ON THE CONSEQUENCES OF THE STATISTICAL MECHANICS THEORY OF LEARNING CURVES FOR THE MODEL SELECTION PROBLEM

MICHAEL J. KEARNS
AT&T BELL LABORATORIES
Murray Hill, New Jersey 07974, USA
E-mail: mkearns@research.att.com

1. Introduction

The Statistical Mechanics (SM) approach to the analysis of learning curves *has enjoyed increased attention and success in the field of computational learning theory over the past several years. In part due to the novelty of its technical methods, and in part due to its identification of interesting learning curve behavior not explained by classical power law theories, the SM theory has emerged as an important complement to the powerful and general Vapnik-Chervonenkis (VC) theory of learning curves. To crudely summarize the differences between the SM and VC theories, we can say that the VC theory requires less knowledge of the problem specifics than the SM theory †, but the VC theory may suffer for this more general approach by predicting learning curves that deviate further from the actual behavior than those predicted by the SM theory. It is worth emphasizing that from a mathematically rigorous point of view, both theories may offer only an upper bound on the learning curve, and these upper bounds will diverge from the true behavior as the implicit assumptions of the theories are violated by the problem under consideration. However, it seems fair to assert that the SM theory has a better chance at capturing the true behavior of the learning curve, since more specifics of the problem are taken into account.

While the SM theory has contributed new methods of analysis and new bounds on learning curves, it has done so for essentially the same learning algorithms that are considered in the VC theory and its variants. For instance, it is common in the VC framework to analyze the learning algorithm that chooses the hypothesis minimizing the training error (breaking ties arbitrarily), and it is now known that the SM theory can also provide learning curve upper bounds for this algorithm [1]. Similarly, many investigations in the SM theory focus on the Gibbs algorithm (which essentially chooses a hypothesis randomly from a distribution that exponentially penalizes training error), but the tools of the VC theory are equally applicable here as well [2]. Thus the SM theory is primarily *descriptive* rather than *prescriptive*: we may obtain new and better bounds on learning curves, but for the same algorithms we have been studying all along.

Are there natural learning problems in which the sometimes rather different predictions made by the VC and SM theories would have algorithmic consequences? Here we argue that *model selection*, in which we must choose the appropriate value for the *complexity* of our hypothesis, is such a problem. An informal example will serve to illustrate the issues. Suppose we are given a set of training data $S = \{\langle x_i, b_i \rangle\}_{i=1}^{m}$

*As an informal working definition, the *learning curve* is the plot of the generalization error as a function of the number of examples, when the hypothesis or student model is chosen by some "natural" learning algorithm such as the Gibbs algorithm.

†More precisely, in the context of supervised learning of boolean functions from independent random examples, the VC theory requires only knowledge of the class of hypothesis or student functions (actually, only the VC dimension of this class is required), while the SM theory additionally requires knowledge of the input distribution, and in certain cases, knowledge of the target or teacher function.

where the x_i are vectors of real numbers and the b_i are binary values. We wish to choose a neural network with a single layer of hidden units on the basis of the data (for simplicity, let us assume full connectivity between adjacent layers). Now if we are given a *fixed* value d for the number of hidden units, both the VC and SM theories could be used to relate the generalization error of h_d, the network of d hidden units minimizing the training error on S, to the normalized sample size m/d. Let us suppose that the VC theory predicts that the resulting generalization error $\epsilon(d) \equiv \epsilon(h_d)$ will be $\epsilon_{\rm VC}(d)$ and the SM theory predicts that it will be $\epsilon_{\rm SM}(d)$. So far, both theories have been applied to the same learning algorithm, although the predicted values $\epsilon_{\rm VC}(d)$ and $\epsilon_{\rm SM}(d)$ may be quite different.

Now suppose that rather than being given a fixed value, we must *choose* the number of hidden units d on the basis of the sample S, which implicitly means that we choose the hypothesis h_d; we would like to choose d to minimize the resulting generalization error $\epsilon(h_d)$. Then straighforward application of the two theories suggest the choices $d_{\rm VC} = argmin_d\{\epsilon_{\rm VC}(d)\}$ and $d_{\rm SM} = argmin_d\{\epsilon_{\rm SM}(d)\}$. These two values for d, and therefore the resulting generalization errors, may be quite different, despite the fact that the underlying training procedure used to choose a network of a *fixed* size is the same in both theories.

