} \tag{48}$$

$$\hat{R}(x) = -[y]_{y|x} \Delta \hat{Q}(x) \tag{49}$$

where $\varepsilon(x) = [(R(x) - y)^2]_{y|x} + V(x, x)$ is the average local generalization error for square loss and $\Delta \hat{Q}(x) = \hat{Q}_0(x) - \hat{Q}(x)$. $G(x, x)$ is the data averaged local posterior variance. It can be interpreted as prediction uncertainty of the model and, with training energy (46), depends solely on the inputs (see appendix B). The set of variational equations (47)–(49) must be solved together with the set of order parameter equations (22) and (24). Both sets of equations depend on the true distribution of the data.

Similar to previous studies in the statistical mechanics of learning [1, 2], we can compute explicit results when the distribution of data is given and simple enough. To illustrate this, we examine in the next two subsections translation invariant, periodic kernels with $K(x, x') = K(||x - x'||)$ and $K(x + 1) = K(x)$ in one spatial dimension and in combination with a uniform input density. We analyse the learning performance (section 4.1) and discuss the quality of the variational Gaussian approximation (section 4.2).

For a uniform input density, the *data averaged* posterior variance $G(x, x)$ (and also $\Delta \hat{Q}(x)$ by equation (47)) is independent of x and equation (27) becomes exact. Similar to equation (27), one can evaluate all order parameter functions as expressions of the eigenvalues λ_k and eigenfunctions $\phi_k(x)$ of the kernel. For periodic kernels, the kernel eigenfunctions $\phi_k(x)$ are given by the Fourier basis and the kernel eigenvalues λ_k can be calculated in a straightforward manner by a Fourier analysis of the kernel function. We consider two types of periodic kernels: first, the periodic radial-basis-function (RBF) kernel $K(x, x') = \sum_{s=-\infty}^{\infty} e^{-|x - x' - s|^2/l^2}$ with the *a priori* assumption that the input data has a characteristic correlation length l. The eigenvalues decay as $\lambda_k \simeq l\sqrt{\pi}e^{-(lk\pi)^2}$. The corresponding random functions are infinitely many times differentiable with probability one. In contrast, we consider the periodic Wiener process for one-dimensional continuous inputs. It has polynomially decaying eigenvalues $\lambda_{2k+1} = \pi^{-2}(2k + 1)^{-2}$ and $\lambda_0 = 0.25$.
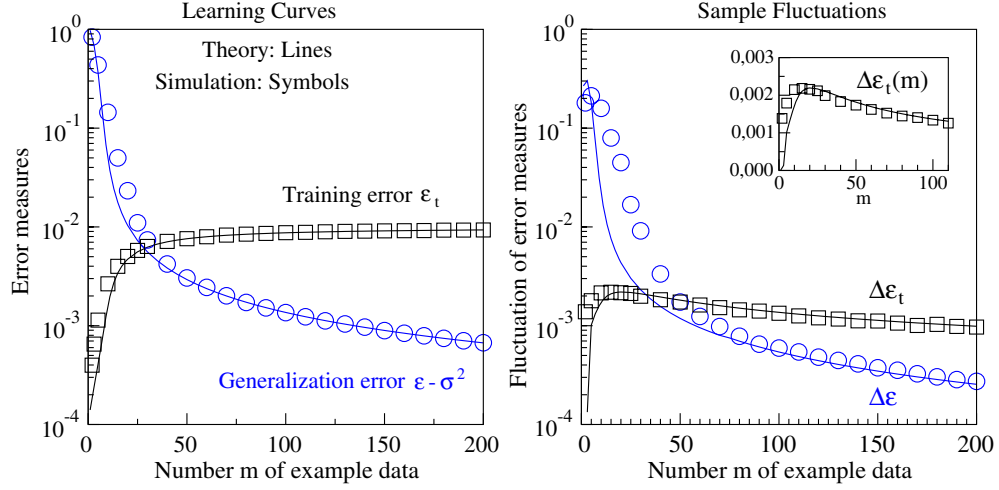
Figure 1. Learning curves (left) and their sample fluctuations (right) for the periodic RBF process with $K(x, x') = \sum_{s=-\infty}^{\infty} \exp(-(x-x'-s)^2/l^2)$ and $l^2 = 0.02$ in one dimension $x \in [0, 1]$, $p(x) = 1$ at noise variance $\sigma^2 = 0.01$. Our theory is given by lines; symbols display simulation results.

The sample paths are rough, non-differentiable random walks. For the evaluation of the quality of the variational approximation, we also consider the non-periodic Wiener process where *exact* analytical results can be obtained for the interesting case $\sigma^2 = 0$ (appendix D).

4.1. Analysis of the learning performance

Learning curves display data averaged prediction errors as a function of the number m of example data. Comparing our theory to simulations, we observe for a broad range of kernels qualitatively the same accuracy of our theoretical results of learning curves and their sample fluctuations. We demonstrate this characteristics for a periodic RBF kernel at $\sigma^2 = 0.01$ with uniformly distributed inputs $x \in [0, 1]$. For simplicity, the model is fully matched to the data which are generated from a random function drawn from the model prior and corrupted by zero-mean Gaussian noise with variance $\sigma_*^2 = 0.01$. The obtained learning curves are shown in the left panel of figure 1. For small sample sizes m, we observe a slight discrepancy between the theoretical prediction (line) and simulations (circles) of the generalization error ε, which is due to the fact that the corrections to the variational free energy are not yet negligible (compare figure 3). The theory becomes highly accurate for sufficiently large sample size m. The success of our theory for space dimension $d = 1$ is remarkable, because one might expect that the Gaussian ansatz (17) would become exact only for the mean field type of models which are realized by infinite space dimension.

It is a further benefit of our approach that we can also compute sample fluctuations (33), (43). The result is displayed in the right panel of figure 1. The theoretical prediction (line) of the fluctuations of the generalization error (circles) becomes accurate for sufficiently large m. Note that we also predict correctly the initial increase of the fluctuations for the training error (inset, right panel of figure 1).

Figure 1 was obtained by an approximate solution of the variational equations, where we approximated $\hat{Q}(x)$ by its average value $[\hat{Q}(x)]_x$. The latter is given by the average of
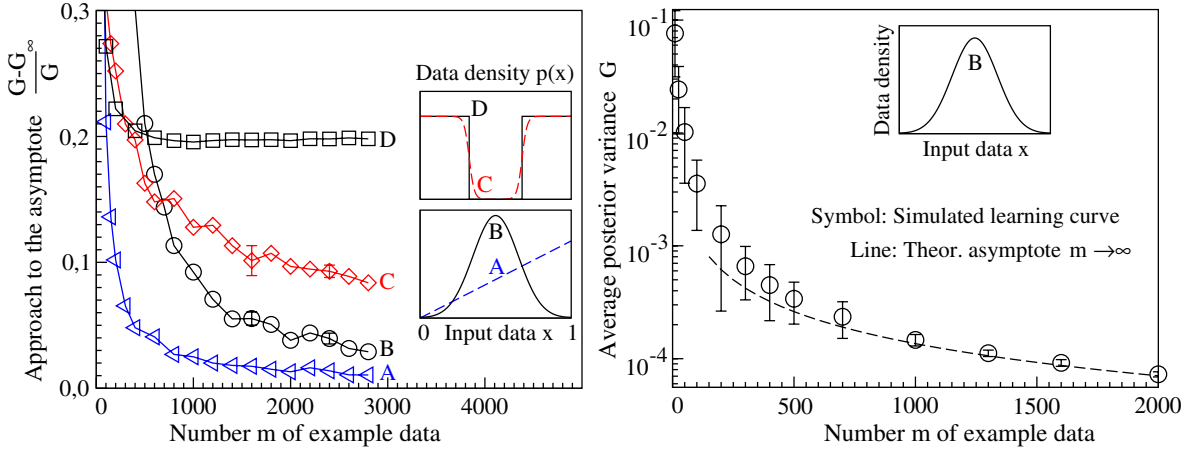
Figure 2. Limit of large sample size $m \to \infty$. Left: symbols show the relative difference $(G - G_\infty)/G$ between simulated learning curve G and theoretical asymptote G_∞, equation (50), for the posterior variance of a periodic RBF process in one spatial dimension using different input densities (insets) on the interval $[0,1]$. A (triangle) density $p(x) = 2x$; B (circle) Gaussian density; C (diamond) homogeneous density with soft slope; D (square) homogeneous density with hard edge towards forbidden region. Right: simulated learning curve (symbols) and theoretical asymptote (dashed line) of the posterior variance for example B (Gaussian inputs).

equation (48). An exact solution of equation (48) is tedious but possible, and we expect that it would improve the accuracy of the theoretical learning curve and particularly of the theoretical sample fluctuations for small and medium sample size m.

We close this section with a few remarks on the limit of large sample size $m \to \infty$. Inserting the variational equations (47)–(49) into equation (28) and assuming $\sigma^2 \neq 0$ yields the asymptote $R(x) \to [y]_{y|x}$ and $V(x,x) \to 0$ for $m \to \infty$. This result simply expresses the asymptotic consistency of the GP approach. The prediction converges to the conditional expectation of the data, which for zero-mean noise is just the true function. Note however that this asymptote might be reached through a series of overfitting maxima.

