Statistical Mechanics of Learning: A Variational Approach for Real Data

Dörthe Malzahn and Manfred Opper
Neural Computing Research Group, School of Engineering and Applied Science,
Aston University, Birmingham B4 7ET, United Kingdom
(Dated: January 10, 2002)

Using a variational technique, we generalize the statistical physics approach of learning from random examples to make it applicable to real data. We demonstrate the validity and relevance of our method by computing approximate estimators for generalization errors that are based on training data alone.

PACS numbers: 84.35.+i,02.50.-r,05.20.-y,87.18.Sn

In recent years, methods of statistical physics have contributed important insights to the theory of learning from example data with neural networks and other learning machines (for recent reviews see e.g. [1, 2]). Such systems are often described by models with a large number of degrees of freedom which interact by a random energy function where the randomness in a learning problem is induced by the data. Exact solutions for the learning performance of a great variety of models have been obtained using tools which were developed for the physics of disordered materials, such as the replica trick [3]. All these case studies assume simple distributions of the data and give invaluable insights into the generic learning behavior. However, they do not (and did not intend to) describe real world learning experiments where one has the additional problem that the distribution of the data is often unknown and idealized assumptions about it seem to be too restrictive. It would be highly desirable if the statistical physics theory could provide practitioners with tools to predict and optimize the performance of a learning algorithm.

In this Letter, we will present a step forward in this direction. We combine the replica approach with a variational approximation which enables us to deal with more realistic distributions of the randomness. For models of supervised learning [1], we demonstrate that the theory can predict relations between experimentally measurable quantities even though details of the data distribution are unknown. As an application, we will derive approximate expressions for data averaged performance measures for state of the art learning algorithms and test them on real data sets. We hope that our approach will inspire similar research in related complex adaptive systems such as, for example, communication systems [2].

We demonstrate the basic idea of our approach on a typical learning scenario where we try to learn an unknown rule which maps inputs $x \in R^d$ to outputs y from a set D of m example data pairs (x_i, y_i), $i = 1, \ldots, m$. The possible rules are modeled by a class of real valued functions $f(x)$. To get a reasonable predictor $\hat{f}(x|D)$ on arbitrary novel inputs x, a popular learning strategy is to balance the goodness of fit on the training data measured by a training energy $E[f; D] = \sum_{i=1}^{m} h(f(x_i), y_i)$ with the prior knowledge about the complexity of rules. In a probabilistic, Bayesian approach to learning [1, 2], both ingredients are combined in a probability distribution

$$\mu_m[f] = \mu[f]\, e^{-E[f;D]}/Z_m \,, \tag{1}$$

which is of the form of a Gibbs equilibrium distribution in statistical physics. It assigns different weights to functions f of being responsible for the observed training examples D. $\mu[f]$ encodes the prior knowledge about the plausibility of different functions. Proper choices of $\mu[f]$ will penalize functions which fit the data well but are too complex to generalize properly on novel inputs x. Parametric models express f by a set of parameters w, which may , e.g., be the weights of a neural network. The distribution $\mu[f]$ is then induced by a distribution over w. Non-parametric models are obtained by assigning an a priori statistical weight $\mu[f]$ directly over the space of functions. Predictions for the output y on a novel input x can be made by suitably averaging the function value $f(x)$ weighted by the distribution (1), where in the course of learning, when more and more data are observed, typically $\mu_m[f]$ will become increasingly concentrated around its mean. For continuous outputs (regression), we predict $y = \hat{f}(x|D) = \langle f(x) \rangle$, and for binary classification problems, we predict $y = \text{sign}[\langle f(x) \rangle]$, where angle brackets denote averages over (1).

We will now specialize on models which are defined by Gaussian prior distributions $\mu[f]$ over functions. Assuming a zero mean, they are fully specified by the correlation *kernel* $K(x, x') \doteq \int d\mu[f]\, \{f(x) f(x')\}$ which must be supplied by the user. It encodes a priori assumptions about the typical variability of model functions f with the input x. Such *Gaussian process* models (GP) have attracted considerable attention in recent years as they represent a flexible and widely applicable concept [4, 5, 6, 7]. GP models can be understood as a limit of Bayesian feed-forward neural networks when the number of hidden units grows to infinity [5]. Non-probabilistic "kernel machines" such as the celebrated *support vector machines* (SVMs) [8] result from GP models by taking appropriate limits. GP also form a basis for field theoretic approaches to density estimation [7]. For all GP models, the mean predictions are expressed as a linear combination of kernel functions centered at the input data $\langle f(x) \rangle = \sum_{j=1}^{m} \alpha_j K(x, x_j)$, where the α_j's are certain Gibbs averages which are independent on the point x [5, 6, 9].