In effect, we are saying the following: fix an underlying training procedure (in this case, training error minimization) that for any sample S and any "complexity" d will produce a hypothesis function h_d, whose unknown generalization error we are denoting $\epsilon(d)$. Then by applying the VC and SM theories for each value of d, we obtain predictions $\epsilon_{\rm VC}(d)$ and $\epsilon_{\rm SM}(d)$ for the behavior of the function $\epsilon(d)$. The "ideal" choice for d satisfies $d = argmin_d\{\epsilon(d)\}$, and the extent to which the values $d_{\rm VC}$ and $d_{\rm SM}$ deviate from this ideal (and the extent to which this deviation results in inferior generalization) is clearly determined by the accuracy of $\epsilon_{\rm VC}(d)$ and $\epsilon_{\rm SM}(d)$ as models of $\epsilon(d)$.

The crux of the model selection problem lies in the fact that, in general, we expect the minimum of $\epsilon(d)$ to be achieved at some "intermediate" value. For values of d that are too small, $\epsilon(d)$ will be large simply because the representational power provided by networks with so few hidden units is insufficient to accurately model the unknown target function. For values of d that are too large, $\epsilon(d)$ will again be large, because even though there is sufficient representational power, we do not have enough data to constrain the weights to values that will result in an accurate hypothesis. In other words, as we increase d for fixed m, at some point the ratio m/d becomes too small for us to reliably increase d any further without suffering greater generalization error. The question, of course, is where exactly is this point? Since the VC and SM theories may sometimes differ radically in their predictions of the generalization error to be expected at any given ratio m/d, we should expect them to sometimes provide radically different solutions to the problem of model selection.

In the remainder of the paper, we first examine more closely the solutions to the model selection problem that are implicit in the two theories. We then propose a specific model selection problem, based on the learning curve for the committee machine, in which the two solutions should be quite different. After a discussion of the relative difficulty of directly applying the SM solution in general, we summarize some recent results [4] in which the SM theory is instead invoked to argue in favor of another well-known model selection method, namely cross validation. We conclude with a discussion of interesting avenues for further research.

2. The VC Solution to Model Selection

Before describing the VC theory solution to model selection, it will be helpful to generalize the simple example discussed above to provide a more abstract setting for model selection problems. In the context of supervised learning of boolean functions from independent random examples, a *model selection problem* consists of an arbitrary *target* or *teacher* function f, a distribution D over the inputs x to f, a sample size m, a *training algorithm* L, and a nested sequence of *hypothesis* or *student* function classes

$$F_1 \subseteq \cdots \subseteq F_d \subseteq \cdots$$

For any function h, the target function f and the input distribution D together define the generalization error $\epsilon(h) \equiv \Pr_{x \in D}[f(x) \neq h(x)]$. The training sample S consists of m randomly chosen examples of f, possibly corrupted by noise. The training algorithm L accepts as input the training sample S and a complexity value d, and outputs a function $h_d \in F_d$; as we range over all the possible values for d, the training algorithm thus yields a sequence $h_1, \ldots, h_d, \ldots$ of increasingly "complex" hypotheses (where complexity is defined by the partial ordering on functions implied by the sequence of hypothesis function classes F_d). The model selection problem is simply that of choosing d to minimize $\epsilon(h_d) = \epsilon(d)$. We assume that the model selection procedure has no information about the target function (and possibly no information about the input distribution), and thus seeks to minimize the resulting generalization error in, say, a minimax sense. In the example discussed above, we did not specify f and D, but the training algorithm L was training error minimization, and F_d was the class of all neural networks with at most d hidden units. As another example, we might consider the case where L is the Gibbs algorithm, and F_d is the class of all committee machines with at most d committee members.

In the VC learning curve theory, the prediction $\epsilon_{\mathrm{VC}}(d)$ for the generalization error $\epsilon(d)$ depends not only on F_d and m, but also on the training error $\hat{\epsilon}(d) \equiv \hat{\epsilon}(h_d) \equiv |\{\langle x_i, b_i \rangle \in S : h_d(x_i) \neq b_i\}|/m$. The resulting expression for $\epsilon_{\mathrm{VC}}(d)$ is

$$\epsilon_{\mathrm{VC}}(d) = \hat{\epsilon}(d) + c_0 \frac{\mathrm{VCD}(F_d)}{m} \log \frac{m}{\mathrm{VCD}(F_d)} \left(1 + \sqrt{1 + \hat{\epsilon}(d) \frac{m}{\mathrm{VCD}(F_d) \log \frac{m}{\mathrm{VCD}(F_d)}}} \right) \quad (1)$$