The *Bayesian* generalization error is obtained by averaging the generalization error with respect to the underlying data generating process. Here, the generalization error is measured directly with respect to the data generating process, i.e., on uncorrupted, noise-free data. For adapted models, the probability density of the data generating process is identical to the GP model prior and the Bayesian generalization error is given by the posterior variance. With equation (27), we obtain for the average posterior variance $[G(x,x)]_x$ in the large sample size limit $m \to \infty$ the following expression:

$$[G(x,x)]_x \doteq \int \mathrm{d}x \, p(x) G(x,x) \approx \sum_k \omega_k \int \mathrm{d}x \, \frac{p(x)\lambda_k}{1 + p(x)\lambda_k(m/\sigma^2)} \tag{50}$$

where we set $\Delta \hat{Q}(x) \approx m/\sigma^2$ for large m (see equation (47)) and $p(x)$ is an arbitrary input density. ω_k denotes the degree of degeneracy of kernel eigenvalue λ_k. Equation (50) is illustrated by figure 2 for a periodic RBF process $K(x,x') = \sum_s \mathrm{e}^{-(x-x'-s)^2/l^2}$ with
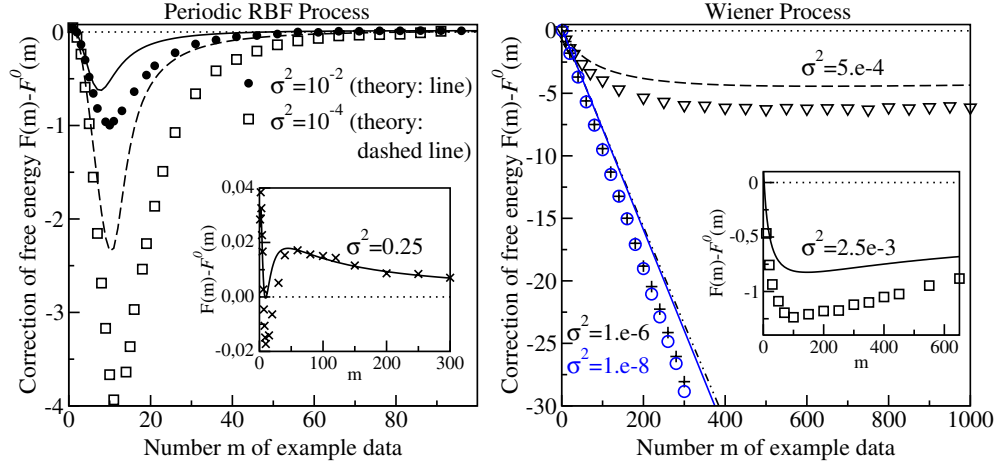
Figure 3. Correction to the free energy. Left: periodic RBF process, $K(x, x') = \sum_s \exp(-(x - x' - s)^2/l^2)$ with $l^2 = 0.02$. Right: Wiener process, $K(x, x') = \max(x, x')$. All y data were set equal to zero, $x \in [0, 1]$ and $p(x) = 1$. Symbols: we subtracted the variational free energy (16) from the true value of the free energy. The latter was obtained by simulations. Lines show the theoretical estimate (53) (except the dotted line at $F - \mathcal{F}^0 = 0$). The noise variance σ^2 decreases from top to bottom. For the Wiener process (right panel), exact analytical results can be obtained for $\sigma^2 = 0$ (appendix D).

$l^2 = 0.02$ and $\sigma^2 = 0.01$ in one spatial dimension using different input densities. The left panel of figure 2 shows the *relative difference* $(G - G_\infty)/G$ between simulated learning curve G and theoretical asymptote G_∞, equation (50), respectively. For Gaussian inputs (example B), we show in the right panel of figure 2 a direct comparison between simulated learning curve G (symbols) and theoretical asymptote G_∞ (dashed line).

The asymptote (50) appears to be valid for continuous data densities $p(x)$ (examples A–C in figure 2, left panel) where we find that the relative difference $(G - G_\infty)/G \to 0$ as $m \to \infty$. Convergence to the asymptote (50) will be slower for continuous data densities with a higher amount of structure, for smaller values of the noise variance σ^2, or for smaller values of the correlation length l in the kernel. Within the framework of our theory, the derivation of equation (50) was based on the additional assumption that the non-commutativity of the operators K^{-1} and U can be neglected in the limit $m \to \infty$. This assumption appears to be violated for non-continuous densities (see example D in figure 2, left panel). In support of these findings, we refer to the Wiener process, where equation (24) becomes a one-dimensional Schrödinger equation with $U(x) = p(x)\Delta\hat{Q}(x)$ playing the role of the potential. For this example it is known [18] that corrections to the approximation (27), (50) contain derivatives of the potential $U(x)$.

4.2. Quality of the variational approximation

To characterize the quality of the Gaussian variational approximation, we calculated the difference $F - \mathcal{F}^0$ approximatively, where F denotes the exact canonical free energy and

$\mathcal{F}^0$ the variational free energy (16) of the corresponding grand-canonical system. We obtain

$$F - \mathcal{F}^0 \approx \lim_{n \to 0} \frac{\partial}{\partial n} \left(\frac{1}{2} \ln \left(\langle H^2 \rangle_0 - \langle H \rangle_0^2 + m \right) - \frac{1}{2} \left(\langle (H - H_0)^2 \rangle_0 - \langle H - H_0 \rangle_0^2 \right) \right). \quad (51)$$

The first contribution to equation (51) accounts for the difference between canonical and grand-canonical free energy. It can be obtained from equation (13) with $\zeta_s = m$. The second term in equation (51) is a correction to the variational approximation $\mathcal{F}_n^0$ of the grand-canonical free energy $-\ln \Xi_n$ using the perturbation expansion

$$-\ln \Xi_n \approx \mathcal{F}_n^0 - \tfrac{1}{2} \left(\langle (H - H_0)^2 \rangle_0 - \langle H - H_0 \rangle_0^2 \right) \pm \cdots \quad (52)$$

for the replicated system. In the following, we consider for simplicity the case where all y data are set equal to zero. This is still an interesting model from which the data averaged posterior variance $G(x, x)$ can be estimated. Equation (51) yields

$$F - \mathcal{F}^0 \approx \tfrac{1}{4}[\Delta \hat{Q}(x) \Delta \hat{Q}(x') G^2(x, x')]_{x', x}$$
$$+ \frac{1}{4}(m^2 - m) \left[\ln \left(1 - \frac{G^2(x, x')}{(G(x, x) + \sigma^2)(G(x', x') + \sigma^2)} \right) \right]_{x', x} \quad (53)$$

where we used equation (47) to simplify expression (53). Figure 3 shows the correction to the free energy for different values of the noise σ^2 and two types of kernels in a one-dimensional input space, $x \in [0, 1]$ and $p(x) = 1$.

The symbols in figure 3 show the difference between the true value of the free energy which is obtained by simulations and the variational free energy (16). The lines correspond to the correction (53). The variational approximation is expected to break down for the noise-free case $\sigma^2 = 0$ where the Gaussian approximation of the posterior Gibbs distribution is no longer appropriate. As we approach this limit $\sigma \to 0$, corrections to the variational free energy become more severe (figure 3 from top to bottom). It is interesting to note that the first correction term (53) allows quantitatively a very reliable assessment of the quality of the variational approximation for all values of σ^2. Until the noise free case $\sigma^2 = 0$ is reached (see appendix D for more details), we see that the difference between the free energy and its variational approximation either saturates at a value of order $\mathcal{O}(1)$ or even decreases with growing m. The free energy itself grows with m, which implies that the correction becomes less and less important for $m \to \infty$.

4.3. Universal relations between error estimates

In praxis, data distributions are usually unknown and may be much more structured than the simple model densities discussed so far. This makes it impossible to solve the set of order parameter and variational equations analytically. However, we can use the replica variational approach to derive *relations* between observables such as data averaged empirical measures and average generalization properties of the algorithm. These relations can be checked in cross-validation experiments where the available data are split into training sets and a hold out data set for testing. In the following, we will give a variety of examples. We will mostly concentrate on regression problems with square error loss.

Equation (41) yields

$$\varepsilon_{\mathrm{t}} = \left[\frac{\varepsilon(x,y)}{(1 + G(x,x)/\sigma^2)^2} \right]_{x,y} \tag{54}$$

which relates the data averaged training error ε_{t} to the average local generalization error $\varepsilon(x,y) = [(\hat{f}_D(x) - y)^2]_D$ and the average local posterior variance $G(x,x)$ at test inputs x. A second relation is obtained from equation (44) for the data averaged empirical posterior variance

$$\sigma_{\mathrm{t}}^2 = \frac{1}{m} \left[\sum_{i=1}^{m} \langle f(x_i)^2 \rangle - \langle f(x_i) \rangle^2 \right]_D \approx \left[\frac{G(x,x)}{1 + G(x,x)/\sigma^2} \right]_x \tag{55}$$

as a function of the average local posterior variance $G(x,x)$ at test inputs x. Within the variational approximation, relations (54) and (55) hold for any density of data. To test their validity we have chosen three common benchmark data sets [20, 21] and find that both relations hold well even for small and medium numbers of training data. Equation (54) is displayed in the left panel of figure 4. Learning starts (with a small number of examples) in the lower left corner. The training error ε_{t} on the noisy data set is initially small and increases with increasing number of example data. Equation (55) is displayed in the right panel of figure 4. Learning starts in the upper right corner as the rescaled empirical posterior variance $\sigma_{\mathrm{t}}^2/\sigma^2$ is initially one and decreases with increasing number of example data. The symbols are simulation results from cross-validation experiments where the right-hand side of equations (54), (55) is computed on a hold-out data set for testing. For GP regression, the input data were preprocessed by rescaling it component-wise to zero mean and unit variance. We used a non-periodic RBF kernel with a single correlation length.

