Bounds for Predictive Errors in the Statistical Mechanics of Supervised Learning

Manfred Opper

Institiut für Theoretische Physik III, Universität Würzburg, Germany

David Haussler

Department of Computer and Information Sciences, UC Santa Cruz, U.S.A.

Abstract

Within a Bayesian framework, by generalizing inequalities known from statistical mechanics, we calculate general upper and lower bounds for a cumulative entropic error, which measures the success in the supervised learning of an unknown rule from examples. Both bounds match asymptotically, when the number m of observed data grows large. We find that the information gain from observing a new example decreases universally like d/m. Here d is a dimension that is defined from the scaling of small volumes with respect to a distance in the space of rules.

PACS numbers: 87.10. +e, 05.90. +m, 89.70. +c

Understanding a neural network's ability to infer an unknown rule from a set of examples has become a fascinating topic in Statistical Mechanics [1, 2, 3]. Using techniques developed in the physics of disordered systems, exact learning curves, which measure the probability of a false prediction on new data, have been calculated for a variety of rules and network models in the last years. Broader reviews can be found e.g. in [4, 5, 6], cases of concrete learning problems are discussed in [7, 8, 9, 10]. These studies revealed a rich behaviour of learning curves, when the number of examples is sufficiently small relative to the number of parameters. Depending on the network's architecture, one can find different types of phase transitions, where the performance changes from poor generalization to a better understanding of the rule. On the other hand, when the number of examples grows large, and the network parameters assume continuous values, the results obtained for many models suggest that learning curves may have universal asymptotic features. For the important case where the rule can be implemented exactly by the network, the decay of the so called generalization error follows an inverse power law in the number of examples, with a constant (often called an "effective dimension"), that is proportional to the number of adjustable parameters. A similar scaling of a suitable error measure extends even to the case when the data are corrupted by a noise process [11].

To what extent can such a scaling behaviour be established in a general, rigorous framework, without making special assumptions like spherical input distributions or the thermodynamic limit? Although significant progress was made recently [4, 12, 13] in treating the asymptotics for certain classes of rules exactly, a proof for a general probabilistic rule has been lacking.

Using the replica method, the asymptotics of learning curves for the so called regular case, where the rule depends smoothly on a set of parameters, was calculated in [4]. A similar result for the scaling of the so called entropic error was given in [12] using asymptotic methods of statistics. These approaches fail in important nonregular cases, e.g., when the rule is given by a deterministic classification problem, or a nonsmooth noise model, where the outputs (classification labels) change discontinuously with the parameters. Using a different method, which is related to the idea of replicas, Amari [13] has performed an analysis of the asymptotics for deterministic binary classification problems. While this is a powerful new technique, it turns out to be more complicated and somewhat less general than the analysis of the regular models.

In the following, we present a new approach to this problem that covers both the regular and nonregular cases in a unified treatment. Our approach is not restricted to the asymptotics of the problem. Within a Bayesian [11, 14, 15] framework, by generalizing inequalities known from statistical mechanics, we derive general upper and lower bounds for a cumulative entropic error measure for a finite number m of examples. We find that both bounds match asymptotically for large m.

In the simplest approach, a rule can be specified by a functional relation $y = f_w(x)$ that maps an input x onto an output y. For a classification problem, we may think of x as vector of features of the object to be classified, and of y as the corresponding classification label. We will assume that f belongs to a parametric class of functions, and w denotes a vector of parameters. E.g., f_w may belong to the class of functions that are realizable by feedforward neural networks of a given architecture, where w specifies the set of network couplings and thresholds. In a more realistic approach, the deterministic relation between x and y should be replaced by a stochastic rule that is described by a probability $P(y|w,x)$. Our progress in learning about the true parameter will be measured by our ability to predict a new output y_{t+1} after having seen the previous observations $y^t = y_1, \ldots, y_t$ and $x^{t+1} = x_1, \ldots, x_{t+1}$. Assume that our estimate for the unknown rule w^* is given by the *predictive distribution* $\hat{P}_{t+1}(y_{t+1}|x_{t+1}) = \hat{P}(y_{t+1}|x_{t+1}, x^t, y^t)$. Then, we can define an entropic error by

$$\Delta I_{t+1} = -\Big\langle \ln \hat{P}_{t+1}(y_{t+1}|x_{t+1}) - \ln P(y_{t+1}|w^*, x_{t+1}) \Big\rangle \qquad (1)$$