where $c_0 > 1$ is a constant and $\mathrm{VCD}(F_d)$ is the Vapnik-Chervonenkis dimension of F_d [3]. The detailed derivation of this expression is rather involved, but the reader is encouraged to consult Vapnik's fascinating and beautiful book on the subject. A few remarks on $\epsilon_{\mathrm{VC}}(d)$ are in order here. First, it can be rigorously shown that with high probability over the random sample S, $\epsilon_{\mathrm{VC}}(d) \geq \epsilon(d)$ [3]. In fairness to the VC theory, this is the only property of $\epsilon_{\mathrm{VC}}(d)$ that is needed for many computational learning theory results — in many applications (such as deriving sample size bounds for particular learning algorithms), it is not necessary to propose a model for the exact behavior of $\epsilon(d)$, and an upper bound will suffice. Second, although it may be difficult to see from Equation (1), the expression for $\epsilon_{\mathrm{VC}}(d)$ is essentially based on a power law theory of learning curves — or more precisely, on power law uniform convergence bounds. By this we mean that the main technical fact underlying Equation (1) is that with high probability over S, for any function h in the class F_d the deviation of the training error of h from the generalization error of h is bounded above by $\sqrt{\mathrm{VCD}(F_d)/m}$

‡ Furthermore, the largest deviation between the training and generalization error over the functions in F_d is closely related to the learning curve for the class F_d. Thus the extent to which $\epsilon_{\rm VC}(d)$ approximates $\epsilon(d)$ is directly related to how much the learning curves for the F_d resemble power laws § In any case, the expression for $d_{\rm VC}$ — the VC theory choice of complexity in the model selection problem — simply becomes

$$argmin_d \left\{ \hat{\epsilon}(d) + c_0 \frac{\mathrm{VCD}(F_d)}{m} \log \frac{m}{\mathrm{VCD}(F_d)} \left(1 + \sqrt{1 + \hat{\epsilon}(d) \frac{m}{\mathrm{VCD}(F_d) \log \frac{m}{\mathrm{VCD}(F_d)}}} \right) \right\} \tag{2}$$

3. The SM Solution to Model Selection

To illustrate how the SM theory can be extended to obtain a prescription for model selection, we will rely on the formalization provided by Haussler, Kearns, Seung and Tishby [1]; this will require a number of technical assumptions. However, we believe that the less rigorous but more general methods that have been in use in the SM theory are also applicable to model selection, and the reader well-versed in that literature should be able to envision the extension.

To apply the formalism, we need to assume that the function classes F_d are finite [1]; as a concrete example, we could consider the case where F_d is the class of all *binary* committee machines with at most d committee members (thus, the weights of each committee member must be either $+1$ or -1). In the following discussion, we assume that the (arbitrary) target function f and input distribution D are fixed (and define the generalization error $\epsilon(\cdot)$).

In this formalization of the SM theory, the crucial quantity to be examined for each class F_d is the *entropy* function $s_d(\epsilon)$, defined by

$$s_d(\epsilon) \equiv \frac{1}{d} \log |F_d(\epsilon)| \tag{3}$$

where $F_d(\epsilon) \subseteq F_d$ is the set of all functions in F_d whose generalization error with respect to f and D is "approximately" ϵ ¶ The function $s_d(\epsilon)$ plays a similar role in the SM theory to that played by the VC dimension in the VC theory, in the sense that the primary dependence that the SM theory has on the specifics of the problem (that is, on f, D and m) is through $s_d(\epsilon)$.

Speaking informally, the generalization error predicted by the SM theory (when the hypothesis is chosen to minimize the training error in F_d on m random examples) is $\epsilon_{\rm SM}(d)$, where $\epsilon_{\rm SM}(d)$ is the largest value of ϵ satisfying

$$s(\epsilon) \geq -\frac{m}{d} \ln \left(1 - \left(\sqrt{\epsilon} - \sqrt{\epsilon_{opt}(d)} \right)^2 \right) . \tag{4}$$

‡ A more refined bound that varies from $\sqrt{\mathrm{VCD}(F_d)/m}$ for small d to $\mathrm{VCD}(F_d)/m$ for large d holds, and is actually the bound used to derive Equation (1).

§ More precise statements of this nature are possible, but this will suffice for our purposes.

¶ See the paper of Haussler, Kearns, Seung and Tisbhy [1] for technical details. A number of variations of this basic definition are allowed, including ones that replace the factor $1/d$ by $1/t(d)$ for some appropriate "scaling function" $t(\cdot)$, and that permit $s_d(\epsilon)$ to merely upper bound the right-hand side of Equation (3).

Here $\epsilon_{opt}(d) \equiv min_{h \in F_d}\{\epsilon(h)\}$ is the best generalization error that can be achieved within the class F_d. The value $\epsilon_{\rm SM}(d)$ has a natural interpretation as a competition between entropy and energy that is a common metaphor in the SM theory.

