

On the Stochastic Complexity of Learning Realizable and Unrealizable Rules

RONNY MEIR
NERI MERHAV

rmeir@ee.technion.ac.il

Department of Electrical Engineering, Technion, Haifa 32000, Israel

Editor: David Haussler

Abstract. The problem of learning from examples in an average case setting is considered. Focusing on the stochastic complexity, an information theoretic quantity measuring the minimal description length of the data given a class of models, we find rigorous upper and lower bounds for this quantity under various conditions. For realizable problems, where the model class used is sufficiently rich to represent the function giving rise to the examples, we find tight upper and lower bounds for the stochastic complexity. In this case, bounds on the prediction error follow immediately using the methods of Haussler et al. (1994a). For unrealizable learning we find a tight upper bound only in the case of learning within a space of finite VC dimension. Moreover, we show in the latter case that the optimal method for prediction may *not* be the same as that for data compression, even in the limit of an infinite amount of training data, although the two problems (i.e. prediction and compression) are asymptotically equivalent in the realizable case. This result may bear consequences for many of the widely used model selection methods.

Keywords: average case learning, stochastic complexity

1. Introduction

We consider in this paper the problem of learning a concept based on a finite set of examples, where learning is based on choosing hypotheses from some hypothesis class. Much of the recent work in the field of mathematical statistics (Vapnik, 1982) and computational learning theory (Natarajan, 1991) has been devoted to the study of worst case bounds on the expected error, where worst case refers to the worst possible choice of function to be learned and the worst possible distribution of training examples. More recently an elegant formulation of the problem of *average-case* learning and its connection to information theory has been made in (Haussler, et al., 1994a), where rather tight bounds for the expected error in terms of information theoretic quantities are obtained when certain assumptions about the function class to be learned are made. Most of the work cited above has explicitly assumed that the so-called Vapnik Chernovenkins dimension (denoted as VC dimension) of the space from which the learning hypotheses are chosen is finite, the bounds mentioned above becoming tight as the number of examples relative to the VC dimension becomes large.

One of the major limitations of the above mentioned results is that bounds are tight only in the limit where the sample size is very large, a typical drawback of the standard statistical analyses. It has become clear in recent years through the use of statistical physics methods that exact results for effectively finite sample sizes can be obtained in the so called *thermodynamic limit*, where the VC dimension of the hypothesis class is allowed to increase in such a way that the ratio between it and the number of examples is finite. This situation, reviewed for example in (Watkin, et al., 1993), has produced a plethora of types of behaviors

which are totally absent in the usual statistical analyses, where the VC dimension of the hypothesis class is finite. The major problem, however, with the statistical physics approach is that in deriving exact results for effectively finite sample sizes they have usually relied on a method known as the replica method (Mezard, et al., 1987) which is notoriously difficult to put on a rigorous basis. Furthermore, in many situations (particularly where the function to be learned is not realizable within the given hypothesis space) it has turned out that even the replica method itself may lead to effectively intractable computations due to the extreme complexity of the solution in these cases. Finally we comment that in claiming typicality of the results (meaning roughly that certain statements can be made with probability 1 in the thermodynamic limit) all the statistical physics results rely on the so-called assumption of self-averaging of extensive quantities (see section 6), having to do with the behavior of various random variables in the thermodynamic limit. This assumption is usually taken for granted, but has never in fact (to our knowledge) been established in the present context.

Our aim in this paper is to derive upper and lower bounds for various quantities which are of interest from a learning theoretic perspective, both in the usual case where the VC dimension is finite as well as in the thermodynamic limit scenario discussed above, where the VC dimension is allowed to increase without limit. It is an explicit goal of this paper to avoid the use of the mathematically problematic replica approach, replacing exact calculations by upper and lower bounds and showing under what conditions these bounds become tight.

The remainder of the paper is organized as follows. Section 2 is devoted to a definition and description of the learning scenario. In section 3 we then describe the (pseudo-) Bayesian framework within which our analysis takes place. The basic bounds studied are derived in section 4 in a general setting and specialized in section 5 to the case of hypothesis spaces of finite VC dimension, where the tightness of the bounds is established. Section 6 then considers the thermodynamic limit scenario where similar results are derived. We proceed then in section 7 to consider issues related to the optimal choice of a certain regularization parameter, and discuss the implications from the point of view of model selection. The problem of the convergence of the loss function to its expected value is considered in section 8 for the case of finite VC dimension. Finally, we summarize our results and mention several open questions in section 9.

2. Definitions and Models

The learning scenario considered in this paper is that of Bayesian decision theory. We follow here the notational conventions of (Haussler & Barron, 1993). Thus, let X , Y and A be sets, called the *instance*, *outcome* and *decision* spaces, respectively. In this paper we consider binary classification and thus take $Y = A = \pm 1$. Learning from examples takes place in this framework through the following steps.

- A function $f: X \rightarrow Y$ chosen and kept fixed thereafter. We assume $f \in \mathcal{F}$ and refer to \mathcal{F} as the target space.
- A sequence of i.i.d. inputs $\mathbf{x}_i \in X$, $i = 1, 2, \dots, n$, is chosen according to some (unknown) probability distribution $D(\mathbf{x})$.
- An outcome sequence is generated according to the target rule, by setting $y_i = f(\mathbf{x}_i)$.
- A ‘learner’ tries to find a hypothesis $h \in \mathcal{H}$ which ‘best’ explains the data. Here \mathcal{H} , the hypothesis space, is a known function space of some finite complexity. Nothing is assumed however about the space \mathcal{F} .

We stress that the learner only has available the data set $\mathbf{x}_1^n = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ and $y_1^n = \{y_1, y_2, \dots, y_n\}$ as well as the function class \mathcal{H} . No knowledge about f , \mathcal{F} or $D(\mathbf{x})$ are assumed. The only vague notion about the previous list concerns the exact meaning of the term ‘best explains the data’. There are at least two possible definitions of this notion.

Prediction. A test sample \mathbf{x}_{n+1} is presented. The objective is to find an h which predicts the label y_{n+1} with as low a probability of error as possible.

Compression. Here the objective of the learner is to provide as concise a description of the data as possible. This idea will be quantified shortly using the notion of stochastic complexity.

In principle there does *not* seem to be any direct connection between the two notions above. However, it turns out that in some cases the two notions can be strongly linked giving rise to a quantitative Occam’s razor principle. We focus in most of this paper on the compression objective, mentioning connections to the prediction objective in sections 7 and 9. We note that the problem addressed in this paper, namely learning with unknown target space \mathcal{F} , is often referred to in the computational learning theory literature as agnostic learning (see Kearns, et al., 1992, for example).

As a concrete example of learning in the above scenario we may consider \mathcal{H} to be a class of feedforward neural networks of limited complexity. The simplest such example would be the single layered perceptron (Rosenblatt, 1962) but in principle any architectural constraints can be assumed. The hypothesis class \mathcal{H} is parametrized (perhaps redundantly) by a weight vector \mathbf{w} . While neural networks have been shown to be universal function approximators (see for example Hornik, et al., 1989) and thus capable of approximating *any* function $f(\cdot)$ to arbitrary accuracy, the question arises as to the performance of limited complexity networks of this nature, especially when trained on finite size data sets.

In an interesting recent paper Haussler, et al. (1994a) have performed an extensive investigation of the average case performance of learning algorithms under the assumption that the hypothesis space \mathcal{H} and the target space \mathcal{F} are identical. In this paper we derive results for the realizable as well as the **unrealizable** problem where the hypothesis space \mathcal{H} is a proper subspace of the target space \mathcal{F} , namely $\mathcal{H} \subset \mathcal{F}$. Thus, even in the case where an infinite amount of data is available the expected error produced by any hypothesis in \mathcal{H} is non-zero. The motivation for our work is threefold: (i) The realizable case has been studied extensively and is relatively well understood, (ii) Since in many cases the target space is unknown, the unrealizable situation arises naturally, (iii) The statistical mechanics results which have been shown to give the correct results in the realizable case, are very difficult to extend to the unrealizable case (mainly due to effects of replica symmetry breaking). Thus, as far is known to the present authors no good bounds are available in the average case setting for this problem. Worst case bounds have been derived in the unrealizable setting by Vapnik (1982).