Equations (55) and (54) have in common that they express empirical error measures as functions of the generalization properties of the model. For a practical purpose it is more convenient to have the inverse relationships, where generalization errors are expressed in terms of empirical error measures. Such results can be used to construct estimates for the generalization ability of the trained model which are computable from the training data D alone. The generalization error $\varepsilon_L(D) = [L(\langle f(x) \rangle; x, y)]_{x,y}$ is measured with a suitable loss function L. Obviously, using the naive approximation $(1/m) \sum_{i=1}^{m} L(\langle f(x_i) \rangle; x_i, y_i)$, i.e., the corresponding training error, will only give an optimistically biased estimate of the generalization error. This bias can be removed by replacing $\langle f(x_i) \rangle$ in this expression by a term $\langle f(x_i) \rangle_{\backslash i}$, which is defined as the posterior mean prediction at x_i, when the input x_i is deleted from the training set. In the language of the *cavity* method, this is the mean cavity field ϕ_{c} (see section 3.4). On the other hand, in the area of *machine learning* the same quantity leads to the so-called *leave-one-out* estimators [19]. An exact computation of $1/m \sum_{i=1}^{m} L(\langle f(x_i) \rangle_{\backslash i}; x_i, y_i)$ would require the time consuming retraining of a GP predictor on a different data set each time a data point x_i is deleted. On the other hand, our variational approximation provides us with explicit (approximate) expressions for this quantity, which can often be computed with an effort that is similar to the computation of the original prediction $\langle f(x_i) \rangle$. A lengthy but straightforward calculation involving integration by parts in equation (44) and a comparison with equation (34) shows that the desired result is given by

$$\langle f(x_i) \rangle_{\backslash i} = \langle f_i \rangle + \gamma_i \langle h'(f_i) \rangle \tag{56}$$
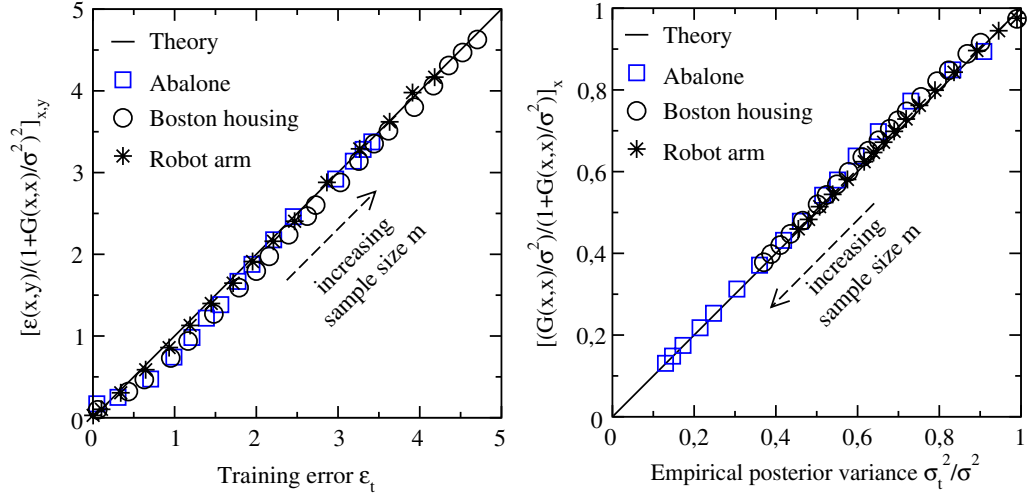
Figure 4. Test of the universal relations (54) (left) and (55) (right) on real data for GP regression with RBF kernel $K(x, x') = \exp(-||x - x'||^2/l^2)$. Circles: Boston housing data [20] ($l^2 = 90$, $\sigma^2 = 0.01$, input dimension $d = 13$). Squares: Abalone data [20] ($l^2 = 10$, $\sigma^2 = 0.1$, $d = 10$). Stars: Puma8nm Robot arm data [21] ($l^2 = 60$, $\sigma^2 = 0.01$, $d = 8$). Details on the three benchmark data sets are given in appendix E.

where

$$\gamma_i = \frac{\langle f_i h'_i \rangle - \langle f_i \rangle \langle h'_i \rangle}{\langle h''_i \rangle - (\langle (h'_i)^2 \rangle - \langle h'_i \rangle^2)} \tag{57}$$

and $h_i = h(f_i, y_i)$, $h'_i = \partial h(f_i, y_i)/\partial f_i$ etc. Here, $\langle \cdots \rangle$ denotes an average over the posterior Gibbs distribution (2). Relations (56) and (44) will yield the unbiased estimates

$$\left[\frac{1}{m} \sum_{i=1}^m L(\langle f(x_i) \rangle_{\backslash i}; x_i, y_i) \right]_D \approx \left[\int \mathrm{D}z \, L\left(R(x) + z\sqrt{V(x, x)}; x, y \right) \right]_{x,y} \approx [\varepsilon_L(D)]_D \tag{58}$$

for generalization errors. We will next test our method for Gaussian regression, i.e., setting $h(f, y) = (1/2\sigma^2)(f - y)^2$. In this case, we find

$$[\varepsilon_L(D)]_D = \left[\frac{1}{m} \sum_{i=1}^m L\left(\frac{\sigma^2 \langle f(x_i) \rangle - \sigma_i^2 y_i}{\sigma^2 - \sigma_i^2}; x_i, y_i \right) \right]_D \tag{59}$$

where $\langle f(x_i) \rangle = \hat{f}_D(x_i)$ denotes the prediction of the model at training point i and $\sigma_i^2 = \langle f^2(x_i) \rangle - \langle f(x_i) \rangle^2$ is the model uncertainty (posterior variance) for this prediction. Equation (59) can be applied without knowledge about the data distribution. It can be shown that equation (59) is in fact *the exact leave-one-out* estimator for GP regression [7,12]. We present tests of equation (59) on regression and binary classification problems. In both cases $\langle f(x) \rangle$ is computed from the squared error h, equation (46), with $\sigma^2 = 0.01$ (Boston data) and $\sigma^2 = 0.1$ (Sonar data), respectively. The left panel of figure 5 compares the generalization errors and their estimates for the publicly available *Boston housing* regression data set [20] using the error function $L(\langle f(x) \rangle; x, y) = |\langle f(x) \rangle - y|^k$ for $k = 1$, $m = 50$ and
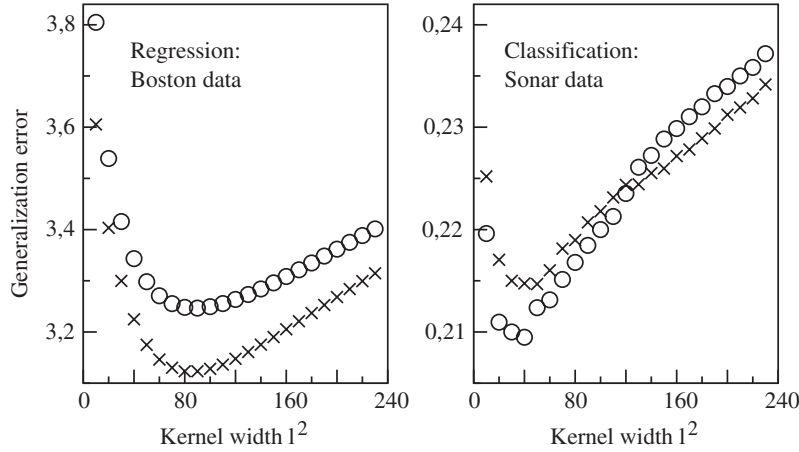
Figure 5. Model selection based on equation (59) for a regression problem (Boston housing data [20]) and a binary classification problem (Sonar data [20]) using the kernel $K(x, x') = \exp(-||x - x'||^2/l^2)$. The right-hand side of equation (59) is an empirical estimate (crosses) which is calculated only on *training* data. It gives a good account of the generalization error (circles) which was calculated on *test* data.

different widths l^2 of the kernel $K(x, x') = \exp(-||x - x'||^2/l^2)$. The results (which are averaged over 20 splits) suggest that our estimators might be well used for *model selection*, i.e., for finding the optimal kernel parameters with smallest generalization error.

The right panel of figure 5 shows corresponding results for binary classification ($y = \pm 1$) on the *Sonar data*, where now the generalization error is the average fraction of misclassified test points, i.e., $L(\langle f(x) \rangle; x, y) = \Theta(-y\langle f(x)\rangle)$, and $\Theta(x)$ is the unit step function with $\Theta(x) = 1$ for $x > 0$ and 0 otherwise.