Like the function $\epsilon_{\rm VC}(d)$, $\epsilon_{\rm SM}(d)$ provides a rigorous upper bound on the true function $\epsilon(d)$ under certain technical conditions [1] ‖ As discussed earlier, we can interpret $\epsilon_{\rm SM}(d)$ as more than just a bound on $\epsilon(d)$ — in the model selection problem, we can hope that $\epsilon_{\rm SM}(d)$ is a good approximation to $\epsilon(d)$ and choose $d_{\rm SM} = argmin_d\{\epsilon_{\rm SM}(d)\}$.

4. An Informal and Illustrative Example

Let us summarize where we are. For fixed sample size m, and for each value of d, the VC theory implicitly predicts that the resulting generalization error $\epsilon(d)$ obeys $\epsilon(d) = \epsilon_{\rm VC}(d)$, where $\epsilon_{\rm VC}(d)$ depends only on m, $\mathrm{VCD}(F_d)$ and the training error $\hat{\epsilon}(d)$. The SM theory predicts that $\epsilon(d) = \epsilon_{\rm SM}(d)$, where $\epsilon_{\rm SM}(d)$ depends on m, F_d, and the input distribution D. It is known that $\epsilon_{\rm SM}(d)$ is a better approximation to $\epsilon(d)$ than $\epsilon_{\rm VC}(d)$, at least in the finite F_d case [1], and we should expect it to be true for the general (non-finite) case — $\epsilon_{\rm SM}(d)$ simply takes into account more of the specifics of the problem. Thus, we should also expect the model selection problem solution derived from $\epsilon_{\rm SM}(d)$ to be superior to that derived from $\epsilon_{\rm VC}(d)$. In this section, we simply wish to describe a simple example where this seems likely to be the case. We will treat the example rather informally, and leave the formal verification as an open problem.

Let us suppose that the target function f is a committee machine with exactly d^* committee members with binary weights, and that the input distribution D is uniform over $\{+1, -1\}^n$ (here n is the number of inputs to the target machine f). Let the class F_d consist of all committee machines with d or fewer committee members with binary weights [**], so $f \in F_{d^*}$.

First consider the functions $\epsilon(d)$ and $\epsilon_{\rm SM}(d)$ when the number of examples m is of order d^*n. There is strong evidence from the SM theory, supported by experimental simulations, that for some constant $\alpha_c > 0$, if $m = \alpha_c d^* n$, then $\epsilon(d^*) = 0$, while if $m < \alpha_c d^* n$ then $\epsilon(d^*) > c_0$ for some constant $c_0 > 0$ [5,6] — in other words, there is a first-order phase transition to perfect generalization at $\alpha_c d^* n$ examples [††] The SM theory correctly predicts this behavior — that is, $\epsilon(d^*) = 0$ for $m = \alpha_c d^* n$ and $\epsilon_{\rm SM}(d^*) > c_0$ for $m < \alpha_c d^* n$. In the context of model selection, this means that $d_{\rm SM} = d^*$, and the generalization error suffered by using the SM theory to choose a value for d is therefore 0.

In contrast, let us examine the function $\epsilon_{\rm VC}(d)$ under these same circumstances. For $m = \alpha_c d^* n$ examples, $\epsilon_{\rm VC}(d^*)$ will be a non-zero constant, because $\mathrm{VCD}(F_{d^*})$ is of order $d^* n$ [7], meaning that the expression $\mathrm{VCD}(F_{d^*})/m$ (which appears additively in the expression for $\epsilon_{\rm VC}(d)$ given in Equation (1)) will be a non-zero constant depending on α_c. The main point here is that *for this particular problem, the VC theory*

‖ Perhaps the most important and restrictive of these conditions is the requirement that we take a thermodynamic limit; however, in certain cases finite sample size bounds on generalization error can also be obtained.

[**] The ensuing argument should also hold qualitatively for the case where the committee members have continuous weights [6].

[††] Again, these results formally rely on a thermodynamic limit, but a rapid increase in the rate of generalization is seen in simulations near $\alpha_c d^* n$ examples even for finite systems.

overpenalizes for complexity, and thus $\epsilon_{\mathrm{VC}}(d^*)$ *greatly overestimates* $\epsilon(d)$. In the context of model selection, this may result in $d_{\mathrm{VC}} << d^*$ — the minimization of $\epsilon_{\mathrm{VC}}(d)$ may yield a value of d that suffers large training error $\hat{\epsilon}(d)$ in exchange for reducing the penalty for complexity. If this were the case, then choosing d_{VC} would cause increased generalization error compared to choosing d_{SM}. We have not carefully determined whether the VC theory's overpenalization for complexity actually results in inferior generalization error for this particular model selection problem; however, it is possible to rigorously demonstrate the phenomenon we have discussed here on other model selection problems, and in fact show that the additional error suffered by choosing d_{VC} instead of d_{SM} can be quite large [4].