3. The Pseudo-Bayesian Framework

In order to describe mathematically the process of learning in a statistical framework we take a Bayesian point of view. According to this view one first assigns a *prior* probability $P_{\mathcal{H}}(h)$ to each hypothesis $h \in \mathcal{H}$. The process of learning can then be viewed as one of modifying the distribution of hypotheses h based on the data (y_1^n, \mathbf{x}_1^n) . Formally, the

well-known Bayes rule allows us to express this idea mathematically as

$$P_{\mathcal{H}}(h \mid y_1^n, \mathbf{x}_1^n) = \frac{P_{\mathcal{H}}(y_1^n \mid h, \mathbf{x}_1^n) P_{\mathcal{H}}(h \mid \mathbf{x}_1^n)}{P_{\mathcal{H}}(y_1^n \mid \mathbf{x}_1^n)}, \quad (1)$$

with

$$P_{\mathcal{H}}(y_1^n \mid \mathbf{x}_1^n) = \int dP_{\mathcal{H}}(h) P_{\mathcal{H}}(y_1^n \mid h, \mathbf{x}_1^n). \quad (2)$$

The integration in Eq. (2) is over the hypothesis space \mathcal{H} and is taken with respect to the prior probability $dP_{\mathcal{H}}(h)$. The function $P_{\mathcal{H}}(y_1^n \mid h, \mathbf{x}_1^n)$ appearing in the numerator of Eq. (1) is usually referred to as the *likelihood* function. The main reason we refer to the above framework as *pseudo*-Bayesian is that the hypothesis space \mathcal{H} is not necessarily equal to the target space \mathcal{F} . In particular it is possible that the learner assigns zero a-priori probability to the ‘true’ function f . The main motivation for this extension is twofold: (i) It is often the case in practical applications that no knowledge is available concerning the target space and thus some assumption must be made, which may very often be inadequate (namely $\mathcal{H} \subset \mathcal{F}$). (ii) The hypothesis space \mathcal{H} may be limited in complexity for computational reasons, since too large a space will be very computationally expensive to search.

Having clarified the reason for our terminology we focus now on the specific form of the likelihood function. First, following common wisdom the learner assumes that the data has been generated according to a Bernoulli process, namely each input \mathbf{x}_i is drawn at random according to $D(\cdot)$ and then labelled as y_i by an unknown function. Thus we have

$$P_{\mathcal{H}}(y_1^n \mid h, \mathbf{x}_1^n) = \prod_{i=1}^n P_{\mathcal{H}}(y_i \mid h, \mathbf{x}_i) \quad (\text{independence assumption}). \quad (3)$$

Since we do not wish to restrict ourselves to realizable problems, we introduce a loss function $\lambda(y, h(\mathbf{x}))$ measuring the loss incurred on predicting an incorrect label $h(\mathbf{x})$. Using this function we express the likelihood as (Levin, et al., 1990):

$$P_{\mathcal{H}}(y_i \mid h, \mathbf{x}_i) = \frac{e^{-\beta \lambda(y_i, h(\mathbf{x}_i))}}{z_{\beta}}; \quad z_{\beta} = \sum_{\{y_i\}} e^{-\beta \lambda(y_i, h(\mathbf{x}_i))}. \quad (4)$$

The parameter β appearing in Eq. (4) is a regularization parameter, which is usually referred to as the inverse temperature in the statistical physics literature. We note in passing that this particular form can be derived using maximum entropy principles. In the remainder of the paper we take $\lambda(y, h(\mathbf{x}))$ to be the 0-1 loss function, defined by $\lambda(y, h(\mathbf{x})) = 0$ if $h(\mathbf{x}) = y$ and unity otherwise. It is easy to see that in this case $z_{\beta} = 1 + e^{-\beta}$ independently of h .

We note that for the 0-1 loss function in the limit $\beta \rightarrow \infty$, the posterior distribution (1) is non-zero only for hypotheses h which minimize the *empirical error*

$$\Lambda(y_1^n, \mathbf{x}_1^n; h) \triangleq \sum_{i=1}^n \lambda(y_i, h(\mathbf{x}_i)), \quad (5)$$

measuring the number of misclassifications. Thus, this framework yields in the appropriate limit the well-known method of minimal empirical loss (Vapnik, 1982). Moreover, if the problem is realizable (i.e. the minimal empirical loss is zero) we recover the problem studied

in (Haussler, et al., 1994a). In this paper, however, we restrict ourselves to the case $\beta < \infty$ throughout.

Having induced a posterior probability distribution $P_{\mathcal{H}}(h \mid y_1^n, \mathbf{x}_1^n)$ on the hypothesis space, it is easy to compute the probability of a value y_{n+1} being assigned to a new input \mathbf{x}_{n+1} . This quantity is given simply by

$$P_{\mathcal{H}}(y_{n+1} \mid y_1^n, \mathbf{x}_1^n, \mathbf{x}_{n+1}) = \frac{P_{\mathcal{H}}(y_1^{n+1} \mid \mathbf{x}_1^{n+1})}{P_{\mathcal{H}}(y_1^n \mid \mathbf{x}_1^n)} = \frac{1}{z_{\beta}} \frac{\int dP_{\mathcal{H}}(h) \exp\{-\beta \Lambda(y_1^{n+1}, \mathbf{x}_1^{n+1}; h)\}}{\int dP_{\mathcal{H}}(h) \exp\{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)\}}. \quad (6)$$

It will also be convenient in what follows to define the *volume ratio*, $\chi(y_1^n, \mathbf{x}_1^n)$, through

$$\chi(y_1^{n+1}, \mathbf{x}_1^{n+1}) \triangleq P_{\mathcal{H}}(y_{n+1} \mid y_1^n, \mathbf{x}_1^n, \mathbf{x}_{n+1}), \quad (7)$$

and the *expected loss* through

$$\bar{\lambda}(f, h) \triangleq E_D[\lambda(f(\mathbf{x}), h(\mathbf{x}))], \quad (8)$$

where the expectation is taken with respect to the distribution $D(\mathbf{x})$.

4. Bounds on the Log-Loss Error Function

In order to quantify the loss incurred on forming predictions using the (possibly incorrect) predictive distribution $P_{\mathcal{H}}(y_{n+1} \mid y_1^n, \mathbf{x}_1^n, \mathbf{x}_{n+1})$, we consider in this section the log-loss error function, defined as

$$L(y_1^n, \mathbf{x}_1^n) = - \sum_{i=0}^{n-1} \log \chi(y_1^{i+1}, \mathbf{x}_1^{i+1}), \quad (9)$$

where the natural logarithm is assumed throughout the paper. Using Eq. (6) we note that the log-loss is given by the simple expression

$$L(y_1^n \mid \mathbf{x}_1^n) = -\log P_{\mathcal{H}}(y_1^n \mid \mathbf{x}_1^n) = -\log \int dP_{\mathcal{H}}(h) e^{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)} + n \log z_{\beta}, \quad (10)$$

where use has been made of the telescopic nature of the sequence $\chi(y_1^i, \mathbf{x}_1^i)$. It is interesting at this point to notice that the log-loss given by Eq. (10) is nothing but the stochastic complexity (Rissanen, 1986) of a class of models \mathcal{H} with respect to a prior distribution $P_{\mathcal{H}}(h)$ and data set y_1^n (see also Amari & Murata, 1993, for a recent contribution). This quantity can be interpreted as the shortest possible code length for the labels y_1^n (for a fixed input sequence \mathbf{x}_1^n) that can be achieved by the models in the class \mathcal{H} . We refer to L interchangeably as either the log-loss or the stochastic complexity. It is also useful to comment that the stochastic complexity is very closely related to the statistical mechanical free energy (see for example Meir & Fontanari, 1993).

Up to this point the inputs \mathbf{x}_1^n have been fixed. Let us check now what can be gained by averaging over the input distribution $D(\mathbf{x})$ as well. To do this we will first need the following lemma¹, which we refer to as the thermodynamic inequality due to its origin in statistical thermodynamics.