5. Support vector machines

In this section, we will test the validity of our variational replica approach for a case where the combination of a non-smooth training energy with a subtle singular limit might not suggest immediately that the trial Gaussian distribution (17) is a good approximation. We consider SVM classifiers [8, 9] which can be understood as generalizations of single-layer neural networks which allow for *nonlinear* separation between classes. Using the expansion of the positive definite SVM kernel $K(x, x') = \sum_k \psi_k(x)\psi_k(x')$ into a suitable set of functions $\psi_k(x)$, the SVM output can be written as $y = \mathrm{sgn}[f(x)]$ where $f(x) = \sum_k w_k \psi_k(x)$. The $\psi_k(x)$ serve as (usually) nonlinear features of the inputs. The vector $\mathbf{w}$ of SVM weights w_i is determined by the following optimization problem: its squared length $||\mathbf{w}||^2 = \sum_k w_k^2$ must be minimized under the condition that the 'hard margin' training energy $h_{\mathrm{HM}} = 0$ at each training data point. This energy is defined as $h_{\mathrm{HM}}(f(x), y) = 0$ if $yf(x) \geq 1$ and $h_{\mathrm{HM}} = \infty$ else. For any data set D, the SVM prediction can be written in the form of equation (8). Using the typical SVM notation [7], we write $\beta_i = y_i \alpha_i$ so that

$$\hat{f}_D(x) = \sum_{i=1}^{m} y_i \alpha_i K(x_i, x) \tag{60}$$

where the coefficients $\alpha_i \geq 0$ are determined by a convex optimization problem (for details see appendix B.2). Only those data points on the *margin*, i.e., for which $\hat{f}_D(x_i)y_i = 1$, the so-called *support vectors* contribute to the sum. For all other data points $\alpha_i = 0$.

To see how SVM models fit into our present GP framework [19] we introduce a Gaussian prior distribution

$$\mu_\epsilon(\mathbf{w}) \propto \exp\left(-\frac{1}{2\epsilon}\sum_k w_k^2\right) \tag{61}$$

over weights. ϵ is a temperature-like parameter, which controls the fluctuations. We define a (pseudo-) posterior distribution

$$\mu_{\epsilon,m}(\mathbf{w}) \propto \mu_\epsilon(\mathbf{w})\exp\left(-\sum_{i=1}^m h_{\mathrm{HM}}(f(x_i), y_i)\right). \tag{62}$$

In the limit $\epsilon \to 0$, the posterior distribution becomes obviously peaked at the minimal length weight vector of the SVM, where the likelihood term takes care of the constraints. Finally, we use the fact that the Gaussian prior distribution over weights (61) defines a Gaussian process measure $\mu_\epsilon[f]$ over functions $f(x) = \sum_k w_k \psi_k(x)$. It is easy to see that its correlation kernel is given by $K_\epsilon(x, x') = \epsilon K(x, x')$. Hence, to extend our framework to SVM models, all we need to do is to control the limit $\epsilon \to 0$ within the variational and order parameter equations.

To do this, we use a proper definition of the grand-canonical free energy as

$$F = -\lim_{n,\epsilon \to 0} \epsilon \ln \int \prod_{a=1}^n \mathrm{d}\mu_\epsilon[f_a]\mathrm{e}^{-H} \tag{63}$$

which remains finite for $\epsilon \to 0$. The Hamiltonian is given by $H = [\mathcal{H}(\{f_a\}, x)]_x$

$$\mathcal{H}(\{f_a\}, x) = -\zeta\left[\prod_{a=1}^n \Theta(yf_a(x) - 1)\right]_{y|x}. \tag{64}$$

Using the Gaussian trial distribution (17), we obtain the order parameter equations (22) and (24) and equations (C.1)–(C.3) for the rescaled variational parameters $\Delta\hat{Q}_\epsilon(x)$, $\hat{Q}_\epsilon(x)$ and $\hat{R}_\epsilon(x)$, which are defined as $\epsilon\Delta\hat{Q}(x)$, $\epsilon^2\hat{Q}(x)$ and $\epsilon\hat{R}(x)$ and have non-trivial limits for $\epsilon \to 0$. Note that the vanishing covariance $G(x, x')$ is replaced by the response function

$$\chi(x, x') = \lim_{\epsilon \to 0} \epsilon^{-1} G(x, x'). \tag{65}$$

Details are given in appendix B.2 and in appendix C. We will next give a few results for the average case performance of SVMs predicted by our theory.

We start with the average number of support vectors mn_{SV}. It can be obtained from the data averaged local field distribution

$$\rho(h) \doteq \frac{1}{m}\left[\sum_{i=1}^m \delta(h - y_i\langle f(x_i)\rangle)\right]_D \tag{66}$$
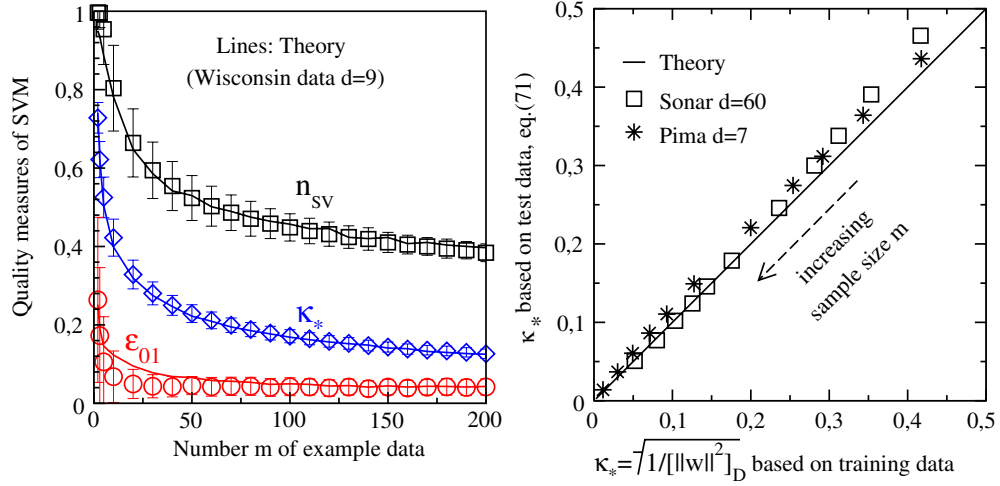
Figure 6. Learning curves for hard margin support vector learning with RBF kernel $K(x, x') = \exp(-||x - x'||^2/l^2)$ on binary classification benchmark data. Left: generalization error ε_{01} (circles), approximate margin κ_* (diamonds) and fraction of support vectors n_{SV} (squares) as computed from their basic definitions on Wisconsin breast cancer data [20] (input dimension $d = 9$, $l^2 = 3$). Lines show the theoretical results (69), (71) and (68) which were obtained on test data. Right: verification of equation (71) on Pima Indian diabetes data [22] (stars, input dimension $d = 7$) and on Sonar data [20] (squares, $d = 60$) with kernel width $l^2 = 3d$. The arrow points into the direction of increasing number m of training data with $m = 4$–7 in steps of 1, $m = 9$–24 in steps of 5 and $m = 33, 50, 100$.

which has a delta-peak at the margin value $h = 1$ realized by all support vectors. Using the general relation (44), we obtain

$$\rho(h) = \delta(h - 1) \left[\Phi\left(\Delta_{xy}\right)\right]_{x,y} + \left[\int_{-\infty}^{-\Delta_{x,y}} \mathrm{D}v \, \delta\left(h - yR(x) + v\sqrt{V(x,x)}\right)\right]_{x,y}, \quad (67)$$

where $\Delta_{xy} = (1 - yR(x))/\sqrt{V(x,x)}$ and $\Phi(x) = \int_{-\infty}^{x} \mathrm{D}v$. Hence, the average percentage of support vectors

$$n_{SV} = \left[\Phi\left(\frac{1 - yR(x)}{\sqrt{V(x,x)}}\right)\right]_{x,y} \quad (68)$$

can be expressed as a function of the SVM's bias and variance at novel test points x. From equation (68) we verify easily the well known result that the average percentage of support vectors n_{SV} is an upper bound to the average generalization error for binary classification, which is, following equation (34),

$$\varepsilon_{01} \doteq \left[\Theta(-y\langle f(x)\rangle)\right]_{D;x,y} = \left[\Phi\left(\frac{-yR(x)}{\sqrt{V(x,x)}}\right)\right]_{x,y}. \quad (69)$$

As a third quantity, we consider the so-called *margin* $\kappa(D) = 1/||\mathbf{w}||$, which (as a minimum distance between positive and negative labelled examples) plays a role as an

indicator for good generalization ability of SVMs [8]. We can obtain some information about the margin within our framework from the free energy, noticing that

$$\frac{1}{\kappa^2(D)} = ||\mathbf{w}||^2 = -2 \lim_{\epsilon \to 0} \epsilon \ln \int d\mu_\epsilon(\mathbf{w}) \, \exp\left(-\sum_{i=1}^{m} h_{\mathrm{HM}}(f(x_i), y_i)\right). \quad (70)$$

Averaging over data, we find that $1/\kappa_*^2 \doteq [1/\kappa^2(D)]_D = 2F$ where F is the grand-canonical free energy (63). Inserting the optimal values for the variational parameters (C.1)–(C.3) into the general expression for the free energy yields

$$\frac{1}{\kappa_*^2} = [\alpha(x,y)]_{xy} \quad (71)$$

where

$$\alpha(x,y) = \frac{m\sqrt{V(x,x)}}{\chi(x,x)} \int_{-\infty}^{\Delta_{xy}} Dv \, (\Delta_{x,y} - v). \quad (72)$$

The function $\alpha(x,y)$ is always non-negative and can be viewed as an averaged version of α_i in equation (60). In fact, we can show that the data averaged prediction at a point x is given by

$$R(x') = [y\alpha(x,y)K(x,x')]_{x,y}. \quad (73)$$

In regions with small $\alpha(x,y)$, data points have a very low probability of being selected as a support vector and will hardly affect the data averaged prediction $R(x)$. These regions are typically far away from the decision surface between positive and negative labelled examples and thus show a high stability of predictions, indicated by a small variance $V(x,x) \approx 0$.

