Machine Learning 25, 195–236 (1996)

Rigorous Learning Curve Bounds from Statistical Mechanics

DAVID HAUSSLER
U.C. Santa Cruz, Santa Cruz, California

MICHAEL KEARNS
AT&T Laboratories Research, Murray Hill, New Jersey

H. SEBASTIAN SEUNG
Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey

NAFTALI TISHBY
Hebrew University, Jerusalem, Israel

Editor: Thomas Hancock

Abstract. In this paper we introduce and investigate a mathematically rigorous theory of learning curves that is based on ideas from statistical mechanics. The advantage of our theory over the well-established Vapnik-Chervonenkis theory is that our bounds can be considerably tighter in many cases, and are also more reflective of the true behavior of learning curves. This behavior can often exhibit dramatic properties such as phase transitions, as well as power law asymptotics not explained by the VC theory. The disadvantages of our theory are that its application requires knowledge of the input distribution, and it is limited so far to finite cardinality function classes.

We illustrate our results with many concrete examples of learning curve bounds derived from our theory.

Keywords: learning curves, statistical mechanics, phase transitions, VC dimension

1. Introduction

According to the Vapnik-Chervonenkis (VC) theory of learning curves (Vapnik, 1982; Vapnik & Chervonenkis, 1971), minimizing empirical error within a function class $\mathcal{F}$ on a random sample of m examples leads to generalization error bounded by $\tilde{O}(d/m)$ (in the case that the target function is contained in $\mathcal{F}$) or $\tilde{O}(\sqrt{d/m})$ plus the optimal generalization error achievable within $\mathcal{F}$ (in the general case)[1]. These bounds are universal: they hold for any class of hypothesis functions $\mathcal{F}$, for any input distribution, and for any target function. The only problem-specific quantity remaining in these bounds is the VC dimension d, a measure of the complexity of the function class $\mathcal{F}$. It has been shown that these bounds are essentially the best distribution-independent bounds possible, in the sense that for any function class, there exists an input distribution for which matching lower bounds on the generalization error can be given (Devroye & Lugosi, 1994; Ehrenfeucht et al., 1989; Simon, 1993).

The universal VC bounds can give the impression that the *true behavior* of learning curves is also universal, and essentially described by the functional forms d/m and $\sqrt{d/m}$. However, it is becoming clear that learning curves exhibit a diversity of behaviors. For instance, some researchers have attempted to fit learning curves from backpropagation experiments with a variety of functional forms, including exponentials (Cohn & Tesauro, 1992). Backpropagation experiments with handwritten digits and characters indicate that good generalization error is sometimes obtained for sample sizes considerably smaller than the number of weights (presumed to be roughly the same as the VC dimension) (Martin & Pittman, 1991), though the VC bounds are vacuous for m smaller than d. Discrepancies between the VC bounds and actual learning curve behavior have also been pointed out and analyzed in other machine learning work (Oblow, 1992; Sarrett & Pazzani, 1992).

Of course, the VC bounds might simply be inapplicable to these experiments, because backpropagation is not equivalent to empirical error minimization. It has been conjectured that backpropagation can access only a limited portion of the function space, so that the "effective dimension" is much smaller than the VC dimension. According to this type of reasoning, learning curves are heavily affected by the specifics of the algorithm. Another possibility is that the VC bounds are applicable, but sometimes fail to capture the true behavior of particular learning curves because of their independence from the distribution. Hence some theorists have sought to preserve the functional form of the VC bounds, but to replace the VC dimension in this functional form by an appropriate distribution-specific quantity, such as the VC entropy (which is the expectation of the logarithm of the number of dichotomies realized by the function class) (Benedek & Itai, 1991; Haussler et al., 1991; Vapnik, 1982). Work on the "empirical VC dimension" has tried to measure the dependence of learning curves on both the algorithm and the distribution via backpropagation experiments (Vapnik et al., 1994).

Perhaps the most striking evidence for the fact that the VC bounds can sometimes fail to model the true behavior of learning curves has come from statistical physics. In recent years, the tools of statistical mechanics have been applied to analyze learning curves with rather curious and dramatic behavior (see the survey of Watkin, Rau and Biehl and the references therein (Watkin et al., 1993)). This has included learning curves exhibiting "phase transitions" (sudden drops in the generalization error) at small sample sizes, as well as asymptotic power law behavior[2] in which the power law exponent is neither 1 nor $1/2$. Although these learning curves do not contradict the VC bounds, it seems fair to say that their behavior is qualitatively different. The theoretical revisions of the VC theory mentioned above cannot explain such behavior, because they conservatively modify only with the constant factors of the same power laws.