LEMMA 1. *The following inequality holds*

$$E_D \left[-\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)} \right] \leq -\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta E_D[\Lambda(y_1^n, \mathbf{x}_1^n; h)]}. \quad (11)$$

PROOF. Let us first assume that the vector \mathbf{x}_1^n can only take on a finite set of K values, $\{\mathbf{x}_\mu\}_{\mu=1}^K$, each occurring with probability p_μ and such that $\sum_{\mu=1}^K p_\mu = 1$. We then find that

$$E_D[\Lambda(y_1^n, \mathbf{x}_1^n; h)] = \sum_{\mu=1}^K p_\mu \Lambda_\mu(f, h),$$

where $\Lambda_\mu(f, h) = \Lambda(f(\mathbf{x}_\mu), \mathbf{x}_\mu; h)$. Thus, we can write the argument of the logarithm on the right-hand side of Eq. (11) as

$$\int dP_{\mathcal{H}}(h) e^{-\beta \sum_{\mu=1}^K p_\mu \Lambda_\mu(f, h)} = \int dP_{\mathcal{H}}(h) \prod_{\mu=1}^K e^{-\beta p_\mu \Lambda_\mu(f, h)}. \quad (12)$$

Now, for any sequence of positive random variables X_1, \dots, X_K we have Hölder's inequality (Hardy, 1952)

$$E \left(\prod_{\mu=1}^K X_\mu \right) \leq \prod_{\mu=1}^K (E X_\mu^{1/q_\mu})^{q_\mu} \quad \left(\sum_{\mu} q_\mu = 1 \right). \quad (13)$$

From (13) we immediately find upon setting $X_\mu = e^{-\beta p_\mu \Lambda_\mu(f, h)}$, $q_\mu = p_\mu$ and using (12) that

$$\begin{aligned} \log \int dP_{\mathcal{H}}(h) e^{-\beta E_D[\Lambda(y_1^n, \mathbf{x}_1^n; h)]} &= \log \int dP_{\mathcal{H}}(h) \prod_{\mu=1}^K e^{-\beta p_\mu \Lambda_\mu(f, h)} \\ &\leq \log \prod_{\mu=1}^K \left(\int dP_{\mathcal{H}}(h) e^{-\beta \Lambda_\mu(f, h)} \right)^{p_\mu} = \sum_{\mu=1}^K p_\mu \log \int dP_{\mathcal{H}}(h) e^{-\beta \Lambda_\mu(f, h)} \\ &= E_D \log \int dP_{\mathcal{H}}(h) e^{-\beta \Lambda(f, h)}. \end{aligned} \quad (14)$$

Multiplying both sides (14) by -1 we obtain the desired inequality (11). The proof for continuous probability densities follows by continuity arguments. \square

Keeping in mind the Definitions (5) and (8) we can express this inequality as

$$E_D \left[-\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)} \right] \leq -\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta n f \bar{\lambda}(f, h)}. \quad (15)$$

The bound (15) is general in that it makes no assumptions about the nature of the space \mathcal{H} . In fact, the expression appearing on the right hand-side of Eq. (15) is just the *high-temperature* free energy as derived previously in (Seung, et al., 1992). The point to note in the present case, however, is that here it is found to be an upper bound for all β , while the usual interpretation treats the right hand side of Eq. (15) as an approximation to the average stochastic complexity, valid for *small* β .

It is also useful at this stage to consider a lower bound to the stochastic complexity, which is easily derived using the so called *annealed approximation* based on Jensen's bound. This bound has often been used as a quick way to obtain useful qualitative results. As we show here however, the annealed approximation may lead to totally inadequate results in the case of learning unrealizable rules. This point has been observed by several authors (see for example Seung, et al., 1992, and Meir & Fontanari, 1992) but seems to have been ignored by many other workers. The lower bound is easily derived using the convexity of the logarithm function and Jensen's bound:

$$E_D \left[-\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)} \right] \geq -\frac{1}{n} \log \int dP_{\mathcal{H}}(h) E_D e^{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)}. \quad (16)$$

Finally, in order to compute many of the integrals appearing in this work we will repeatedly make use of a variant of the well-known Laplace method for the asymptotic evaluation of integrals. Since the particular conditions used in this paper are not the standard ones, we have included in Appendix B a proof of the following result, which we refer to as the extended Laplace method (see Amari, et al., 1992, for similar results). Consider a function $f(\mathbf{x})$ defined over $X \subset R^k$ and let $f(\mathbf{x})$ achieve its global minimum, f_{\min} , at some point \mathbf{x}_0 , for which

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \frac{|f(\mathbf{x}) - f(\mathbf{x}_0)|}{\|\mathbf{x} - \mathbf{x}_0\|^s} = c \quad (0 < c < \infty) \quad (17)$$

for some positive s , referred to as the index of continuity of f at \mathbf{x}_0 (in the usual case where a quadratic expansion around the minimum is legitimate we have $s = 2$). Under appropriate conditions, spelled out in detail in Appendix B, one can show that the following asymptotic result holds:

$$\int_X d\mathbf{x} g(\mathbf{x}) e^{-nf(\mathbf{x})} \sim \frac{2\pi^{k/2}}{s(cn)^{k/s}} \frac{\Gamma(\frac{k}{s})}{\Gamma(\frac{k}{2})} g(\mathbf{x}^*) e^{-nf_{\min}}, \quad (18)$$

where $a_n \sim b_n$ signifies that $a_n/b_n \rightarrow 1$ for $n \rightarrow \infty$.

5. Finite-Dimensional Hypothesis Space

Having established upper and lower bounds for the average stochastic complexity the question naturally arises as to whether they are *asymptotically tight*, namely whether they both approach the *same* limit as $n \rightarrow \infty$. In this section we assume that the hypothesis space \mathcal{H} is of finite dimension, taken to mean that it can be parameterized by a parameter-vector of finite dimension. We also assume that \mathcal{H} is of finite VC dimension, noting that the assumed finite-dimensionality of \mathcal{H} does *not* necessarily imply finite VC dimension (see Sontag, 1992, for a counter-example). Using these assumptions we can show that the asymptotic tightness of the bounds (15) and (16) occurs only under very special circumstances. In fact, using the result (18) it is easy to see that as $n \rightarrow \infty$ the right hand side of (15) converges to

$$\beta \lambda^*(f) = \beta \min_h \bar{\lambda}(f, h) \quad (19)$$

while the right hand side of Eq. (16) converges to

$$-\max_h \log E_D e^{-\beta \lambda(f(\mathbf{x}), h(\mathbf{x}))} = -\log \max_h E_D e^{-\beta \lambda(f(\mathbf{x}), h(\mathbf{x}))}. \quad (20)$$

It is clear that these two expressions are different in general except in the limit $\beta \rightarrow 0$.

It turns out in fact that in the case of the 0-1 loss-function considered in this paper, both the annealed and the thermodynamic bounds approach the same limit for *any* β as long as the problem is *realizable*. In particular we have:

LEMMA 2. *The upper bound (15) and the lower bound (16) are asymptotically tight in the case of a 0-1 loss function, if the problem is realizable, namely $\mathcal{H} = \mathcal{F}$.*

PROOF. For a 0-1 loss function λ one can verify the simple identity

$$e^{-\beta \lambda} = 1 - (1 - e^{-\beta})\lambda. \quad (21)$$

Using this expression in Eq. (20) we see that the annealed bound is asymptotically equal to

$$-\log[1 - (1 - e^{-\beta})\lambda^*(f)]. \quad (22)$$

It is immediately obvious that Eqs. (22) and (19) agree for any β only if $\lambda^*(f) = 0$ which corresponds to the realizable case. \square

In fact, it will become apparent shortly that both the annealed and the thermodynamic bounds give rise to the same asymptotic rate of convergence to their limiting value in the case of realizable rules. For the case of unrealizable rules (even for 0-1 loss) it is clear that the two bounds are asymptotically incompatible. Having established the gap between the two bounds in general, the question arises as to whether either of the bounds leads asymptotically to the correct value. As we see from the next theorem, the upper bound (15) is in fact asymptotically tight as long as the hypothesis class \mathcal{H} is of a finite VC dimension, whether the problem is realizable or not.

THEOREM 1. *Let \mathcal{H} be a class of binary valued decision functions of finite VC dimension. Then the upper bound (15) becomes an equality for $n \rightarrow \infty$. In particular we have*

$$\lim_{n \rightarrow \infty} E_D \left[-\frac{1}{n} \log P_{\mathcal{H}}(y_1^n | \mathbf{x}_1^n) \right] = \beta \min_h E_D [\lambda(f(\mathbf{x}), h(\mathbf{x}))] + \log z_{\beta}. \quad (23)$$

PROOF. From the obvious inequality $-\Lambda(y_1^n, \mathbf{x}_1^n; h) \leq -\min_h [\Lambda(y_1^n, \mathbf{x}_1^n; h)]$ it follows that

$$\begin{aligned} E_D \left[-\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)} \right] &\geq E_D \left[-\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta \min_h \Lambda(y_1^n, \mathbf{x}_1^n; h)} \right] \\ &= \beta E_D \left[\min_h \frac{1}{n} \Lambda(y_1^n, \mathbf{x}_1^n; h) \right], \end{aligned} \quad (24)$$

where we have used the fact that $\int dP_{\mathcal{H}}(h) = 1$. Combining this result with the bound (11) we have

$$\begin{aligned} \beta E_D \left[\min_h \frac{1}{n} \Lambda(y_1^n, \mathbf{x}_1^n; h) \right] &\leq E_D \left[-\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)} \right] \\ &\leq -\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta n E_D \lambda(f(\mathbf{x}), h(\mathbf{x}))} \end{aligned} \quad (25)$$