where the brackets denote an average over the distribution of previous examples x^t, y^t and the new example (x_{t+1}, y_{t+1}). We adopt a Bayesian approach [11, 14, 15] and assume that nature has drawn the true parameter w^* of the rule at random from a prior distribution. Hence we can measure the performance of any learning method that produces an estimated probability $\hat{P}_{t+1}(y_{t+1}|x_{t+1})$ for a new output by including an additional average over the prior in the definition (1). In this sense, one can show that the Bayes optimal choice for the predictive distribution is given by the posterior probability

$$\hat{P}_{t+1}(y_{t+1}|x_{t+1}) = P(y_{t+1}|x_{t+1}, x^t, y^t) = \frac{\int dw \; P(y^t|w, x^t) P(y_{t+1}|w, x_{t+1})}{\int dw \; P(y^t|w, x^t)}, \qquad (2)$$

where $\int dw$ denotes the expectation over the prior. Here and in the sequel we assume that the inputs x_t are independent and the outputs y_t are independent given w^* and x_t. If we assume that Bayes optimal method is applied sequentially each time a new example is observed, the *cumulative* entropic error on m examples is found to be

$$I_m = \sum_{t=1}^{m} \Delta I_t = \left\langle \ln \frac{P(y^m|w^*,x^m)}{\int dw \; P(y^m|w,x^m)} \right\rangle_{all} \tag{3}$$

$$= \left\langle \int dw^* \sum_{y^m} P(y^m|w^*,x^m) \ln \frac{P(y^m|w^*,x^m)}{\int dw \; P(y^m|w,x^m)} \right\rangle_{x^m} .$$

Here, the the subscript *all* means an average taken over all random variables, i.e. the outputs y^m for fixed w^* and x^m, the parameter w^* and the inputs x^m. A second interpretation of Eq.(3) comes from information theory. I_m is the (input) average of the so called mutual information [16] between the data y^m and the parameter w^*, given the inputs x^m. Hence, it tells us how much information we have gained about the unknown parameter w^* by observing the outputs y^m. Accordingly ΔI_t is the instantaneous information gain from the t–th example.

To motivate the bounds on I_m that are derived in the following, we discuss a special case for the mutual information that makes its relation to statistical mechanics evident. Let us assume for the moment, that the likelihood $P(y^m|w,x^m)$ of the data can be written in the form of a Boltzmann factor $P(y^m|w,x^m) \propto e^{-\beta E(w,y^m,x^m)}$, where $E(w,y^m,x^m)$ may be interpreted as a suitable measure for the goodness of fit of the data using a deterministic model and the temperature β^{-1} measures the strength of the noise in the data. In the context of learning, E is called the training energy. Then the mutual information can be written

$$I_m = \Big\langle - \ln \int dw\, e^{-\beta E(w,y^m,x^m)} - \beta\, E(w^*, y^m, x^m) \Big\rangle_{all}.$$

Apart from the trivial second term, a calculation of the mutual information becomes essentially equivalent to a calculation of the quenched average of a free energy in statistical mechanics.

Besides the replica method, two other approaches are well known in statistical mechanics to estimate the average free energy. These are the *annealed* and the *high temperature* approximations. Both are summarised in the rigorous inequalities

$$- \ln \int dw \Big\langle e^{-\beta E} \Big\rangle \leq - \Big\langle \ln \int dw\, e^{-\beta E} \Big\rangle \leq - \ln \int dw\, e^{-\beta \langle E \rangle}. \tag{4}$$

The lower bound gives the annealed free energy, and is derived from Jensen's inequality. On the other hand, the expression on the righthand side has been used to estimate the free energy in the high temperature limit, but the fact that it is a rigorous upper bound [17] on the quenched average appears not to be widely known. It can also be easily proved from Jensen's inequality by utilizing that $\ln \int dw\, e^{-\beta E}$ is convex in E. Both bounds have been recently applied to the stochastic complexity of learning a rule [18].

The lower bound is too weak for noisy or nonrealizable rules, and the upper bound applied to the deterministic case ($\beta = \infty$) yields the trivial result $I_m \leq \infty$. In order to derive expressions for I_m that are useful in general, new techniques are required.