We study the typical learning performance of this kernel approach by averaging over different drawings of training data

sets D where all examples (x_i, y_i) are generated independently from the same distribution $p(y, x) = p(y|x)p(x)$. We use the replica approach for computing the averaged free energy $F = [-\ln Z_m]_D$ which serves as a generating function for useful data averages. Here, we denote data averages by square brackets. We get $F = -\lim_{n\to 0} \frac{\partial \ln [Z_m^n]_D}{\partial n}$. To facilitate subsequent calculations, we use a *grand canonical* formulation where the number of examples m is only fixed on average by a chemical potential μ. An elementary calculation which uses the independence of the data, yields the grand canonical partition function for the n times replicated system

$$\Xi_n(\mu) \doteq \sum_{m=0}^{\infty} \frac{e^{\mu m}}{m!} [Z_m^n]_D = \int \prod_{a=1}^{n} d\mu[f_a] \ \exp[-H] \quad (2)$$

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

$$\mathcal{H}(\{f_a\}, x) = -e^{\mu} \left[\exp\left\{ -\sum_{a=1}^{n} h(f_a(x), y) \right\} \right]_{(y|x)} . \quad (3)$$

Here, the square brackets $[\ldots]_x$ and $[\ldots]_{y|x}$ denote expectations with respect to the distribution of inputs x and with respect to the conditional distribution of outputs y given the inputs. The chemical potential μ is adjusted such that $m = \frac{\partial \ln \Xi_n(\mu)}{\partial \mu}$ for all n which gives the simple result $\mu = \ln m$ in the limit $n \to 0$. For sufficiently large m, we replace the sum over m in Eq. (2) by its dominating term and recover the original (canonical) free energy as $F = -\lim_{n\to 0} \frac{\partial \ln \Xi_n(\ln m)}{\partial n}$.

Rather than evaluating (2) for simple and artificial distributions $p(x, y)$ in the limit of high input dimensionality, we resort to a variational approximation which can adapt to practically relevant situations. A similar approach was found to be useful in studying low dimensional disordered systems such as polymers or interfaces in random media [10]. We approximate H by a trial replica Hamiltonian H_0 which minimizes the variational bound (see e.g. [11])

$$-\ln \Xi_n(\mu) \leq -\ln \int \prod_{a=1}^{n} d\mu[f_a] \ e^{-H_0} + \langle H - H_0 \rangle_0 \ . \quad (4)$$

The brackets $\langle \ldots \rangle_0$ denote an average with respect to the distribution induced by $\prod_{a=1}^{n} \mu[f_a] e^{-H_0}$. Equation (4) is the leading term in a systematic perturbation expansion of the free energy with respect to the difference $H - H_0$. For Gaussian measures $\mu[f]$, a local trial Hamiltonian of the form

$$H_0 = \left[\sum_{a \leq b} \hat{Q}_{ab}(x) f_a(x) f_b(x) + \sum_a \hat{R}_a(x) f_a(x) \right]_x \quad (5)$$

is an appropriate choice. The most important feature of (5) is the explicit dependence of the variational parameters $\hat{Q}_{ab}(x)$ and $\hat{R}_a(x)$ on the input variable x . This enables us to take the nonuniformity of realistic input densities into account. 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 (1) becomes increasingly concentrated around its mean for large m.

Expressing the averaged Hamiltonian $\langle H \rangle_0$ by local order parameter fields $R_a(x) = \langle f_a(x) \rangle_0$ and $Q_{ab}(x, x)$ [the latter being a special case of the two point function $Q_{ab}(x, x') = \langle f_a(x) f_b(x') \rangle_0$], a straightforward variation yields

$$\frac{d\langle \mathcal{H} \rangle_0}{dR_a(x)} = \hat{R}_a(x) \ ; \qquad \frac{d\langle \mathcal{H} \rangle_0}{dQ_{ab}(x, x)} = \hat{Q}_{ab}(x) \ . \quad (6)$$