Keeping in mind that $\Lambda(y_1^n, \mathbf{x}_1^n; h) = \sum_i \lambda(f(\mathbf{x}_i), h(\mathbf{x}_i))$ and using the Eq. (18) to evaluate the integral on the right hand side of the equation, we find in the limit $n \rightarrow \infty$ that

$$\begin{aligned} \beta E_D \min_h \frac{1}{n} \sum_{i=1}^n \lambda(f(\mathbf{x}_i), h(\mathbf{x}_i)) &\leq E_D \left[-\frac{1}{n} \log \int dP_{\mathcal{H}}(h) e^{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)} \right] \\ &\leq \beta \min_h E_D[\lambda(f(\mathbf{x}), h(\mathbf{x}))] \end{aligned} \quad (26)$$

where terms of order $\log n/n$ have been suppressed on the right hand side of the equation. At this point we make use of the well known result of Vapnik and Chernovenkins (see Vapnik, 1982) stating that a sufficient condition for the uniform convergence of the empirical loss to the expected loss is that the VC dimension of \mathcal{H} be finite. Thus, the order of the operations on the left hand side of the equation can be interchanged demonstrating that the two sides of the equation become identical in the limit $n \rightarrow \infty$. \square

The above observations help to explain why many of the results obtained using the annealed approximation yield qualitatively the correct results obtained from the full quenched theory using the (non-rigorous) replica method. Moreover, the lack of asymptotic tightness of the annealed bound in the case of unrealizable rules helps explain why this approximation has usually yielded completely wrong results when compared with the exact replica results. Examples of both of these observations can be found in Seung, et al. (1992) and Meir & Fontanari (1992).

Having established the asymptotic tightness of the upper bound (15), we proceed now to establish the rate of convergence to its limiting value. Assume \mathcal{H} is finite-dimensional, parameterized by some vector \mathbf{w} of dimension k . It is of interest here to compute the asymptotic behavior of the stochastic complexity. Using Eqs. (15) and (18) we may directly proceed to evaluate the relevant integral in our case (slightly abusing the notation by identifying $\bar{\lambda}(f, h)$ and $\bar{\lambda}(f, \mathbf{w})$),

$$\begin{aligned} \int dP_{\mathcal{H}}(h) e^{-\beta n \bar{\lambda}(f, h)} &= \int d\mathbf{w} P(\mathbf{w}) e^{-\beta n \bar{\lambda}(f, \mathbf{w})} \\ &\sim \frac{2\pi^{k/2}}{s(\beta cn)^{k/s}} \frac{\Gamma(\frac{k}{s})}{\Gamma(\frac{k}{2})} P(\mathbf{w}^*) e^{-\beta n \min_{\mathbf{w}} \bar{\lambda}(f, \mathbf{w})}, \end{aligned} \quad (27)$$

where \mathbf{w}^* is the minimum point of $\bar{\lambda}(f, \mathbf{w})$, and we have assumed that

$$\bar{\lambda}(f, \mathbf{w}) = \bar{\lambda}(f, \mathbf{w}^*) + c \|\mathbf{w} - \mathbf{w}^*\|^s \quad (s > 0)$$

in the neighborhood of \mathbf{w}^* . Taking the logarithm of this equation, dividing by n and using Eq. (15) we obtain an upper bound valid for realizable as well as unrealizable problems,

$$E_D \left[-\frac{1}{n} \log P_{\mathcal{H}}(y_1^n | \mathbf{x}_1^n) \right] \leq \beta \min_{\mathbf{w}} \bar{\lambda}(f, \mathbf{w}) + \log z_{\beta} + \frac{k \log n}{s} + O\left(\frac{1}{n}\right). \quad (28)$$

In view of the annealed lower bound, Eq. (16), we can similarly compute a lower bound to the stochastic complexity. It is clear from the above derivation, that the only modification to the derivation of the upper bound, Eq. (28), is in the function appearing in the exponent of the integral (27). In view of the observations of the previous section concerning the asymptotic equality of the two bounds in the case of realizable rules, and making use of Eq. (21) together with $\log(1-x) \approx -x$ for small x , it is easy to see that in the realizable case the upper and lower bounds agree to order $\log n/n$. Thus we have:

THEOREM 2. *The stochastic complexity for the 0-1 loss function in the realizable case is asymptotically given by*

$$E_D \left[-\frac{1}{n} \log P_{\mathcal{H}}(y_1^n | \mathbf{x}_1^n) \right] = \log z_{\beta} + \frac{k \log n}{s} + \Theta\left(\frac{1}{n}\right), \quad (29)$$

where s is the index of continuity of $\bar{\lambda}(f, \mathbf{w})$ at $\mathbf{w}^* = \arg \min_{\mathbf{w}} \bar{\lambda}(f, \mathbf{w})$.

6. The Thermodynamic Limit

From the previous analysis we note that the extended Laplace method, Eq. (18), used to derive the asymptotic expression (29) was strictly permissible only if two conditions are met: (i) The dimension of the hypothesis space \mathcal{H} is finite, and (ii) The sample size n increases to infinity. An interesting question arises as to whether any useful bounds can be derived for *finite* sample size n . Bearing in mind the stochastic nature of the problem it is hard to believe that any useful results can be derived in the case where both d (the dimension of \mathcal{H}) and the sample size n are finite. However, recent work using ideas from statistical physics (see Watkin, et al., 1993, for a review) has focused on the so-called *thermodynamic limit* where both d and n are allowed to increase without limit, keeping their ratio fixed to some finite value α . Specifically we consider the limit

$$\alpha \triangleq \lim_{d \rightarrow \infty} \frac{n}{d} < \infty. \quad (30)$$

It is clear from this definition that α is a measure of the normalized sample size. For example, consider a simple perceptron with d inputs and weights trained on a sample of size $n = \alpha d$ with $0 < \alpha < \infty$.

While many results have been derived in recent years using the thermodynamic limit, most of these have relied either on the replica method or the annealed approximation. Unfortunately, the replica method is notoriously difficult to justify in a rigorous way, while the annealed approximation produces *incorrect* results even asymptotically in the unrealizable case as we have seen in the previous section. Very recently, Haussler, et al. (1994b), have been able to derive rigorous upper bounds on learning curves in the thermodynamic limit under the assumption that the hypothesis space is countable. It is our aim here to show

that similar bounds can be derived for the somewhat different problem studied in this paper. In fact, the discreteness of \mathcal{H} plays no role in our analysis.

Although the extended Laplace method cannot be used as it stands in this case, it turns out that under rather mild conditions we can transform the expression for the stochastic complexity into an expression for which extended Laplace integration is permissible in the thermodynamic limit. In order to demonstrate this point we make the following assumption.

ASSUMPTION 1. The following transformation holds

$$\int d\mathbf{w} P(\mathbf{w}) e^{-\beta n \bar{\lambda}(f, \mathbf{w})} = C_d \int d\mathbf{q} e^{-dG(f, \mathbf{q}) - \beta n \bar{\lambda}(f, \mathbf{q})}, \quad (31)$$

where C_d depends on d polynomially, \mathbf{q} is a k -dimensional vector with k finite and $G(f, \mathbf{q})$ is singular only at a finite number of points.

We can in fact view the transformation in Eq. (31) as a *coordinate transformation* from \mathbf{w} to \mathbf{q} , in which case the function $G(f, \mathbf{q})$ is related to the Jacobian of the transformation. The exact details of the transformation depend on the problem at hand through the specific form of the functions $G(f, \mathbf{q})$ and $\bar{\lambda}(f, \mathbf{q})$ assumed to be independent of d and n . Many examples where the specific form of the function $G(f, \mathbf{q})$ can be computed have been studied in the literature (see for example Seung, et al., 1992, and Watkin, et al., 1993) and as far as we know all comply with the assumption. We note that the computation of $G(f, \mathbf{q})$, referred to in the statistical physics community as the *entropy* function, can be done in a perfectly rigorous way and depends in no way on the replica method. In order to clarify the transformation in Eq. (31) we describe in Appendix A an explicit calculation for a simple realizable problem as well as an unrealizable case, where the exact asymptotic scaling of the stochastic complexity is derived. As a final comment we remark that the variable d appearing in (31) need *not* be the Euclidean dimension of the parameter vector \mathbf{w} , although this is the case in all cases we know of.