While the bounds (4) hold for general distributions of data y^m and x^m, and are also true for fixed parameter w^*, the special type of expectation involved in the definition (3) allows for a better type of bound. Note that the same probabilities (both the prior and $P(y^m|w,x^m)$) that define the average also appear inside the logarithm to be averaged. This allows us to utilize the following simple inequality for the upper bound. For an arbitrary probability $Q(y^m|w,x^m)$ one has

$$\int dw^* \sum_{y^m} P(y^m|w^*,x^m) \ln \frac{P(y^m|w^*,x^m)}{\int dw\, P(y^m|w,x^m)} \leq \int dw^* \sum_{y^m} P(y^m|w^*,x^m) \ln \frac{P(y^m|w^*,x^m)}{\int dw\, Q(y^m|w,x^m)}. \tag{5}$$

This follows from the fact that the difference between the right and left expressions forms a relative entropy, and thus is always nonnegative [16]. If we define Q such that $Q(y^m|w^*,x^m) = \prod_{t=1}^m Q(y_t|w^*,x_t)$, then using the independence of the inputs x_t and conditional independence of the outputs y_t, we can combine (5) with the upper bound of (4) and get

$$I_m \leq -\int dw^* \ln \int dw\, e^{-mD_Q(w^*,w)}. \tag{6}$$

where $D_Q(w^*,w) = \left\langle \sum_y P(y|w^*,x) \ln \frac{P(y|w^*,x)}{Q(y|w,x)} \right\rangle_x$. Optimising with respect to Q, inequality (6) defines a variational method to bound I_m. Clearly, by setting $Q = P$, one will be faced with a diverging D_Q, when the ratio $\frac{P(y|w,x)}{P(y|w^*,x)}$ is not bounded (this occurs for deterministic classification rules where $P(y|w^*,x) \in \{0,1\}$). Such divergences can be avoided by slightly perturbing P. For the case where the output variable y assumes only $K < \infty$ different values, the ansatz $Q_m(y|w,x) = (1-\delta_m)P(y|w,x) + \delta_m K^{-1}$ is sufficient to express the relative entropy $D_Q(w^*,w)$ in terms of a bounded quantity, defined by $\Delta(w,w^*) =$

$\frac{1}{2}\left\langle \sum_y \left(\sqrt{P(y|w,x)} - \sqrt{P(y|w^*,x)}\right)^2 \right\rangle_x$, which is known in statistics as (the average of) the so-called squared *Hellinger distance.* For the case of deterministic rules, $\Delta(w, w^*)$ reduces to the standard 0-1 generalization error, i.e. the probability that w and w^* disagree on a new input. A relation between D_Q and Δ can be established from the elementary inequality $\sum_y P_*(y) \ln \frac{P_*(y)}{Q(y)} \leq \frac{1}{2} G\left(\min_y [\frac{Q(y)}{P_*(y)}]\right) \sum_y \left(\sqrt{P_*(y)} - \sqrt{Q(y)}\right)^2$, valid for any distributions P_* and Q, where $G(z) = 2\,\frac{z - \ln z - 1}{(1-\sqrt{z})^2}$. ¿From the fact that $G(z)$ is a decreasing function of z, and by using the triangular inequality, one can prove that $D_{Q_m}(w^*, w) \leq G(\delta_m/K)\left(\Delta(w,w^*) + 3\sqrt{\delta_m}\right)$. Hence, since $G(z)$ is proportional to $-\ln z$ for small z, choosing $\delta_m = m^{-4}$, it is possible to obtain the upper bound

$$I_m \leq -\int dw^* \ln \int dw\, e^{-m[c(m)\Delta(w,w^*)] + o(1)}, \tag{7}$$

where $c(m) = \mathcal{O}(ln(m))$ and $o(1) \rightarrow 0$ as $m \rightarrow \infty$. If on the other hand $\frac{P(y|w,x)}{P(y|w^*,x)}$ is bounded, we can chose $Q = P$ and a bound analogous to (7) is obtained, where now $c(m) \leq$ *const* and no $o(1)$ term is needed.

It is possible to obtain a lower bound that has a structure similar to the upper bound. This is, to our knowledge, new and is based on the following extremal property of the mutual information. Define

$$J_m(\lambda) = -\int dw^* \sum_{y^m} P(y^m|w^*, x^m) \ln \int dw \left(\frac{P(y^m|w, x^m)}{P(y^m|w^*, x^m)}\right)^\lambda. \tag{8}$$

One can show that $J_m(\lambda)$ is *maximal* at $\lambda = 1$, where it equals the mutual information. An application of this result with $\lambda = \frac{1}{2}$ yields

$$I_m = \langle J_m(1) \rangle_{x^m} \geq \left\langle J_m(\frac{1}{2}) \right\rangle_{x^m} \geq - \int dw^* \ln \int dw \, e^{m \ln(1-\Delta(w,w^*))} \geq - \int dw^* \ln \int dw \, e^{-m\Delta(w,w^*)}. \tag{9}$$

The second inequality is again derived from Jensen's inequality. For a deterministic rule, the second to the last expression coincides with the result of the annealed average.