Assuming replica symmetry, i.e., $R_a(x) = R(x)$, $Q_{ab}(x, x') = Q(x, x')$ for $a \neq b$ and $Q_{aa}(x, x') = Q_0(x, x')$, the order parameter fields have a simple physical meaning in terms of averages over example data sets D. We have $Q(x, x') = [\langle f(x) \rangle \langle f(x') \rangle]_D$ and $Q_0(x, x') = [\langle f(x) f(x') \rangle]_D$. $G(x, x') = [\langle f(x) f(x') \rangle - \langle f(x) \rangle \langle f(x') \rangle]_D$ is the average posterior correlation function. $R(x) = [\langle f(x) \rangle]_D$ is the bias of the predictor, whereas $V(x, x') = Q(x, x') - R(x)R(x')$ is its data covariance. A direct calculation of order parameters from (5) expresses these in terms of the variational parameters as $R(x) = -[G(x, x')\hat{R}(x')]_{x'}$ and $V(x, x') = -[G(x, x'')G(x', x'')\hat{Q}(x'')]_{x''}$. Finally, G is found as the operator inverse $G = \left(K^{-1} + u\right)^{-1}$, where K is the kernel integral operator and $u(x, x') = p(x)(\hat{Q}_0(x) - \hat{Q}(x))\delta(x - x')$.

There are various ways to make use of our variational framework. We can use the probability measure defined by the trial Hamiltonian (5) in order to compute the average case performance of the learning algorithm at test points (x, y) not contained in the data set D. For example, the data average of the (mean square) prediction error $\varepsilon_2(D) = [(\langle f(x) \rangle - y)^2]_{x,y}$ is $[\varepsilon_2(D)]_D = [(R(x) - y)^2 + V(x, x)]_{x,y}$. Its sample fluctuations are $[\varepsilon_2(D)^2]_D - [\varepsilon(D)]_D^2 = [4(R(x) - y)(R(x') - y')V(x, x') + 2V^2(x, x')]_{x,x',y,y'}$. Similar to previous studies in the statistical mechanics of learning [1, 2], we could compute explicit learning curves from the variational equations, when the distribution of data is given. In general, explicit analytical solutions are possible for simple distributions, or in the asymptotic limit when the number of data grows large. We will give examples of such results elsewhere. Since in practice, however, data distributions are usually unknown, a more important application of our approach is in the derivation of explicit *relations* between data averaged observables, which will hold (within the framework of our approximation) for *any* such distribution. Such relations can help to estimate the performance of an algorithm on novel data which were not contained in the training set D.

To demonstrate this idea, we consider the problem of finding empirical estimates for generalization errors on novel test data (x, y). These estimates should be computable from the training data D only. Since in many real applications, test errors may be measured with an error function different from the original training energy h, we consider general error functions $L(\langle f(x) \rangle, x, y)$. Obviously, using the naive approximation $\frac{1}{m} \sum_{i=1}^{m} L(\langle f_i \rangle, x_i, y_i)$ to estimate the expected error $\varepsilon_L(D) = [L(\langle f(x) \rangle, x, y)]_{x,y}$, where $f_i \doteq f(x_i)$, will give an

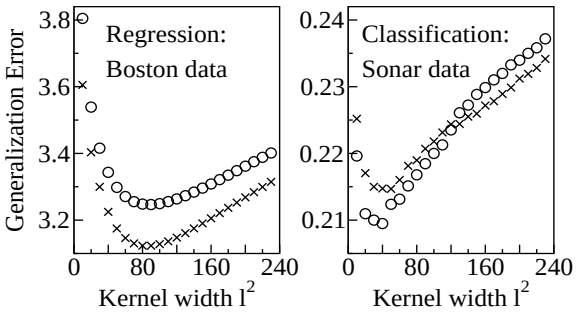

FIG. 1: Model selection based on Eq. (9) for a regression problem (Boston housing) and a binary classification problem (Sonar) using the kernel $K(x,x') = \exp(-||x-x'||^2/l^2)$. The right hand side of Eq. (9) is an empirical estimate (crosses) which is calculated only on *training* data. It gives a good account of the generalization error (circles) which was calculated on *test* data.

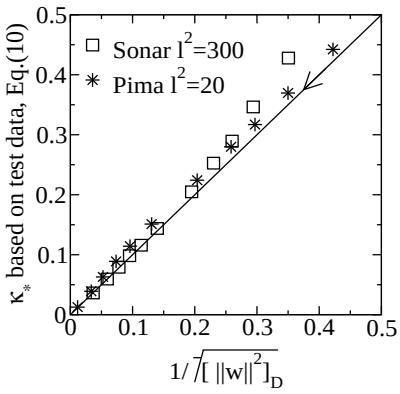

FIG. 2: Verification of Eq. (10) on two benchmark data sets for binary classification with input dimension $d = 7$ (Pima Indian diabetes, stars) and $d = 60$ (Sonar data, squares). The arrow points into the direction of increasing number m of training data with $m = 4-7$ in steps of 1, $m = 9-24$ in steps of 5 and $m = 33, 50, 100$.