Using the Definition (30) above we may express the integral in Eq. (31) as

$$C_d \int d\mathbf{q} e^{-n[\beta \bar{\lambda}(f, \mathbf{q}) + \frac{1}{\alpha} G(f, \mathbf{q})]}. \quad (32)$$

Since \mathbf{q} is finite-dimensional, taking the limit $n \rightarrow \infty$ we can use Eq. (18) to evaluate the upper bound on the stochastic complexity (see Eq. (15)), obtaining the following result:

THEOREM 3. *Under the conditions of Assumption 1, the stochastic complexity is upper bounded in the thermodynamic limit as follows:*

$$\begin{aligned} \lim_{n, d \rightarrow \infty} E_D \left[-\frac{1}{n} \log P_{\mathcal{H}}(y_1^n | \mathbf{x}_1^n) \right] &\leq \lim_{n, d \rightarrow \infty} \left[-\frac{1}{n} \log \int d\mathbf{w} P(\mathbf{w}) e^{-\beta n \bar{\lambda}(f, \mathbf{w})} \right] + \log z_{\beta} \\ &= \min_{\mathbf{q}} \left[\beta \bar{\lambda}(f, \mathbf{q}) + \frac{1}{\alpha} G(f, \mathbf{q}) \right] + \log z_{\beta}. \end{aligned} \quad (33)$$

For $\alpha \rightarrow \infty$ the error term $\bar{\lambda}(f, \mathbf{q})$ becomes dominant and we obtain the same limit derived previously in Eq. (28) for the finite-dimensional case. For finite values of α , however, one must take into account the interplay between the error term $\bar{\lambda}(f, \mathbf{q})$ and the

entropy term $G(f, \mathbf{q})$ arising from the integration over parameter space. We note that the entropy term appearing in this equation is unique to the thermodynamic limit and does not appear in the finite dimensional case, since all volumes in that case are finite and thus contribute negligibly as compared to the error term $\bar{\lambda}$, in the limit of large sample size. We thus expect, and indeed find, that the asymptotics of the stochastic complexity can display a variety of different forms depending on the interplay between the loss-function $\bar{\lambda}(f, \mathbf{q})$ and the entropy term $G(f, \mathbf{q})$. For comparison we note that the bound in the finite-dimensional case depended on the space \mathcal{H} only through its dimension k .

A natural question that arises at this point is whether the upper bound (33) is tight for $\alpha \rightarrow \infty$. A very simple analysis shows that the asymptotic tightness of (33) holds in the case where the problem is realizable, as long as the conditions of Assumption 1 are met.

THEOREM 4. *If the conditions in Assumption 1 above hold and the problem is realizable, then the bound (33) is asymptotically tight for any loss function $\lambda(\cdot, \cdot)$, becoming an equality for $\alpha \rightarrow \infty$.*

PROOF. First let us assume that $G(f, \mathbf{q})$ possesses no singularities. Then for $\alpha \rightarrow \infty$ the upper bound of Eq. (33) becomes simply $\min_{\mathbf{q}} \bar{\lambda}(f, \mathbf{q}) + \log z_{\beta}$. Using an argument similar to that used in Theorem 1, we then have the following upper and lower bounds on the stochastic complexity, valid in the limit $\alpha \rightarrow \infty$:

$$\beta E_D \min_h \frac{1}{n} \Lambda(y_1^n, \mathbf{x}_1^n; h) \leq E_D \left[-\frac{1}{n} \log P_{\mathcal{H}}(y_1^n | \mathbf{x}_1^n) \right] - \log z_{\beta} \leq \beta \min_{\mathbf{q}} \bar{\lambda}(f, \mathbf{q}). \quad (34)$$

In the realizable case both the left-hand side and the right-hand sides of Eq. (34) are identically zero, thus establishing the claim, irrespective of what particular loss-function is used. In the case where $G(f, \mathbf{q})$ is singular at $\mathbf{q}^* = \arg \min \bar{\lambda}(f, \mathbf{q})$ we can always take a small δ such that $G(f, \mathbf{q}^* + \delta)$ is finite. Then previous argument still holds and an upper bound differing at most by the arbitrarily small term δ from the previous case is obtained. \square

Using Eq. (16) and very similar arguments to those used to derive the upper bound (33), we can show that in the case of a 0-1 loss function the annealed bound provides a tight lower bound in the case of realizable problems. As in the finite-dimensional case, the upper and lower bounds agree in the realizable case since $\min_{\mathbf{q}} \bar{\lambda}(f, \mathbf{q}) = 0$. It should be stressed however that for a general loss function (differing from the 0-1 loss function) the annealed bound does *not* provide a tight lower bound even for realizable problems, as was the case in the finite-dimensional problem studied in section 5.

While we have shown that in the finite-dimensional case the upper bound (15) is tight whether the problem is realizable or not, we have only been able to demonstrate the tightness of the upper bound in the thermodynamic limit in the case of realizable problems. Although Theorem 4 made no use of Vapnik and Chernovenkins' results on uniform convergence, which are not guaranteed to hold in the present setting, we can immediately see that the proof applies only to the realizable case in which the minimal empirical error is identically zero for *any* input sequence \mathbf{x}_1^n . The proof of the asymptotic tightness (or lack thereof) in the unrealizable case remains an open problem.

7. On the Optimal Temperature for Compression

As we have shown in the previous section the expected stochastic complexity can be bounded from above using Eq. (28) in the finite dimensional case and Eq. (33) in the thermodynamic limit. Since the temperature-like variable β is a free parameter, the question arises whether there is an optimal value which minimizes the stochastic complexity and thus the description length in the coding interpretation.

Considering first the finite-dimensional case, Eq. (28), and computing the value of β which minimizes the stochastic complexity we find to leading order that

$$\beta^*(n) \approx \log \left(\frac{1 - \lambda_{\min}}{\lambda_{\min}} \right) - \frac{b}{n}, \quad (35)$$

where $\lambda_{\min}(f) = \min_{\mathbf{w}} \bar{\lambda}(f, \mathbf{w})$ is the minimal value of the expected error and b is a constant. It is interesting to observe that for the realizable case $\beta^*(\infty) = \infty$ since $\lambda_{\min} = 0$ while $\beta^*(\infty)$ is finite for the unrealizable case. Observing that $\beta = \infty$ corresponds to minimizing the empirical loss defined in (5), it seems that in the unrealizable case the minimal description length is achieved by a strategy that does *not* aim at minimizing the empirical error even asymptotically. Since minimizing the empirical error is asymptotically equivalent to minimizing the expected error, our results imply that in the unrealizable case the notions of minimal description length and minimal expected error are obtained by different strategies, at least in the case of noise free learning considered here. This result should have implications for model selection schemes in cases where the hypothesis space is inadequate. Note that in the realizable case we find that the optimal temperature for compression is $\beta = \infty$ for any sufficiently large n (where the asymptotic expansion holds). It is interesting to note that Oppen & Haussler (1991) have obtained the optimal temperature in the case of learning a realizable rule where the labels y_i are corrupted by noise such that $y_i \rightarrow -y_i$ with probability $1 - \eta$. They find $\beta^* = \log((1 - \eta)/\eta)$, which is exactly the same asymptotic expression we obtained if we identify η and λ_{\min} .

Having discussed the finite-dimensional case we proceed now to the case where the dimension of \mathcal{H} is allowed to increase with n , the so called thermodynamic limit discussed in section 6. Starting from the bound (33) and minimizing it over β we find the rather simple result that the minimum is achieved for

$$\beta^* = \log \left(\frac{1 - \lambda^*}{\lambda^*} \right), \quad (36)$$

where $\lambda^* = \bar{\lambda}(f, \mathbf{q}^*)$, and \mathbf{q}^* is obtained from

$$\mathbf{q}^* = \arg \min_{\mathbf{q}} \left[H(\bar{\lambda}(f, \mathbf{q})) + \frac{1}{\alpha} G(f, \mathbf{q}) \right]. \quad (37)$$

Here $H(\bar{\lambda})$ is the binary entropy function given by

$$H(\bar{\lambda}) = -\bar{\lambda} \log \bar{\lambda} - (1 - \bar{\lambda}) \log(1 - \bar{\lambda}).$$

The comments made above about the behavior of β_{∞} in the realizable and unrealizable cases apply here as well.

8. On the Almost Sure Convergence of the Stochastic Complexity

Since the stochastic complexity depends explicitly on the realization of the random sequence \mathbf{x}_1^n , two basic questions arise: (i) Does the stochastic complexity converge to a limit? (ii) In the case that convergence can be guaranteed, is the convergence to a random or deterministic limit?

In order to show almost sure convergence to a *deterministic* limit, we limit ourselves in this section to the case where the space \mathcal{H} is *finite* dimensional and of finite VC dimension, bearing in mind the comments made in section 5 concerning the relationship between finite-dimensionality and finite VC dimension. Before addressing our particular problem we recall a basic result of Vapnik and Chernovenkins which we reproduce here for convenience (see Vapnik, 1982, for a proof).