Both bounds (7) and (9) have the form of a free energy for a system described by coordinates w, which is attracted towards w^* via a nonrandom potential $\Delta(w, w^*)$. The corresponding "temperature" that determines the fluctuations of w away from w^*, decreases inversely proportional (apart from logarithmic corrections) with the number of examples m. Hence, the dominant contributions to the integrals over w come from small neighbourhoods of w^* in the large m limit. Denote by $V_\epsilon(w^*)$ the volume (probability) of parameters w that satisfy $\Delta(w, w^*) \leq \epsilon$, measured with respect to the prior. One can show that when the *local dimension* $d(w^*)$ defined by $d(w^*) = \lim_{\epsilon \to 0} \ln V_\epsilon(w^*) / \ln \epsilon$ exists, one has

$$\lim_{m \to \infty} \frac{-\ln \int dw \, e^{-m\Delta(w,w^*)}}{\ln m} = d(w^*). \tag{10}$$

Assuming that the limit and the integration over w^* can be exchanged [19], we can use (10) together with (7,9) to obtain the asymptotic scaling $I_m \simeq d \ln(m)$ for $m \to \infty$, where $d = \int dw^* \, d(w^*)$. As a consequence, it can be shown that whenever the limit $\lim_{m \to \infty} m \, I_m$ exists, the asymptotic value of the information gain is given by

$$\Delta I_m \simeq \frac{d}{m} \tag{11}$$

(see [20].) The values for the dimension d will depend on the type of learning problem and can be calculated for several models. For a regular noise model, when the probability P depends smoothly on an N component parameter vector w, a quadratic expansion $\Delta(w, w^*) \simeq \sum_{i,j} C_{ij}(w_i - w_i^*)(w_j - w_j^*)$ is valid for w close to w^*. Inserted into (7,9), assuming a well behaved prior, this shows that $d = \frac{1}{2}N$. On the other hand, for deterministic, binary classification rules, using polar coordinates, other types of expansions for Δ around w^* are often possible [13]. Here, for fixed angle variables the scaling $\Delta(w, w^*) \simeq ||w - w^*||^\gamma$ with $\gamma = 1$ is a typical case, resulting in $d = N$. This type of behaviour was found for rules that are defined by simple neural networks, like the perceptron and the committee and parity machines [21, 22] with tree architecture, explaining the somewhat surprising result that the asymptotics of their learning curves did not depend on the number of hidden units (for recent results, see [23]). A different scaling with $\gamma = \frac{1}{2}$ is obtained in a special thermodynamic limit of the committee machine with tree architecture, N weights and h hidden units [21]. When $N \to \infty$, $h \to \infty$, but $\frac{h}{N} \to 0$, the dimension scales like $d \simeq 2N$. Further work has to show in which cases other types of universality classes will appear.

The above results also enable us to bound the error of the so called *Gibbs algorithm*. This algorithm, which has been extensively studied for various models using methods of statistical mechanics [4, 5, 6], chooses an estimate $\hat{w}$ for the unknown parameter w^* at random from the posterior distribution $p(w|y^m, x^m)$. This can be realized by a random walk in the space of parameters, which reduces to a well known Monte Carlo process in case where the posterior is of the Gibbs type $p(w|y^m, x^m) \propto e^{-\beta E(w,y^m,x^m)}$. We will be interested in the question of how close the estimate $\hat{w}$ comes to the true parameter w^* on average. Thus we define the Gibbs–error by

$$\varepsilon_m = \left\langle \frac{\int dw\, P(y^m|w, x^m)\Delta(w, w^*)}{\int dw\, P(y^m|w, x^m)} \right\rangle_{all} . \tag{12}$$

This definition of ε_m reduces to the average 0-1 generalization error of the Gibbs algorithm after m training examples in the case of a deterministic rule.