optimistically biased estimate in most cases. In order to compute better, *unbiased* estimates for test errors which are based on the training data, we use the general result

$$\frac{1}{m}\left[\sum_{i=1}^{m} g(\langle F_1(f_i)\rangle,\ldots,\langle F_k(f_i)\rangle, x_i, y_i)\right]_D = \int Dz\,\left[g(\langle F_1(\phi)\rangle_\phi,\ldots,\langle F_k(\phi)\rangle_\phi, x, y)\right]_{x,y} \quad (7)$$

where $Dz = e^{-z^2/2}dz/\sqrt{2\pi}$ and $\langle\ldots\rangle_\phi$ denotes an expectation with respect to the distribution

$$P(\phi) = \frac{\exp\left[-h(\phi,y) - \frac{(R(x)+z\sqrt{V(x,x)}-\phi)^2}{2G(x,x)}\right]}{\sqrt{2\pi G(x,x)}\; Z(x,y,z)} \quad (8)$$

with norm $Z(x,y,z)$. Eq. (7) is easily proved within our variational framework and holds for arbitrary functions g, $F_1,\ldots,F_k$ [12]. To compute the desired unbiased estimates, we try to find a set of functions g and $F_1,\ldots F_k$ such that the right hand side of (7) can be rewritten as the data average of $\varepsilon_L(D)$.

We will demonstrate this idea for the GP model with mean square training error $h_2(f(x),y) = \frac{1}{2\sigma^2}(f(x)-y)^2$. This model finds widespread applications [4, 6] and has the advantage that the computations of Gibbs averages (1) such as means and correlation functions can be performed analytically in closed form. Nevertheless, the analysis of the data average is nontrivial, because averaging leads to a non Gaussian model in replica space which makes the application of our variational approximation still necessary. Using $F_1(f) = f$ and $F_2(f) = f^2$ in Eq. (7) as well as $[L(\langle f(x)\rangle, x, y)]_D = \int Dz\,[L(R(x)+z\sqrt{V(x,x)}, x, y)]_{x,y}$ yields

$$[\varepsilon_L(D)]_D = \frac{1}{m}\sum_{i=1}^{m}\left[L\left(\frac{\sigma^2\langle f_i\rangle - \sigma_i^2 y_i}{\sigma^2-\sigma_i^2}, x_i, y_i\right)\right]_D, \quad (9)$$

where $\sigma_i^2 = \langle f_i^2\rangle - \langle f_i\rangle^2$. Generalizations to other GP models and support vector machines are possible and will be given elsewhere. We present tests of Eq. (9) on regression and binary classification problems. In both cases $\langle f(x)\rangle$ is computed from the squared error h_2 with $\sigma^2 = 0.01$ (Boston data) and $\sigma^2 = 0.1$ (Sonar data), respectively. The left panel of Fig. 1 compares the generalization errors and their estimates for the publicly available *Boston housing* regression data set [13] using the error function $L(\langle f(x)\rangle, x, y) = |\langle f(x)\rangle - y|^k$ for $k = 1$, $m = 50$ and different widths l^2 of the kernel $K(x,x') = \exp(-||x-x'||^2/l^2)$. The results (which are averaged over 20 splits of the entire data set into training and test examples) suggest that our estimators might be well used for *model selection*, i.e., for finding the optimal kernel parameters with smallest generalization error. It is interesting to note that the case $k = 2$ leads to the exact *leave one out estimator* for the square error described in [4]. The right panel of Fig. 1 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.

In the following, 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 (5) is a good approximation. We consider SVM classifiers [8] which can be understood as generalizations of single layer neural networks which allow for *nonlinear* separation between classes. Expanding the positive definite SVM kernel $K(x,x') = \sum_k \psi_k(x)\psi_k(x')$ in a set of implicit features $\psi_k(x)$, the SVM output can be written as $y = \mathrm{sign}[f(x)]$ where $f(x) = \sum_k w_k\psi_k(x)$. The vector $\mathbf{w}$ of SVM weights w_i is determined by minimizing its length $||\mathbf{w}||$ under the condition that the "hard margin" training energy $h_{HM} = 0$. The latter is defined as $h_{HM}(f(x),y) = 0$ if $yf(x) \geq 1$ and $h_{HM} = \infty$ otherwise. This SVM model is contained within the present GP framework [9] by introducing a Gaus-