THEOREM 5 (Vapnik 82). *Let $h(\mathbf{x})$ be a class of decision rules of bounded VC dimension d_{VC} , and let $v(h)$ be the frequency of errors computed from the sample (y_1^n, \mathbf{x}_1^n) . Then for any $\delta > 0$ with probability at least $1 - \delta$ one may assert that for all $n \geq d_{VC}$, and simultaneously for all rules $h \in \mathcal{H}$, the probability of erroneous classification, $P_{\mathcal{H}}(h)$, is within the limits*

$$v(h) - 2\sqrt{\frac{d_{VC} \left(\log \frac{2n}{d_{VC}} + 1 \right) - \log \frac{\delta}{9}}{n}} < P_{\mathcal{H}}(h) < v(h) + 2\sqrt{\frac{d_{VC} \left(\log \frac{2n}{d_{VC}} + 1 \right) - \log \frac{\delta}{9}}{n}} \quad (38)$$

With this theorem in hand we can now prove the almost sure convergence of the stochastic complexity. Before presenting the theorem we note that in the case of the 0-1 loss function the normalized empirical error $\frac{1}{n} \Lambda(y_1^n, \mathbf{x}_1^n; h)$ is simply the fraction of errors, while the expected error is simply $\bar{\lambda}(f, h)$.

THEOREM 6. *In the case of a 0-1 loss function, the random variable $-\frac{1}{n} \log P_{\mathcal{H}}(y_1^n | \mathbf{x}_1^n)$ converges almost surely to its mean value for $n \rightarrow \infty$ if the hypothesis space \mathcal{H} is finite-dimensional. In particular we have*

$$-\frac{1}{n} \log \int dP_{\mathcal{H}}(h) \exp\{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)\} \xrightarrow{a.s., -D} \beta \min_h \bar{\lambda}(f, h) \quad (39)$$

PROOF. Since the empirical error assumes only discrete values, it divides the space \mathcal{H} into subspaces characterized by the finite set of values $\frac{1}{n} \Lambda(y_1^n, \mathbf{x}_1^n; h) = \lambda$ where $\lambda = 0, \frac{1}{n}, \dots, 1$. We can thus decompose the integral of interest as follows

$$\int dP_{\mathcal{H}}(h) e^{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)} = \sum_{\lambda} P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n) e^{-\beta n \lambda}, \quad (40)$$

with λ assuming $n + 1$ discrete values as above. Here

$$P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n) = \int_{\{h: \frac{1}{n} \Lambda(y_1^n, \mathbf{x}_1^n; h) = \lambda\}} dP_{\mathcal{H}}(h), \quad (41)$$

is the measure of hypotheses $h \in \mathcal{H}$ with empirical error frequency λ . Now, using Eq. (40), the following upper and lower bounds are immediate:

$$\begin{aligned} \max_{\lambda} [P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n) e^{-\beta n \lambda}] &\leq \int dP_{\mathcal{H}}(h) \exp\{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)\} \\ &\leq (n+1) \max_{\lambda} [P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n) e^{-\beta n \lambda}] \end{aligned} \quad (42)$$

Taking logarithms, dividing by n and using $\max f(x) = -\min(-f(x))$ we obtain

$$\begin{aligned} \min_{\lambda} \left[-\frac{1}{n} \log P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n) + \beta \lambda \right] &= \frac{\log(n+1)}{n} \\ &\leq -\frac{1}{n} \log \int dP_{\mathcal{H}}(h) \exp\{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)\} \\ &\leq \min_{\lambda} \left[-\frac{1}{n} \log P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n) + \beta \lambda \right] \end{aligned} \quad (43)$$

Now, according to the results of Vapnik and Chernovenkins described in the previous theorem, the normalized empirical error, $\frac{1}{n} \sum_{i=1}^n \lambda(f(\mathbf{x}_i), h(\mathbf{x}_i))$ can be bounded with probability $1 - \delta$ for large n as follows

$$\bar{\lambda}(f, h) - a_{\delta} \sqrt{\frac{\log n}{n}} \leq \frac{1}{n} \sum_{i=1}^n \lambda(f(\mathbf{x}_i), h(\mathbf{x}_i)) \leq \bar{\lambda}(f, h) + a_{\delta} \sqrt{\frac{\log n}{n}},$$

where a_{δ} can be read off from the Theorem 2. From this result it is easy to see that with probability $1 - \delta$

$$\begin{aligned} \int_{\{h: \bar{\lambda}(f, h) = \lambda\}} dP_{\mathcal{H}}(h) - b_{\delta} \sqrt{\frac{\log n}{n}} &\leq P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n) \\ &\leq \int_{\{h: \bar{\lambda}(f, h) = \lambda\}} dP_{\mathcal{H}}(h) + b_{\delta} \sqrt{\frac{\log n}{n}}, \end{aligned} \quad (44)$$

where b_{δ} is a constant (independent of n). In any event we have found that with probability 1 as $n \rightarrow \infty$

$$P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n) = \int_{\{h: \bar{\lambda}(f, h) = \lambda\}} dP_{\mathcal{H}}(h) + O\left(\sqrt{\frac{\log n}{n}}\right) \quad (n \rightarrow \infty), \quad (45)$$

where the first term on the right hand side of the equation is a deterministic finite number independent of n . Taking logarithms of the inequality (44) and dividing by n we note that two cases may arise depending on whether $P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n)$ is zero or whether it can be bounded below by a constant $c > 0$. In the prior case the logarithm becomes infinite, and it is clear from (43) that the minimum over λ cannot be achieved. Assuming therefore that $P_{\mathcal{H}}(\lambda; y_1^n, \mathbf{x}_1^n)$ can be bounded below by a positive constant as long as $\lambda > \lambda_{\min}(f) = \min_h \lambda(f, h)$ it is easy to see from (43) that

$$-\frac{1}{n} \log \int dP_{\mathcal{H}}(h) \exp\{-\beta \Lambda(y_1^n, \mathbf{x}_1^n; h)\} \xrightarrow{a.s.-D} \beta \min_h \bar{\lambda}(f, h) \quad (46)$$

□

It is clear from the proof of Theorem 5 that the realizability or lack thereof did not enter the proof. It seems therefore that the distinction between these two problems in the finite-dimensional case is not of major importance, although it seems to play a major role in the thermodynamic limit scenario described in section 6.

9. Discussion

We have focused in this paper on bounding the stochastic complexity for learning binary valued classification problems, both for realizable as well as unrealizable problems. Considering the widespread use of the minimum description length principle in model selection methods, we have found it useful to focus our attention, following Rissanen (1986), on the shortest description length of the data, given a class of models. Our main results can be summarized as follows. First, we have shown that a widely used bound, namely the annealed bound of Eq. (16) becomes tight only under very stringent conditions, in particular for 0-1 loss functions and realizable problems. Thus, our results imply that use of this lower bound should be avoided unless its tightness can be established. Second, we have shown that for the case of realizable rules, the upper bound (15) is asymptotically tight and thus yields a consistent limiting behavior for all sample sizes. This result is true both for finite-dimensional hypothesis classes as well as for the thermodynamic limit scenario considered in section 6. The case of unrealizable learning, however, poses some difficulties. While we have shown (Theorem 1) that the upper bound (15) is asymptotically tight for finite-dimensional spaces we have not been able to demonstrate this in the thermodynamic limit framework for unrealizable problems. A further result of our work is related to the optimal choice of the regularization parameter β . As we have shown, in the case of learning unrealizable rules the best choice for β is given by $\beta = \log(1 - \lambda_{\min})/\lambda_{\min}$ where λ_{\min} is the minimal possible expected error within the hypothesis space \mathcal{H} . We have argued in this case that the method of minimizing the stochastic complexity and that of minimizing the expected error are incompatible, even asymptotically. This result is significant for model selection methods. Finally, we have shown that the stochastic-complexity approaches asymptotically a deterministic limit, a property often referred to as self-averaging in the statistical physics community (Mezard, et al., 1987). The more difficult problem of establishing self-averaging for the thermodynamic limit scenario remains open.

The most obvious inadequacy of our results concerns the analysis of unrealizable learning in the thermodynamic limit. We believe that an important step still needs to be made in understanding this situation. The work of Haussler, et al. (1994b) constitutes an important step in this direction, although limited to finite cardinality hypothesis spaces. A further inadequacy of our results is concerned with the relationship between compression and prediction alluded to in section 2. One of the most interesting questions one may ask is whether models which compress the data optimally are those which make the best predictions. This question becomes particularly important when minimum description length principles are used to select models from a *complexity-limited* class of functions, based on a finite data set. In the case of learning realizable rules, tight upper and lower bounds relating the prediction error to the stochastic complexity have been established in (Haussler, et al., 1994a), thus effectively answering the above question. While some of the results in (Haussler, et al., 1994a) can be used in the unrealizable case, they turn out not to be asymptotically tight. We have been able to improve on their bounds relating the prediction error to the stochastic

complexity for unrealizable rules (Meir & Merhav, 1994), but have unfortunately not been able to establish the asymptotic tightness of the bound. In view of the ambiguity of unrealizable problems and the importance of model selection criteria, we believe that establishing asymptotically tight bounds in this case is of paramount importance.