Equation (71) relates the averaged inverse squared margin on the training data to the SVM's *bias*, its *variance* and the response function χ at novel test data (x,y). It generalizes Gardner's famous result [23] for the optimal stability of linear perceptrons to SVMs under general data distributions. The left panel of figure 6 compares the approximate margin κ_*, the relative number of support vectors n_{SV} and the generalization error for binary classification ε_{01} as computed from their basic definition (symbols) with our theoretical prediction (lines) given by equations (71), (68) and (69). We used a publicly available classification data set [20] and averaged over random splits of the entire data set into training and test examples. The input data was preprocessed by rescaling it component-wise to zero mean and unit variance. Using an RFB kernel $K(x,x') = \exp(-||x-x'||^2/l^2)$, we observe qualitatively similar results for several other common benchmark classification data sets [20, 22] and a broad range of values of the kernel correlation length l. The variational Gaussian approximation improves with higher data dimensionality (right panel of figure 6) and becomes very good for sufficiently large training sizes m.

For a small number m of example data and very narrow kernel functions $K(x,x')$ (which show a fast decay with respect to the distance $||x-x'||$), one finds $|R(x)| \ll 1$, $V(x,x) \ll 1$ and $\chi(x,x) \approx 1$ for the majority of test inputs x. For this extreme case, we find theoretically and experimentally for the approximate margin the data-independent result $\kappa_* = 1/\sqrt{m}$. Recently, universal behaviour has been reported for some asymptotic $m \to \infty$ learning properties of hard margin SVM on noisy data [11] under simple artificial

data distributions. It would be interesting also to establish such universal features with the variational approach.

Finally, we would like to mention that it is possible to generalize the result (58) for approximate leave-one-out estimators to SVMs. The calculation is sketched in appendix B.2 and the result agrees with that of a previous independent approach of Opper and Winther [19], who used a TAP mean field approach.

6. Discussion

We have presented a theoretical framework which allows us to analyse the average case performance of learning algorithms under arbitrary data distributions. The method yields explicit results for learning curves, when such a distribution is given. More important, one can derive useful *relations between observables* which can be tested on real data and can be used to optimize the learning performance.

In this paper, we have performed a extensive analysis for the case of Gaussian process models and their relatives, the hard margin support vector classifiers. We expect that this represents only a small selection of the possible applications of our approach. Obvious future extensions could include an assessment of model selection and optimization criteria which are based on estimators for generalization errors such as equation (59) or on a free energy type of criterion for the model complexity. For this task, one would have to investigate sample fluctuations of such quantities, which is possible within our approach.

It will also be interesting to generalize the variational approach to models with statistically *dependent* data such as (hidden) Markov processes and their relatives, which are highly relevant for time series prediction. Taking into account non-Gaussian prior distributions and other trial Hamiltonians will extend the applicability of our method to the wider field of modern probabilistic data modelling. Applying our methods to fields like combinatorial optimization will also require an extension to *replica symmetry breaking* solutions of the variational equations.

We also believe that an interesting future direction of our variational replica approach will be in the development of inference (i.e. learning) algorithms for complex probabilistic data models. The efficient approximate computation of posterior mean predictions for such models using a cavity type of idea [25]–[27] is at present an active area of research within statistical physics. These approaches are known to become exact in a 'thermodynamic limit' framework, but at present there is not much control over their accuracy in real applications. We hope that our approach could be used to construct algorithms which are *guaranteed* to work well *on average* for realistic system sizes and data distributions. The possibility of improving our approximations perturbatively might then be turned into a method for testing and improving the accuracy of approximate inference methods.

Acknowledgments

of Denmark, Denmark, and from the postgraduate programme 'Natural disasters' at the University of Karlsruhe, Germany.

Appendix A. Computation of the average square loss training error and its fluctuations

This appendix provides details for the computation of the average square loss training error and its sample fluctuations (37), (38) by a linear response method. Equation (13) relates the free energy $F_n = -\ln[Z_m^n(\lambda_1, \lambda_2)]_D$ of the replicated and data averaged canonical system (36) to the corresponding grand-canonical free energy $\mathcal{F} = -\ln \Xi_n(\lambda_1, \lambda_2) = -\ln \int \prod_{a=1}^n \mathrm{d}\mu[f_a]\mathrm{e}^{-H(\lambda_1,\lambda_2)}$ with Hamiltonian

$$H(\lambda_1, \lambda_2) = -\zeta_{\mathrm{s}}\left[\exp\left(-\sum_{a=1}^n h(f_a(x), y) + \lambda_1 \prod_{a=1}^2 \mathcal{L}_a(x,y) + \lambda_2 \prod_{b=3}^4 \mathcal{L}_b(x,y)\right)\right]_{x,y} \quad (A.1)$$

where $\mathcal{L}_a(x,y) = f_a(x) - y$. We use equation (13) and obtain

$$\frac{\partial \ln[Z_m^n(\lambda_1, \lambda_2)]}{\partial \lambda_1} = -\mathcal{F}_1 + \frac{\mathcal{F}_1''}{2\mathcal{G}''(\zeta_{\mathrm{s}})} + \frac{\partial \zeta_{\mathrm{s}}}{\partial \lambda_1}\left(\mathcal{G}'(\zeta_{\mathrm{s}}) + \frac{1}{2\mathcal{G}''(\zeta_{\mathrm{s}})}\left(\frac{2m}{\zeta_{\mathrm{s}}^3} + \mathcal{F}'''\right)\right) \quad (A.2)$$

where $\mathcal{F}_1 = -(\partial \ln \Xi_n(\lambda_1, \lambda_2))/\partial \lambda_1$ denotes the linear response of the replicated and averaged grand-canonical system. The third term of equation (A.2) is due to the implicit λ_1-dependence of the saddle point ζ_{s}. In the following, we neglect the small contribution $\mathcal{F}''' = -(\partial^3 \ln \Xi_n(\lambda_1, \lambda_2))/\partial^3 \zeta_{\mathrm{s}} \approx 0$. (The energy $\mathcal{F}$ is approximatively linear in ζ_{s}.) Within this approximation

$$\frac{\partial^2 \ln[Z_m^n(\lambda_1, \lambda_2)]}{\partial \lambda_1 \partial \lambda_2}$$

$$\begin{aligned}= & -\mathcal{F}_{12} + \frac{\mathcal{F}_{12}''}{2\mathcal{G}''(\zeta_{\mathrm{s}})} + \frac{\partial^2 \zeta_{\mathrm{s}}}{\partial \lambda_1 \partial \lambda_2}\left(\mathcal{G}'(\zeta_{\mathrm{s}}) + \frac{m}{\zeta_{\mathrm{s}}^3 \mathcal{G}''(\zeta_{\mathrm{s}})}\right) \\
& + \frac{\mathcal{F}_1'' \mathcal{F}_2''}{2(\mathcal{G}''(\zeta_{\mathrm{s}}))^2} + \sum_{(i,j)=(1,2),(2,1)} \frac{\partial \zeta_{\mathrm{s}}}{\partial \lambda_i}\left(-\mathcal{F}_j' + \frac{m}{\zeta_{\mathrm{s}}^3 (\mathcal{G}''(\zeta_{\mathrm{s}}))^2} \mathcal{F}_j''\right) \\
& + \frac{\partial \zeta_{\mathrm{s}}}{\partial \lambda_1} \frac{\partial \zeta_{\mathrm{s}}}{\partial \lambda_2}\left(\mathcal{G}''(\zeta_{\mathrm{s}}) - \frac{3m}{\zeta_{\mathrm{s}}^4 \mathcal{G}''(\zeta_{\mathrm{s}})} + \frac{2m^2}{\zeta_{\mathrm{s}}^6 (\mathcal{G}''(\zeta_{\mathrm{s}}))^2}\right).\end{aligned} \quad (A.3)$$

Using the saddle point condition $\mathcal{G}'(\zeta_{\mathrm{s}}) = 0$, which yields

$$\zeta_{\mathrm{s}} = -\frac{m}{\mathcal{F}'} \quad (A.4)$$

we can write

$$\begin{aligned}\frac{\partial \zeta_{\mathrm{s}}}{\partial \lambda_1} &= \frac{\zeta_{\mathrm{s}}^2}{m} \mathcal{F}_1' \\
\frac{\partial^2 \zeta_{\mathrm{s}}}{\partial \lambda_1 \partial \lambda_2} &= \frac{\zeta_{\mathrm{s}}^2}{m} \mathcal{F}_{12}' + \frac{2\zeta_{\mathrm{s}}^3}{m^2} \mathcal{F}_1' \mathcal{F}_2'.\end{aligned} \quad (A.5)$$