5. Remarks on the SM Theory and Cross Validation

The remarks of the preceding section indicate how the solutions to model selection suggested by the VC and SM theories may differ, and why it may result in inferior performance by the VC method on certain problems. Despite this, we do not expect the SM model selection method outlined in Section to become a serious competitor to the VC method of Section for one simple and obvious reason: the general SM approach requires too much knowledge of the problem specifics (knowledge that may be difficult or impossible to attain), and the calculations required are extremely difficult. Nevertheless, even if we do not choose to directly implement the SM solution to model selection, the SM theory still has algorithmic consequences for the model selection problem, namely by providing concrete arguments for favoring the use of cross validation over the VC method (and similar "penalty-based" methods such as the minimum description length principle) for certain problems. Here we just briefly remark on why SM theory discoveries lend support for cross validation, and refer the reader to an extensive paper quantifying and verifying the ideas given informally here [4].

Suppose that we began by assuming that for every problem, $\epsilon_{\mathrm{VC}}(d)$ were a reasonably good model of $\epsilon(d)$, denoted $\epsilon_{\mathrm{VC}}(d) \approx \epsilon(d)$. Using Equation (1), we see that this is tantamount to assuming that there is a *universal* relationship between the training error $\hat{\epsilon}(d)$ and the generalization error $\epsilon(d)$. Perhaps the main contribution of the SM theory to research on learning curves has been its conclusive demonstration of the fact that there is no such universal relationship — there is a great diversity of possible relationships between $\hat{\epsilon}(d)$ and $\epsilon(d)$. The example given in the last section is based on this diversity — it is simply a problem in which the relationship between $\hat{\epsilon}(d)$ and $\epsilon(d)$ that is assumed by the VC theory does not hold.

Given the diversity of learning curves established by work in the SM theory, it seems reasonable to assert that *any* model selection method based on the assertion of a universal relationship between $\hat{\epsilon}(d)$ and $\epsilon(d)$ would be doomed to failure on at least some problems. This is in fact the case, and can be verified rigorously [4]. From this viewpoint, the strength of cross validation lies in the fact that it assumes no such universal relationship — rather, regardless of the behavior of $\hat{\epsilon}(d)$, cross validation directly estimates $\epsilon(d)$ using an independent test sample. Of course, the issue of how much generalization ability is lost due to excluding part of the sample from the training process is important here, and for some problems may in fact cause cross validation to have inferior performance. However, our claim here is that under fairly common circumstances, the flexibility cross validation enjoys by directly estimating $\epsilon(d)$ compensates for the diversity of learning curves that proves fatal for many other model selection methods. Again, these ideas are made considerably more precise in

a recent paper [4].

6. Topics for Further Research

Despite the difficulty of the required calculations in general, it would be interesting to work out the details of the SM prescription for model selection outlined in Section for particular problems, such as choosing the best number of committee members. Experimental comparisons with other approaches such as the VC theory, cross validation and the minimum description length principle could then be carried out.

7. Acknowledgements

Thanks to all of the participants of the workshop *Neural Networks: The Statistical Mechanics Perspective* for a lively meeting, and special thanks to the workshop's mastermind and organizer, Prof. Jong-Hoon Oh. I am also grateful to my colleague H.S. Seung for many valuable and fascinating discussions on learning curves and model selection.

1. D. Haussler, M. Kearns, H.S. Seung, and N. Tishby, in *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory* (ACM Press, 1994), p. 76.
2. D. Haussler, M. Kearns, and R.E. Schapire, *Machine Learning* 14 (1994), p. 83.
3. V. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer-Verlag, New York, 1982).
4. M. Kearns, Y. Mansour, A.Y. Ng, and D. Ron, to appear in *Proceedings of the Eighth Annual ACM Conference on Computational Learning Theory* (ACM Press, 1995).
5. K. Kang, J. Oh, C. Kwon, and Y. Park, *Phys. Rev E* 48(6) (1993), p. 4805.
6. H. Schwarze and J. Hertz, *J. Phys. A: Math. Gen.* 26 (1993), p. 4919.
7. E. Baum and D. Haussler, *Neural Computation* 1(1) (1989), p. 151.