Appendix A

We show here, following the derivation in (Seung, et al., 1992), how the coordinate transformation in Eq. (31) may be derived for a simple model. Although this derivation appears in the statistical physics literature, we present it here for completeness. In particular, we assume the target function is given by a single-layer perceptron $f(\mathbf{x}) = \text{sgn}(\mathbf{w}_0^T \cdot \mathbf{x})$ with the same expression applying to the hypothesis h , namely $h(\mathbf{x}) = \text{sgn}(\mathbf{w}^T \cdot \mathbf{x})$. Since the signum function is unaffected by the normalization of its argument we fix the norms of the respective weight vectors, taking in particular $\|\mathbf{w}_0\|^2 = \|\mathbf{w}\|^2 = d$. The expected loss for any spherically symmetric distribution, $D(\mathbf{x})$, can easily be computed (Watkin, et al., 1993) and yields

$$\bar{\lambda}(\mathbf{w}_0, \mathbf{w}) = \frac{1}{\pi} \cos^{-1} \left(\frac{1}{d} \mathbf{w}_0^T \cdot \mathbf{w} \right). \quad (47)$$

From Eq. (31) we need to evaluate the following integral:

$$\mathcal{J} = \int d\mathbf{w} P(\mathbf{w}) e^{-\beta n \bar{\lambda}(\mathbf{w}_0, \mathbf{w})}. \quad (48)$$

In view of the normalization requirement $\|\mathbf{w}\|^2 = d$ we let $P(\mathbf{w})$ be a uniform distribution on the hypersphere of radius \sqrt{d} , given specifically as

$$P(\mathbf{w}) d\mathbf{w} = \delta \left(\sum_i w_i^2 - d \right) \prod_{i=1}^d \frac{dw_i}{\sqrt{2\pi e}}. \quad (49)$$

Introducing an integration variable R through the identity

$$1 = \int_{-1}^1 dR \delta \left(R - \frac{1}{d} \mathbf{w}_0^T \cdot \mathbf{w} \right), \quad (50)$$

and making use of the Fourier representation of the Dirac delta function,

$$\delta(x) = \int_{-i\infty}^{i\infty} \frac{d\hat{x}}{2\pi i} e^{\hat{x}x}, \quad (51)$$

we may transform the integral (48) into the following form:

$$\begin{aligned} \mathcal{J} &= \int_{-1}^1 dR \int_{-i\infty}^{i\infty} \frac{d\hat{R}}{2\pi i/d} \int_{-i\infty}^{i\infty} \frac{dE}{4\pi i} e^{-d\hat{R}R + dE/2 - \beta n \bar{\lambda}(R)} \\ &\quad \times \int \prod_{i=1}^d \left(\frac{dw_i}{\sqrt{2\pi e}} \right) e^{\frac{1}{2}E \sum_i w_i^2 + \hat{R} \sum_i w_i w_{0i}}. \end{aligned} \quad (52)$$

Note that the integration region in Eq. (50) is $R \in [-1, 1]$ since by the Schwarz inequality $|\frac{1}{d} \mathbf{w}_0^T \cdot \mathbf{w}| \leq 1$. The integrals over w_i factorize, and since they are simple Gaussian integrals may be calculated exactly, giving rise to the expression

$$\mathcal{J} = \int_{-1}^1 dR \int_{-\infty}^{\infty} \frac{d\hat{R}}{2\pi i/d} \int_{-\infty}^{\infty} \frac{dE}{4\pi i} e^{-d[\hat{R}R - \frac{1}{2}E + \alpha\beta \bar{\lambda}(R) - \hat{R}^2/2E + \frac{1}{2}\log(eE)]}. \quad (53)$$

Note that the dependence on the target vector \mathbf{w}_0 has dropped out from the calculation, by using the normalization requirement $\|\mathbf{w}_0\|^2 = d$. Now, in order to compute the 3-dimensional integral in Eq. (53) we note that in the limit $d \rightarrow \infty$ it takes the form of a saddle-point integral in the complex planes of \hat{R} and E , and can thus be computed by deforming the contours of integration so that they pass through the point where the argument of the exponent is minimal (de Bruijn, 1981). The minimum with respect to the variables \hat{R} and E can be obtained easily, and is given simply by $\hat{R} = R/(1 - R^2)$ and $E = 1/(1 - R^2)$, from which we are left with the simple one-dimensional integral

$$\mathcal{J} = C \int_{-1}^1 dR e^{-n[\beta \bar{\lambda}(R) - \frac{1}{2\alpha} \log(1 - R^2)]}, \quad (54)$$

where C is a constant which is independent of d in the present case. It is clear that this expression is of the correct form of Eq. (31) and fulfills the conditions in Assumption 1. Note that although the function $\log(1 - R^2)$ is singular, this singularity occurs at a single point, as stipulated. From these results one may proceed to compute the upper bound to the stochastic complexity, which in this case yields

$$E_D \left[-\frac{1}{n} \log P_{\mathcal{H}}(y_1^n | \mathbf{x}_1^n) \right] \leq \log z_{\beta} + \frac{\log \alpha}{\alpha} \quad (\alpha \rightarrow \infty), \quad (55)$$

a result previously derived by Györfi & Tishby (1990) (replacing the inequality by an equality sign) using the replica method. Although the above derivation can be obtained via a simple geometrical argument in the present case (Seung, et al., 1992), we have presented the full derivation here since the geometrical argument does not generalize to more complex situations.

As a second example we follow Watkin & Rau (1992) and consider a simple unrealizable problem, differing from the previous case in that the target function $f(\mathbf{x})$ is a perceptron with a threshold δ , namely $f(\mathbf{x}) = \text{sgn}(\mathbf{w}_0^T \cdot \mathbf{x} - \delta\sqrt{d})$. The hypothesis space again consists of simple perceptrons without the threshold term. Clearly the problem is unrealizable for $\delta \neq 0$. Taking $D(\mathbf{x})$ to be a multivariate normal distribution with zero mean and unit $d \times d$ covariance matrix, one finds

$$\bar{\lambda}(\mathbf{w}_0, \mathbf{w}) = \frac{1}{2} - \int_{|\delta|}^{\infty} Dx \operatorname{erf} \left(\frac{Rx}{\sqrt{2(1 - R^2)}} \right), \quad (56)$$

where $\operatorname{erf}(u) = \frac{2}{\sqrt{\pi}} \int_0^u dt e^{-t^2}$, Dx is the Gaussian measure $(e^{-x^2/2}/\sqrt{2\pi})dx$ and $R = \frac{1}{d} \mathbf{w}_0^T \cdot \mathbf{w}$. The minimal error is obtained by letting $R \rightarrow 1$ in Eq. (56) obtaining $\lambda_{\min} = \frac{1}{2} \operatorname{erf}(|\delta|/\sqrt{2})$. Note that $\bar{\lambda} = 1/2$ for $\delta = \pm\infty$, as expected. The calculation discussed previously for the realizable case, $\delta = 0$, follows almost unchanged, giving rise to Eq. (54)

with the only modification being that $\bar{\lambda}(R)$ is replaced by Eq. (56). Computing the upper bound on the stochastic complexity to leading order in $1/\alpha$ we find

$$E_D \left[-\frac{1}{n} \log P_{\mathcal{H}}(y_1^n | \mathbf{x}_1^n) \right] \leq \beta \lambda_{\min}(\delta) + \log z_{\beta} + \frac{\log \log \alpha}{2\alpha} \quad (\alpha \rightarrow \infty). \quad (57)$$

Note that the decay of the stochastic complexity to its limiting behavior is *faster* in this case than in previous realizable case. These calculations may be repeated for many different models (see Watkin, et al., 1993, and the references therein), resulting in similar expressions which depend of course on the details of the problem, as discussed in section 6.

Appendix B

In this appendix, we extend the usual Laplace method of asymptotic integration to deal with integrals whose integrand may be non-analytic, and thus not expandable into the usual Taylor series. In particular, we consider integrals of the form

$$I(n) = \int_X d\mathbf{x} g(\mathbf{x}) e^{-nf(\mathbf{x})}, \quad (58)$$

where $X \subset R^k$ is the integration region and the function $f(\mathbf{x})$ attains its global minimum at a *finite* set of points $\{\mathbf{x}_i\} \in X$. We now establish the following result (see Amari, et al. (1992) for a different proof under slightly different conditions).