In [24], the inequality $\varepsilon_m \leq \frac{1}{2\ln 2}\Delta I_m$ was found to hold for any deterministic, binary rule. At present, for the general case of arbitrary P, we are able to prove the somewhat weaker result that $\varepsilon_m \leq 2\Delta I_m$. The first inequality combined with (11) shows that the generalization error for the Gibbs–algorithm is asymptotically upper bounded by $0.721\,\frac{d}{m}$ in the deterministic binary case. This value is close to the asymptotics $0.62\,\frac{d}{m}$, which was obtained for many special models using the replica trick. As an example for the second inequality, consider a model of independent output noise for a binary classification problem, which is defined by the output $y = \eta \cdot f_w(x) \in \{-1, +1\}$, where $\eta \in \{-1, +1\}$ is a noise variable [11]. Using our bound it can be shown that the Gibbs algorithm is able to predict the correct *noise free output* $f_{w^*}(x)$ with an average probability that is not larger than *const* $\frac{d}{m}$ asymptotically. Here, the *const* grows with the noise rate.

So far our results hold on *average* over the prior distribution on the rules. It will be interesting to see under which conditions these results will also hold pointwise for almost all rules (see recent work of Merhav and Feder [25] and Feder, Freund and Mansour [26]). It is also a challenge to extend our methods to the important case of learning unrealizable rules.

Acknowledgments

D. Haussler was supported by NSF grant IRI-9123692 and M. Opper by a Heisenberg fellowship of the DFG.

References

[1] D. Hansel and H. Sompolinsky, Europhys. Lett. 11, 687 (1990).

[2] G. Györgyi, Phys. Rev. Lett. 64, 2957 (1990).

[3] H. Sompolinsky, N. Tishby and H.S. Seung, Phys. Rev. Lett. 65, 1683 (1990).

[4] H. Seung, H. Sompolinsky, and N. Tishby; Physical Review A 45, 6056 (1992).

[5] T. L. H. Watkin, A. Rau and M. Biehl; Rev. Mod. Phys. 65, 499 (1993).

[6] M. Opper and W. Kinzel; *Statistical Mechanics of Generalization*, to appear in: *Physics of Neural Networks*, ed. by J. L. van Hemmen, E. Domany and K. Schulten, published by Springer Verlag.

[7] H. Sompolinsky and N. Tishby, Europhys. Lett. 13, 567 (1990).

[8] I. Kocher and R. Monasson, Int. J. of Neural Syst. 2, 115 (1991).

[9] A. Romeo, Phys. Rev. E47, 2162 (1993).

[10] G. Parisi and F. Slanina, Europhys. Lett. 17, 497 (1992).

[11] M. Opper and D. Haussler; contribution to the *4th ACM Conference on Computational Learning Theory (COLT91)* (Santa Cruz 1991); pages 75- 87, published by Morgan Kaufmann, San Mateo. Neural Computation 4, 604 (1992).

[12] S. Amari and N. Murata; Neural Computation 5, 140 (1993).

[13] S. Amari; Neural Networks 6, 161 (1993).

[14] M. Opper and D. Haussler; Phys. Rev. Lett. 66, 2677 (1991).

[15] D. Haussler and A. Barron, *How well do Bayes methods work for on-line prediction of* $\{+1, -1\}$ *values?*, Proceedings of the Third NEC Symposium on Computation and Cognition, SIAM (1992).

[16] T. Cover and Joy A. Thomas; *Elements of Information Theory*, Wiley Series in Telecommunications, New York, 1991.

[17] K. Symanzik, J.Math. Phys. 6, 1155 (1965).

[18] R. Meir and N. Merhav, *Stochastic Complexity of Learning Realizable and Unrealizable Rules*, Machine Learning 19(3), 241 (1995).

[19] For example, a sufficient condition, somewhat weaker than uniform convergence in w^*, under which this exchange can be justified, is the existence of a positive $c(w^*)$, with $\int dw^* \, c(w^*) < \infty$, such that for all ϵ smaller than some ϵ_0, $V_\epsilon(w^*) \geq \epsilon^{c(w^*)}$.

[20] D. Haussler and M. Opper, *8th ACM Conference on Computational Learning Theory (COLT95)* (Santa Cruz 1995); published by ACM Press.

[21] H. Schwarze, J. Hertz; Europhys. Lett. 20, 375 (1992).

[22] M. Opper; Phys. Rev. Lett. 72, 2113 (1994).

[23] B. Schottky, preprint.

[24] D. Haussler, M. Kearns, and R. Schapire; Machine Learning 14, 84 (1994).

[25] Merhav and Meir Feder, *A Strong Version of the Redundancy-Capacity Theorem of Universal Coding*, IEEE Trans. Inform. Theory 41(3), 714 (1995).

[26] M. Feder, Yoav Freund and Yishay Mansour, *Optimal Universal Learning and Prediction of Probabilistic Concepts*, to appear in Proc. IEEE Information Theory Conference, 1995; Published by IEEE.