sian prior distribution of variance ϵ independently for each weight w_k. In the the limit $\epsilon \to 0$ the posterior Gibbs distribution becomes concentrated at the minimal length weight vector of the SVM. The Gaussian distribution over weights is equivalent to a Gaussian process over functions f with correlation kernel $K_\epsilon(x,x') = \epsilon K(x,x')$. We test our method on the SVM margin $\kappa(D) = 1/||\mathbf{w}||$ between positive and negative labeled examples which plays an important role as an indicator for good generalization ability of SVMs [8]. It can be shown that $\kappa(D)$ is related to the free energy as $1/\kappa_*^2 \doteq [1/\kappa^2(D)]_D = 2\lim_{\epsilon\to 0}\epsilon F_\epsilon$. Using our variational approximation we get

$$1/\kappa_*^2 = \left[\frac{m\sqrt{V(x,x)}}{\chi(x,x)} \int_{-\infty}^{\Delta(x,y)} Dz\,(\Delta(x,y)-z) \right]_{x,y} \qquad (10)$$

where $\Delta(x,y) = \frac{1-yR(x)}{\sqrt{V(x,x)}}$ and $\chi(x,x') = \lim_{\epsilon\to 0}\epsilon^{-1}G_\epsilon(x,x')$ is a response function which can be computed using the SVM algorithm. Equation (10) relates the averaged inverse squared margin on the training data to functions of the SVM's *bias*, its *variance* and the response function χ on novel test data (x,y). It generalizes Gardner's famous result [14] for the optimal stability of linear perceptrons to SVMs under general data distributions. Figure 2 compares κ_* as computed from its basic definition, with our prediction given by Eq. (10) for two publicly available classification data sets [13]. We used the kernel $K(x,x') = \exp(-||x-x'||^2/l^2)$ [15]. Figure 2 is obtained for suboptimal choices of the correlation length l and different sizes m of the training data sets D. For each m we have performed an average over 20 splits of the entire data set into training and test examples. Our variational Gaussian approximation improves with growing training size m and higher data dimensionality. The accuracy of our result (10) improves further, if the kernel width l^2 is optimized to achieve small generalization errors.

The two examples presented so far are only a small selection of the possible applications of our approach. Obvious future extensions will include an assessment of the reliability of estimators Eq. (9) for model selection by taking their sample fluctuations into account. Also alternative, Bayesian criteria which use the minimization of the free energy for model selection fit naturally into the statistical physics framework. Generalizing the variational approach to models with statistically *dependent* data such as e.g. (hidden) Markov processes, which are relevant for time series prediction, and to non Gaussian priors and trial Hamiltonians, will extend its applicability to the wider field of modern probabilistic data modeling. Finally, it will be necessary to estimate and improve the accuracy of our approximations. This can be done by computing perturbative corrections to the variational free energy (4), but also in a nonperturbative framework by including the possibility of *replica symmetry breaking*. While the latter phenomenon is not expected to be relevant for most of the interesting kernel models with convex error functions, applications of the method in other fields like combinatorial optimization might definitely benefit from this extension.

Acknowledgement: This work was supported by EPSRC grant GR/M81601.

[1] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[2] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing* (Oxford Science Publications, Oxford, 2001).

[3] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond*, Lecture Notes in Physics Vol. 9 (World Scientific, Singapore, 1987).

[4] G. Wahba, *Splines Models for Observational Data*, Series in Applied Mathematics Vol. 59 (SIAM, Philadelphia, 1990).

[5] R. Neal, *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics Vol. 118 (Springer, New York, 1996).

[6] C. K. I. Williams and C. E. Rasmussen, in *NIPS* edited by D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (MIT, Cambrigde, MA, 1996), Vol. 8, p. 514.

[7] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. **77**, 4693 (1996).

[8] See, for example, *Advances in Kernel Methods Support Vector Learning* edited by B. Schölkopf, C. J. C. Burges, and A. J. Smola (MIT, Cambridge, MA, 1999).

[9] M. Opper and O. Winther, Neural Computation **12**, 2655 (2000).

[10] M. Mézard and G. Parisi, J. Phys. I (France) **1**, 809 (1991).

[11] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (Mc Graw-Hill, New York, 1965).

[12] Similarly, one can compute sample fluctuations.

[13] The datasets can be downloaded from `http://www1.ics.uci.edu/~mlearn/MLSummary.html`.

[14] E. Gardner, J. Phys. A **21**, 257 (1988).

[15] One can model decision surfaces by polynomials of arbitrary order.