We can now calculate the linear response of the canonical system from the linear response $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_{12}$ of the grand-canonical system. With equation (39), we obtain at $\lambda_1, \lambda_2 = 0$

$$\mathcal{F}_1(0,0) = \zeta_{\mathrm{s}}\left[\left\langle \exp\left(-\sum_{a=1}^n h(f_a(x), y)\right) \prod_{b=1}^2 \mathcal{L}_b(x,y)\right\rangle\right]_{x,y} \quad (A.6)$$

and

$$\mathcal{F}_{12}(0,0) = \zeta_{\mathrm{s}} \left[\left\langle \exp\left(-\sum_{a=1}^{n} h(f_a(x), y) \right) \prod_{b=1}^{4} \mathcal{L}_b(x, y) \right\rangle \right]_{x,y}$$
$$+ \zeta_{\mathrm{s}}^2 \mathrm{VAR}\left(\left[\exp\left(-\sum_{a=1}^{n} h(f_a(x), y) \right) \prod_{b=1}^{2} \mathcal{L}_b(x, y) \right]_{x,y}, \right.$$
$$\left. \left[\exp\left(-\sum_{a=1}^{n} h(f_a(x'), y') \right) \prod_{c=3}^{4} \mathcal{L}_c(x', y') \right]_{x',y'} \right) \tag{A.7}$$

where $\mathrm{VAR}([\cdot], [\cdot]) = \langle[\cdot][\cdot]\rangle - \langle[\cdot]\rangle\langle[\cdot]\rangle$. Variations of the replica measure $\langle\cdot\rangle$, equation (10), with respect to ζ_{s} are very small. That is, we neglect derivatives of the measure $\langle\cdot\rangle$ with respect to ζ_{s}, for example

$$\mathcal{F}_1'(0,0) \approx \left[\left\langle \exp\left(-\sum_{a=1}^{n} h(f_a(x), y) \right) \prod_{b=1}^{2} \mathcal{L}_b(x, y) \right\rangle \right]_{x,y} \tag{A.8}$$

and $\mathcal{F}_1''(0,0) \approx 0$. This yields from equations (A.2) and (A.3) the final results (41) and (43). We used the exact relations $\mathcal{G}'(\zeta_{\mathrm{s}}) = 0$, $\lim_{n\to 0} \mathcal{G}''(\zeta_{\mathrm{s}}) = m/\zeta_{\mathrm{s}}^2$, $\lim_{n\to 0} \zeta_{\mathrm{s}} = m$, and approximated the replica measure $\langle\cdot\rangle$ by $\langle\cdot\rangle_0$, equation (17).

Appendix B. Posterior mean and covariance of GP models for a fixed training set D

In this section we derive formulae for the mean prediction and prediction uncertainty of GP models trained on a *specific* data set D.

We define $f_i \equiv f(x_i)$ for the Gaussian process fields, where $i = 1, \dots, m$ denotes training inputs and x_{t} denotes a test input. We are interested in Gaussian averages of functions of the vector $\mathbf{f}_+ = (f_1, \dots, f_m, f_{\mathrm{t}})$, e.g., the predictions $\hat{\mathbf{f}}_+ = (\hat{f}_1, \dots, \hat{f}_m, \hat{f}_{\mathrm{t}})$, which are given by the posterior means

$$\hat{\mathbf{f}}_+ \doteq \langle \mathbf{f}_+ \rangle = \frac{1}{Z(\mathbf{y})} \int \mathrm{d}\mathbf{f}_+ \, \mathbf{f}_+ \mu[\mathbf{f}_+] p(\mathbf{y}|\mathbf{f}). \tag{B.1}$$

Since the likelihood $p(\mathbf{y}|\mathbf{f})$ depends only on the values of the field at the training points $\mathbf{f} = (f_1, \dots, f_m)$ we have replaced the infinite dimensional Gaussian process measure $\mu[f]$ by the joint Gaussian distribution $\mu[\mathbf{f}_+]$

$$\mu[\mathbf{f}_+] = \frac{\exp\left(-\frac{1}{2}\epsilon^{-1}\mathbf{f}_+^T \mathbf{K}_+^{-1} \mathbf{f}_+ \right)}{\sqrt{\det(2\pi\epsilon\mathbf{K}_+)}} \tag{B.2}$$

where $\mathbf{K}_+^{-1}$ is the inverse to the kernel matrix $K_{ij} = K(x_i, x_j)$ on the $m+1$ inputs. The parameter ϵ has been introduced to treat the SVM model, for which we have to discuss the asymptotic behaviour of quantities as $\epsilon \to 0$.

We will use the following relation, which is easily proved using integration by parts and the identity $f_i = \sum_j \epsilon K_{ij}(\sum_k \epsilon^{-1} K_{jk}^{-1} f_k)$. Let A be a function of $\mathbf{f}_+$. Then for any

$i \in \{1, \ldots, m, t\}$ we have

$$\langle f_i A(\mathbf{f}_+) \rangle_\mu = \epsilon \sum_j \left\langle \frac{\partial A(\mathbf{f}_+)}{\partial f_j} \right\rangle_\mu K_{ij} \tag{B.3}$$

where the averages are over the prior density $\mu[\mathbf{f}_+]$ (equation (B.2)). This yields immediately the *exact* relation

$$\hat{f}(x_i) = \sum_{j=1}^m K(x_i, x_j) y_j \beta_j \tag{B.4}$$

for all $j = 1, \ldots, m$ in terms of the 'embedding strength'

$$y_j \beta_j = \epsilon \int \frac{\mathrm{d}\mathbf{f} \mu[\mathbf{f}]}{Z(\mathbf{y})} \frac{\partial p(\mathbf{y}|\mathbf{f})}{\partial f_j} = -\epsilon \langle h'(f_j, y_j) \rangle, \tag{B.5}$$

where in the last equality we have set $p(\mathbf{y}|\mathbf{f}) = \prod_{k=1}^m \mathrm{e}^{-h(f_k, y_k)}$ and the average is over the posterior density $\mu_m[\mathbf{f}] = (1/Z(\mathbf{y}))\mu[\mathbf{f}]p(\mathbf{y}|\mathbf{f})$. An alternative expression is obtained by introducing external fields $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ which are added to $\mathbf{f}$ in the likelihood term

$$y_j \beta_j = \epsilon \frac{\partial \ln \int \mathrm{d}\mathbf{f} \mu[\mathbf{f}] p(\mathbf{y}|\mathbf{f} + \boldsymbol{\xi})}{\partial \xi_j} \bigg|_{\boldsymbol{\xi}=0}. \tag{B.6}$$

In a similar fashion we can compute the predictive uncertainty, i.e., the covariance $\langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle$, of the model at any (test or training) inputs x_i and x_j. We obtain for the second moments

$$\langle f_i f_j \rangle = \epsilon K(x_i, x_j) + \epsilon^2 \sum_{k,l=1}^m K(x_i, x_k) K(x_j, x_l) \int \frac{\mathrm{d}\mathbf{f} \mu[\mathbf{f}]}{Z(\mathbf{y})} \frac{\partial^2 p(\mathbf{y}|\mathbf{f})}{\partial f_k \partial f_l}. \tag{B.7}$$

We finally end up with the linear response expression

$$\langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle = \epsilon K(x_i, x_j) + \epsilon^2 \sum_{k,l=1}^m K(x_i, x_k) K(x_j, x_l)$$

$$\times \left(\frac{\partial^2}{\partial \xi_k \partial \xi_l} \ln \int \mathrm{d}\mathbf{f} \mu[\mathbf{f}] p(\mathbf{y}|\mathbf{f} + \boldsymbol{\xi}) \right) \bigg|_{\boldsymbol{\xi}=0}. \tag{B.8}$$

Appendix B.1. The Gaussian process regression model

The likelihood for the training data is

$$p(\mathbf{y}|\mathbf{f}) = \prod_{j=1}^m \exp\left(-\frac{1}{2\sigma^2}(f_j - y_j)^2 \right) \tag{B.9}$$

with real valued targets $y_i \in \mathbb{R}$ and $\epsilon = 1$. We can compute the 'embedding strength' (B.5) and prediction uncertainty (B.8) analytically. This yields the known results [14]

$$y_i \beta_i = \sum_{j=1}^m \mathbf{C}_{ij}^{-1} y_j \tag{B.10}$$

$$\langle f(x) f(x') \rangle - \langle f(x) \rangle \langle f(x') \rangle = K(x, x') - \sum_{i,j=1}^m K(x, x_i) K(x', x_j) \mathbf{C}_{ij}^{-1}. \tag{B.11}$$

Matrix $\mathbf{C}$ has elements $C_{ij} = K(x_i, x_j) + \sigma^2 \delta_{ij}$.

Appendix B.2. Hard margin support vector classification

The likelihood for the training data is

$$p(\mathbf{y}|\mathbf{f}) = \prod_{j=1}^{m} \Theta(y_j f_j - 1) \tag{B.12}$$

with binary targets $y_i = \pm 1$. For $\epsilon \to 0$, the posterior density is infinitely peaked and we recover from the definition (B.5) the well known fact that $\alpha_i \doteq \beta_i y_i$ equals zero unless $y_i f_i = 1$, i.e., unless data point i is a *support vector*. One can show [7] that the $\alpha_i \geq 0$ are solutions to the quadratic optimization problem

$$\min_{\{\alpha_j\} \geq 0} \left(\tfrac{1}{2} \sum_{i,j}^{m} \alpha_i y_i K(x_i, x_j) y_j \alpha_j - \sum_{j}^{m} \alpha_j \right). \tag{B.13}$$

We will not discuss the derivation of this result from the $\epsilon \to 0$ limit of equation (B.5) but refer the reader to [19]. Equation (B.13) defines a convex optimization problem, i.e., every local solution is a global one. It can, for example, be solved iteratively using the AdaTron algorithm [24].