THEOREM. *Let $f(\mathbf{x})$ and $g(\mathbf{x})$ be continuous real functions with domain $X \subset R^k$. We assume the following to hold:*

1. *$f(\mathbf{x})$ attains its global minimum, f_{\min} , at a finite set of points $\{\mathbf{x}_i\}_{i=1}^K$.*
2. *The function $f(\mathbf{x})$ may be expanded near its respective minima as a homogeneous (but not necessarily analytic) function of degree s_i in the Euclidean norm $\|\mathbf{x} - \mathbf{x}_i\|$, namely*

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_i} \frac{|f(\mathbf{x}) - f(\mathbf{x}_i)|}{\|\mathbf{x} - \mathbf{x}_i\|^{s_i}} = c_i \quad (0 < c_i < \infty).$$

3. *The integral (58) exists for all positive integers n .*
4. *There is a positive constant b such that $f(\mathbf{x}) - f_{\min} > b$ for \mathbf{x} outside some bounded subset of X .*
5. *The function $g(\mathbf{x})$ is integrable and can be expanded in a convergent power series around any of the minima of $f(\mathbf{x})$.*

Then, the leading asymptotic behavior of $I(n)$ is given by

$$I(n) \sim \frac{2\pi^{k/2}}{s(cn)^{k/s}} \frac{\Gamma(\frac{k}{s})}{\Gamma(\frac{k}{2})} g(\mathbf{x}^*) e^{-nf_{\min}}, \quad (59)$$

where $s = \max(s_1, \dots, s_K)$, c is the value of c_i at the minimum with the highest value of s_i and \mathbf{x}^ is the corresponding value of \mathbf{x} .*

PROOF. We first establish the theorem in the one-dimensional case assuming a single global minimum at x_0 . The extension of the theorem to the multi-dimensional case comprising

multiple global minima will then be sketched. Our proof follows that of de Bruijn (1981), except where details specific to our case apply. Thus, assume the integration region in the one-dimensional case is $a \leq x \leq b$, with $a < x_0 < b$ (the case of a minimum at a boundary can be similarly treated but will not be spelled out here). In order to simplify the proof we assume that $f_{\min} = 0$ and $g(\mathbf{x}) = 1$. The former restriction is easily removed by defining $f' = f - f_{\min}$, and the latter causes no problems due to the regularity assumption (5) above. Now, given $\epsilon > 0$ one can find a $\delta > 0$ such that

$$|f(x) - c|x - x_0|^s| < \epsilon |x - x_0|^s \quad (|x - x_0| \leq \delta),$$

keeping in mind that $f(x_0) = 0$. As in the standard approach to Laplace integration, one splits the integration region as follows

$$I(n) \triangleq \int_a^b dx e^{-nf(x)} = \left(\int_a^{x_0-\delta} + \int_{x_0-\delta}^{x_0+\delta} + \int_{x_0+\delta}^b \right) dx e^{-nf(x)}. \quad (60)$$

It is easy to see at this point that under the assumptions of the theorem (see de Bruijn, 1981) the first and third integrals in (60) contribute vanishingly small terms in the limit $n \rightarrow \infty$. The integral around the minimum x_0 may then be bounded as follows:

$$\int_{x_0-\delta}^{x_0+\delta} dx e^{-n(c+\epsilon)|x-x_0|^s} < \int_{x_0-\delta}^{x_0+\delta} dx e^{-nf(x)} < \int_{x_0-\delta}^{x_0+\delta} dx e^{-n(c-\epsilon)|x-x_0|^s} \quad (61)$$

□

Once again the standard arguments can be applied showing that the limits of all integrals may be extended to infinity at the price of introducing exponentially small terms. Using the identity

$$\int_0^\infty dx x^{k-1} e^{-nx^s} = \frac{1}{sn^{k/s}} \Gamma\left(\frac{k}{s}\right) \quad k > 0, n > 0, s > 0, \quad (62)$$

we immediately find that

$$\frac{2}{s(n(c+\epsilon))^{1/s}} \Gamma\left(\frac{1}{s}\right) < I(n) < \frac{2}{s(n(c-\epsilon))^{1/s}} \Gamma\left(\frac{1}{s}\right). \quad (63)$$

Since ϵ is arbitrarily small we immediately obtain the desired result. The extension to the multi-dimensional case can be easily made by noting that since the integrand near the minimum, \mathbf{x}_0 , depends by assumption only on the Euclidean norm $\|\mathbf{x} - \mathbf{x}_0\|$, one may transform to spherical coordinates, $r^{k-1} dr d\Omega$, obtaining

$$I(n) \sim \int_0^\infty dr \int d\Omega r^{k-1} e^{-cnr^s} = \frac{2\pi^{k/2}}{s(cn)^{k/s}} \frac{\Gamma\left(\frac{k}{s}\right)}{\Gamma\left(\frac{k}{2}\right)}, \quad (64)$$

where we have used the result $\int d\Omega = 2\pi^{k/2}/\Gamma(k/2)$. Including the function $g(\mathbf{x})$ and relaxing the requirement that $f_{\min} = 0$, immediately yields the desired result, Eq. (59). Finally, the extension to the case of multiple discrete global minima can be obtained by splitting the integration region into small volumes around each minimum, applying the above argument to each such integral and retaining only the leading contribution in the limit $n \rightarrow \infty$. The standard result for Laplace integration (where the function $f(\mathbf{x})$ can be expanded quadratically around the minimum) can be obtained by setting $s = 2$ in Eq. (59).

Acknowledgment

We are grateful to Manfred Oppel and Ofer Zeitouni for helpful discussions. R. Meir thanks Haim Sompolinsky and David Wolpert for useful comments which helped to clarify the paper. Finally, we thank the two anonymous reviewers for their very perceptive remarks about the proper application of Laplace type methods.

Notes

1. We are grateful to Manfred Oppel for pointing out this inequality to us.

References

- Amari, S., Fujita, N., & Shinomoto, S. (1992). Four Types of Learning Curves, *Neural Computation*, 4:605–618.
- Amari, S., & Murata, N. (1993). Statistical theory of learning curves under entropic loss, *Neural Computation*, 5:140–153.
- de Bruijn, N.G. (1981). *Asymptotic Methods in Analysis*, Dover Publications, New York.
- Györfyi, G., & Tishby, N. (1990). Statistical theory of learning a rule, in W.K. Theumann and R. Köberle, (Eds.), *Neural Networks and Spin Glasses*.
- Hardy, G., Littlewood, L.E., & Polya, G. (1952). *Inequalities*, Cambridge University Press, Cambridge.
- Haussler, D., & Barron, A.D. (1993). How well do Bayes methods work for on-line prediction of $\{\pm 1\}$ values, preprint.
- Haussler, D., Kearns, M.J., & Schapire, R.E. (1994a), Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension, *Machine Learning*, 14:83–113.
- Haussler, D., Kearns, M.J., Seung, H.S., & Tishby, N., (1994b). Rigorous learning curve bounds from statistical mechanics, preprint.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal function approximators, *Neural Networks*, 2:359–366.
- Kearns, M.J., Schapire, R.E., & Sellie, L.M. (1992). Towards efficient agnostic learning. In *Proceedings of the fifth annual ACM conference on computational learning theory*, ACM Press, New York.
- Levin, E., Tishby, N., & Solla, S. (1990). A statistical approach to learning and generalization in layered neural networks., *Proc. IEEE*, 78:1568–1574.
- Meir, R., & Fontanari, J.F. (1992). Calculation of learning curves for inconsistent algorithms, *Phys. Rev. A*, 45(12):8874–8884.
- Meir, R., & Fontanari, J.F. (1993) Data compression and prediction in neural networks, *Physica A*, 200:644–654.
- Meir, R., & Merhav, N. (1994). Unpublished.
- Mezard, M., Parisi, P., & Virasoro, M.A. (1987). *Spin Glass Theory and Beyond*, World Scientific, Singapore.
- Natarajan, B.K. (1991). *Machine Learning: A Theoretical Approach*, Morgan Kaufmann, San Mateo, California.
- Oppel, M., & Haussler, D. (1991). Calculation of the learning curve of Bayes optimal classification algorithm for learning a perceptron with noise. In *Proceedings of the fourth annual ACM conference on computational learning theory*, Morgan Kaufmann, San Mateo.
- Rissanen, J. (1986). Stochastic complexity and modeling, *Ann. Stat.* 14(3):1080–1100.
- Rosenblatt, F. (1962). *Principles of Neurodynamics*, Spartan, New York.
- Seung, H.S., Sompolinsky, H., & Tishby, N. (1992). Statistical mechanics of learning from examples, *Phys. Rev. A*, 45:6056–6091.
- Sontag, E.D. (1992). Feedforward nets for interpolation and classification, *J. Comp. Sys. Sci.*, 45:20–48.
- Vapnik, V.N. (1982). *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, Berlin.
- Watkin, T.L.H., & Rau, A. (1992). Learning unlearnable problems with perceptrons, *Phys. Rev. A*, 45(6):4102–4110.
- Watkin, T.L.H., Rau, A., & Biehl, M. (1993). The statistical mechanics of learning a rule, *Rev. Mod. Phys.*, 65(2):499–556.