We further note that the problem of perfectly separating the training data into the desired classes using an SVM is always solvable if the selected correlation kernel has a sufficient modelling complexity. For example, polynomial kernels $K(x, x') = (xx'/d)^r$ allow us to model decision surfaces between the classes by polynomials up to order r. The RBF kernel $K(x, x') = \exp(-||x - x'||/l^2)$ has infinite modelling complexity, i.e., it always allows us to separate arbitrary data.

The linear response relations (B.6) and (B.8) can be used to derive an expression for the response function $\chi(x, x')$, equation (65), which can be directly computed from simulations of the SVM algorithm. By introducing external fields ξ_i within the quadratic programming problem (B.13) one can show that

$$\frac{\partial^2}{\partial \xi_k \partial \xi_l} \ln \int \mathrm{d}\mathbf{f} \mu[\mathbf{f}] p(\mathbf{y}|\mathbf{f} + \boldsymbol{\xi}) = \epsilon^{-1} y_k \frac{\partial \alpha_k}{\partial \xi_l} = -\epsilon^{-1} (\mathbf{K}_{\mathrm{SV}}^{-1})_{lk}. \tag{B.14}$$

From this we get

$$\chi(x, x') = \lim_{\epsilon \to 0} \epsilon^{-1} [\langle f(x) f(x') \rangle - \langle f(x) \rangle \langle f(x') \rangle]_D$$

$$= K(x, x') - \left[\sum_{i,j \in \mathrm{SV}} K(x, x_i) K(x_j, x') (\mathbf{K}_{\mathrm{SV}}^{-1})_{ij} \right]_D \tag{B.15}$$

where $(\mathbf{K}_{\mathrm{SV}}^{-1})$ denotes the inverse of the kernel matrix restricted to the set of support vectors. The right-hand side of equation (B.15) is calculated from the SVM algorithm and the sum runs over the support vectors of the respective training data set D.

In a similar way it is possible to compute an approximate leave-one-out estimator $(1/m) \sum_{i=1}^{m} \Theta(-y_i \langle f_i \rangle_{\backslash i})$ analogous to equations (58) and (56). It can be shown that the expression (57) is equivalent to

$$\gamma_i = \frac{\partial \langle f_i \rangle / \partial \xi_i}{(\partial^2/\partial \xi_i^2) \ln \int \mathrm{d}\mathbf{f} \mu[\mathbf{f}] p(\mathbf{y}|\mathbf{f} + \boldsymbol{\xi})}. \tag{B.16}$$

Using equations (56) and (B.16), we obtain for hard margin support vector classification

$$\langle f_i \rangle_{\setminus i} = \langle f_i \rangle - \frac{y_i \alpha_i}{(\mathbf{K}_{\mathrm{SV}}^{-1})_{ii}}, \tag{B.17}$$

if $x_i \in \mathrm{SV}$ (which implies $\langle f_i \rangle = y_i$). Further, $\langle f_i \rangle_{\setminus i} = \langle f_i \rangle$ if x_i is not a support vector, i.e., if $y_i \langle f_i \rangle > 1$. This coincides with the result previously derived in [19] by using a different approach.

Appendix C. Variational equations for the hard margin support vector machine

Using the Gaussian trial distribution (17), we obtain the variational equations

$$\Delta \hat{Q}_\epsilon(x) = \frac{m}{\chi(x,x)} \left[\Phi\left(\Delta_{xy}\right) \right]_{y|x} \tag{C.1}$$

$$\hat{Q}_\epsilon(x) = -\Delta \hat{Q}_\epsilon(x) \frac{V(x,x)}{\chi(x,x)} - \frac{mV(x,x)}{\chi^2(x,x)} \left[\Delta_{xy} \int_{-\infty}^{\Delta_{xy}} \mathrm{D}v \left(\Delta_{xy} - v\right) \right]_{y|x} \tag{C.2}$$

$$\hat{R}_\epsilon(x) = -\Delta \hat{Q}_\epsilon(x) R(x) - \frac{m\sqrt{V(x,x)}}{\chi(x,x)} \left[y \int_{-\infty}^{\Delta_{xy}} \mathrm{D}v \left(\Delta_{x,y} - v\right) \right]_{y|x} \tag{C.3}$$

where $\Delta_{xy} = (1 - yR(x))/\sqrt{V(x,x)}$ and $\Phi(x) = \int_{-\infty}^{x} \mathrm{D}v$. $\Delta \hat{Q}_\epsilon(x)$, $\hat{Q}_\epsilon(x)$ and $\hat{R}_\epsilon(x)$ denote the value of the rescaled variational parameters $\epsilon \Delta \hat{Q}(x)$, $\epsilon^2 \hat{Q}(x)$ and $\epsilon \hat{R}(x)$ for $\epsilon \to 0$ and $\chi(x,x') = \lim_{\epsilon \to 0} \epsilon^{-1} G(x,x')$. Note that the variational parameter $\Delta \hat{Q}_\epsilon(x)$ (equation (C.1)) is proportional to the local probability of a test point x being a support vector (compare with equation (68)).

Appendix D. Wiener process regression: analytical study for $\sigma = 0$

The Wiener process regression model offers the exceptional opportunity to compare the replica variational approximation with *exact analytical* results for the free energy and the learning curve of the posterior variance at the point $\sigma = 0$ where the variational approximation fails to become asymptotically correct. In the following, we consider the Wiener process, which is a Gaussian Markov process defined by the correlation kernel $K(x,x') = \min(x,x')$ for non-negative inputs and initial condition $f(0) = 0$. Input data will be generated in the interval $[0,1]$ with the density $p(x) = 1$. Since the posterior variance for regression is independent of outputs y it can therefore be estimated from a simplified model where all y data are set to zero.

Appendix D.1. Exact results

The canonical partition function of a GP regression model is a functional integral which for $\sigma = 0$ reads

$$Z|_{\sigma^2=0} = \int \mathrm{d}\mu[f] \lim_{\sigma^2 \to 0} \frac{\exp\left(-\sum_{k=1}^{m} \frac{f^2(x_k)}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}^m} = \int \mathrm{d}\mu[f] \prod_{k=1}^{m} \delta(f(x_k)). \tag{D.1}$$

It can be rewritten as an m-dimensional integral over the probability distribution of the process on all m training data points,

$$Z|_{\sigma^2=0} = \int \prod_{i=1}^{m} \mathrm{d}f_i \, p(f_m = 0, f_{m-1} = 0, \ldots, f_1 = 0) \tag{D.2}$$

where $f_i \doteq f(x_i)$. The Wiener process is a *Markov process*, i.e., its joint probability density factorizes as $p(f_m, f_{m-1}, \ldots, f_1) = \prod_i p(f_{i+1}|f_i)p(f_1)$ where

$$P(f_{i+1}|f_i) = \frac{\exp\left(-\frac{1}{2}\frac{(f_{i+1}-f_i)^2}{(x_{i+1}-x_i)}\right)}{\sqrt{2\pi(x_{i+1}-x_i)}} \tag{D.3}$$

where we assume ordered inputs $x_{i+1} > x_i$. Introducing $x_0 \doteq 0$, this yields

$$\ln Z|_{\sigma^2=0} = \sum_{i=0}^{m-1}\left\{-\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}\ln(x_{i+1}-x_i)\right\}. \tag{D.4}$$

We can now perform the data average over the inputs by using the statistics of their differences $\Delta x = (x_{i+1} - x_i)$. For large m, it is well known that these have exponential density $p(\Delta x) \approx m\mathrm{e}^{-m\Delta x}$. This yields

$$[\ln Z]_D \simeq \frac{m}{2}\ln\left(\frac{m}{2\pi}\right) - \frac{m}{2}\int_0^m \mathrm{e}^{-y}\ln y \, \mathrm{d}y. \tag{D.5}$$

To calculate the average posterior variance $[G(x,x)]_x$, we note that due to the Markov property of the Wiener process, the posterior probability for a function value f_x at a position x is entirely determined by its bracketing data points $x_i < x < x_{i+1}$ from the training data set (compare with equation (D.3))

$$P(f(x)|f_m = 0, \ldots, f_1 = 0) = \frac{P(f(x)|f_i = 0)P(f_{i+1} = 0|f(x))}{P(f_{i+1} = 0|f_i = 0)} = \frac{\exp\left(-\frac{f^2(x)}{2\sigma^2(x)}\right)}{\sqrt{2\pi\sigma^2(x)}}, \tag{D.6}$$

i.e., the posterior probability is a Gaussian with variance $\sigma^2(x) = ((x_{i+1}-x)(x-x_i))/(x_{i+1}-x_i)$. Hence, we have to calculate

$$[G(x,x)]_x = \sum_i \int_{x_i}^{x_{i+1}} \mathrm{d}x \, \sigma^2(x) \simeq m\left[\int_{x_1}^{x_2} \mathrm{d}x \, \sigma^2(x)\right]_{\Delta x} \tag{D.7}$$

where $[\cdots]_{\Delta x}$ refers to the average with respect to the (asymptotically exponential) statistics of distances $\Delta x = x_2 - x_1$. Since $\int_{x_1}^{x_2}\sigma^2(x)\,\mathrm{d}x = (x_2-x_1)^2/6$ we find

$$[G(x,x)]_x \simeq \frac{m}{6}[(\Delta x)^2]_{\Delta x} \simeq \frac{1}{3m}. \tag{D.8}$$

Equations (D.5) and (D.8) are valid for sufficiently large m and can be compared to the results from the replica variational approach.

Appendix D.2. Comparison to the replica variational approximation

In the case where all y data are set to zero, replicas do not couple and $\hat{Q}(x) = 0$. The two variational equations (48) and (49) are trivially fulfilled with $R(x) = 0$ and $V(x, x) = 0$. Since the Wiener process is not a periodic process, all order parameters and variational parameters will depend explicitly on x. However, for large m, we can safely neglect this dependence and work with constant functions $\hat{Q}_0$ and $G \equiv [G(x,x)]_x$. The variational free energy simplifies to

$$-\frac{\partial \ln \Xi_n(m)}{\partial n} = -\ln \int \prod_a^n \mathrm{d}\mu[f_a] \exp\left(-\frac{\hat{Q}_0}{2}[f^2(x)]_x\right) - \frac{\hat{Q}_0 G}{2} + \frac{m}{2}\ln\left(1 + \frac{G}{\sigma^2}\right) \qquad (D.9)$$

with the remaining variational equation (47)

$$\hat{Q}_0 = \frac{m}{\sigma^2 + G}. \qquad (D.10)$$

In the following, we use the Karhunen–Loeve expansion (7) of the process with respect to the set of eigenfunctions on the interval $[0, 1]$ with eigenvalues λ_k. We get

$$[G(x,x)]_x = -2\frac{\partial}{\partial \hat{Q}_0}\ln \int \prod_a^n \mathrm{d}\mu[f_a] \exp\left(-\frac{\hat{Q}_0}{2}[f^2(x)]_x\right) = \frac{\partial}{\partial \hat{Q}_0}\ln \prod_k (1 + \hat{Q}_0 \lambda_k). \quad (D.11)$$

The eigenvalues of the Wiener process on the interval $[0, 1]$ are $\lambda_k = 1/\pi^2 \left(k - \frac{1}{2}\right)^2$ and $k = 1, 2, \ldots, \infty$. One gets $\sum_{k=1}^{\infty} \ln\left(1 + (\hat{Q}_0/\pi^2 (k - (1/2))^2)\right) = \ln \cosh \sqrt{\hat{Q}_0}$ and

$$[G(x,x)]_x = \frac{\tanh \sqrt{\hat{Q}_0}}{2\sqrt{\hat{Q}_0}}. \qquad (D.12)$$

The limit $\sigma \to 0$ is interesting due to the expected breakdown of the variational approximation for the GP regression model which is no longer asymptotically exact. Combining equation (D.10) and (D.12) yields for $\sigma^2 = 0$ and large m

$$\hat{Q}_0 \approx 4m^2 \qquad (D.13)$$

$$[G(x,x)]_x \approx \frac{1}{4m}. \qquad (D.14)$$

Comparison of equation (D.14) with (D.8) shows that the variational approach still gives the correct $1/m$ decay of the posterior variance but with a wrong prefactor.

In order to compare the variational free energy (D.9) with equation (D.5), we need to account for the normalization factor $1/\sqrt{2\pi\sigma^2}^m$ and take the limit $\sigma^2 \to 0$,

$$-\frac{\partial \mathcal{F}_n}{\partial n} - \frac{m}{2}\ln(2\pi\sigma^2) \to \frac{m}{2}\ln\left(\frac{m}{2\pi}\right) - \frac{m}{2}(1 - \ln 4) \qquad (D.15)$$

where we used equation (D.13) and (D.14). The difference between the true value of the free energy (D.5) and the variational free energy (D.15) amounts to $-(m/2)(1 - \ln 4 - \int_0^m e^{-y} \ln y \, \mathrm{d}y) \approx -0.0955m$, which is in very good agreement with the right panel of figure 3 (symbols: crosses and circles)!

31

Appendix E. Brief characterization of the used benchmark data

Abalone data, *Boston housing* data and *Robot arm* data are regression data sets. For Abalone data [20], the task is to predict the age of Abalone from physical measurements (input dimension $d = 10$, where we translated the gender encoding into binary inputs $M/F/I = 100/010/001$). For Boston housing data [20], the task is to predict house prices in the Boston, MA, area from census data ($d = 13$). For Robot arm data (Puma8nm) [21], the task is to predict the angular acceleration of one link of a Puma 560 robot arm. The $d = 8$ inputs are angular positions, velocities and torques of the robot arm.

Wisconsin breast cancer data, *Pima Indians diabetes* data and *Sonar* data are binary classification data sets. For Wisconsin breast cancer data [20], the task is to distinguish between benign and malignant forms of cancer based on $d = 9$ tissue characteristics. For Pima Indians diabetes data [22], the task is to detect diabetes or its absence from a set of $d = 7$ physical parameters. For Sonar data [20], the task is to distinguish between rocks and mines from sonar backscatter measured at $d = 60$ different angles.

References

[1] Engel A and Van den Broeck C P L, 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
[2] Nishimori H, 2001 *Statistical Physics of Spin Glasses and Information Processing* (Oxford: Oxford University Press)
[3] Mézard M, Parisi G and Virasoro M A, 1987 *Spin Glass Theory and Beyond* (*Lecture Notes in Physics* vol 9) (Singapore: World Scientific)
[4] Malzahn D and Opper M, *A variational approach to learning curves*, 2002 *Advances in Neural Information Processing Systems* vol 14, ed T G Dietterich, S Becker and Z Ghahramani (Cambridge, MA: MIT Press), http://www.nips.cc/NIPS2001/papers/#part3
[5] Malzahn D and Opper M, 2002 *Phys. Rev. Lett.* **89** 108302
[6] Boser B E, Guyon I M and Vapnik V M, *A training algorithm for optimal margin classifiers*, 1992 *Proc. 5th Annual ACM Workshop on Computational Learning Theory* (Pittsburgh, PA: ACM Press) pp 144–52, http://portal.acm.org/citation.cfm?id=130401
[7] Vapnik V, 1995 *The Nature of Statistical Learning Theory* (Berlin: Springer)
[8] Bartlett P J, Schölkopf B, Schuurmanns D and Smola A J (ed), 2000 *Advances in Large-Margin Classifiers* (Cambridge, MA: MIT Press)
[9] Cristianini N and Shawe-Taylor J, 2000 *Support Vector Machines* (Cambridge: Cambridge University Press)
[10] Dietrich R, Opper M and Sompolinsky H, 1999 *Phys. Rev. Lett.* **82** 2975
[11] Opper M and Urbanczik R, 2001 *Phys. Rev. Lett.* **86** 4410
[12] Wahba G, 1990 *Splines Models for Observational Data* (*Series in Applied Mathematics* vol 59) (Philadelphia, PA: SIAM)
[13] Neal R, 1996 *Bayesian Learning for Neural Networks* (*Springer Lecture Notes in Statistics* vol 118) (Berlin: Springer)
[14] Williams C K I and Rasmussen C E, 1996 *Advances in Neural Information Processing Systems* vol 8, ed D S Touretzky, M C Mozer and M E Hasselmo (Cambridge, MA: MIT Press) p 514, http://books.nips.cc/nips08.html
[15] Bialek W, Callan C G and Strong S P, 1996 *Phys. Rev. Lett.* **77** 4693
[16] Papoulis A, 1991 *Probability, Random Variables, and Stochastic Processes* (New York: McGraw-Hill)
[17] Feynman R P and Hibbs A R, 1965 *Quantum Mechanics and Path Integrals* (New York: McGraw-Hill)
[18] Parisi G, 1988 *Statistical Field Theory* (New York: Addison-Wesley)
[19] Opper M and Winther O, 2000 *Neural Comput.* **12** 2655
[20] The UCI Repository of machine learning databases, 1998 University of California, Department of Information and Computer Science (Irvine, CA) http://www1.ics.uci.edu/~mlearn/MLRepository.html
[21] Data for Evaluating Learning in Valid Experiments (Delve), 1995 The University of Toronto, Department of Computer Science, http://www.cs.toronto.edu/~delve

[22] Ripley B D, 1996 *Pattern Recognition and Neural Networks* (Cambridge: Cambridge University Press), http://www.stats.ox.ac.uk/pub/PRNN/
[23] Gardner E, 1988 *J. Phys. A: Math. Gen.* **21** 257
[24] Anlauf J K and Biehl M, *Properties of an adaptive perceptron algorithm*, 1990 *Parallel Processing in Neural Systems and Computers* (Amsterdam: Elsevier)
[25] Opper M and Winther O, 2001 *Phys. Rev. Lett.* **86** 3695
[26] Opper M and Saad D (ed), 2001 *Advanced Mean Field Methods: Theory and Practice* (Cambridge, MA: MIT Press)
[27] Mezard M, Parisi G and Zecchina R, 2002 *Science* **297** 812