In this paper, we show that ideas from statistical mechanics (namely, the annealed approximation (Amari et al., 1992; Levin et al., 1989; Schwartz et al., 1990; Sompolinsky et al., 1991) and the thermodynamic limit (Sompolinsky et al., 1991)) can be used as the basis of a mathematically precise and rigorous theory of learning curves[3]. This theory will be distribution-specific, but will not attempt to force a power law form on learning curves. Speaking coarsely, there are two main ideas behind our theory that are novel to someone familiar with the VC theory. The first new idea is related to the annealed approximation. It is based on the simple observation that in the VC theory and its proposed distribution-dependent variants, all hypotheses of generalization error greater than ϵ are treated equally by the analysis—for instance, by assigning $(1 - \epsilon)^m$ to all such hypotheses as an upper bound on the probability of being consistent with m random examples. We undertake a more refined analysis that decomposes the function class into *error shells* that actually attribute the correct generalization error to each hypothesis, and give uniform convergence bounds on each shell. The resulting bounds already predict learning curve behavior not explained by the VC theory, but are difficult to interpret.

The second new idea is to formalize a particular mathematical limit known to statistical physicists as the *thermodynamic limit*. The goal of this limit is to express the error shell decomposition bounds in a form that is both useful and intuitive. The thermodynamic limit accomplishes this goal by introducing the notion of the correct *scale* at which to analyze a learning curve, and by expressing the learning curve as a competition between an entropy function (measuring the logarithm of number of hypotheses as a function of their generalization error ϵ) and an energy function (measuring the probability of minimizing the empirical error on a random sample as a function of generalization error).

The resulting theory provides a formalized variant of the statistical physics approach that is able to predict and explain many nontrivial behavioral phenomena of learning curves, including phase transitions. It is far from being the last word on learning curves, and indeed, the task of providing a truly universal theory of learning curves—one that applies to all function classes, input distributions, and target functions, and is furthermore *tight* in all cases—appears to be a daunting if not unreasonable task. Furthermore, this paper concentrates on the case of finite cardinality function classes (although we provide some discussion of possible extensions to the infinite case). For someone familiar with the VC theory, it may be somewhat surprising that we devote so much effort to the finite case, since in the VC theory a power law uniform convergence bound can be obtained trivially for finite classes. Briefly, it turns out that in our formalism, it can be nontrivial to translate a collection of separate uniform convergence bounds, one for each error shell, into a learning curve bound, even in the finite case. By concentrating on this translation step, our methods can yield much tighter learning curve bounds than the VC theory in some cases.

The reader should regard the current paper as having three primary goals. First, we aim to derive from first principles a formal theory retaining the spirit of the statistical mechanics approach. Second, we aim to provide evidence in the form of specific examples and a general lower bound that the new theory truly is closer to modeling the actual behavior of learning curves than the standard VC theory. Third, we aim to precisely relate the statistical mechanics approach to the VC theory.

2. The finite and realizable case

We begin with the most basic model of learning an unknown boolean target function. We assume that the target function f is chosen from a known class $\mathcal{F}$ of $\{0, 1\}$-valued functions over an input space X. We refer to this as the *realizable* setting, since the learning algorithm knows a class of functions that contains or *realizes* the target function. We also assume that $\mathcal{F}$ has finite cardinality.

The learning process consists of giving a learning algorithm a fixed finite number m of independent random *training examples* of f. Thus, let D be any fixed probability distribution over X. The learning algorithm receives as input a training sample $S = \{\langle x_i, f(x_i)\rangle\}_{1 \leq i \leq m}$. Each input x_i in the training sample is chosen randomly and independently according to the fixed distribution D. For any boolean function h, the *generalization error* of h is the probability of disagreement between h and f : $\epsilon_{\text{gen}}(h) = \mathbf{Pr}_{x \in D}[h(x) \neq f(x)]$. Note that the training sample S depends on f and m and $\epsilon_{\text{gen}}(h)$ depends on f and D. Throughout the paper we will consider these quantities as fixed and suppress such dependencies.

If we let h denote the *hypothesis* function output by a "reasonable" learning algorithm following training on m examples, what is the behavior of $\epsilon_{\text{gen}}(h)$ as a function of the sample size m? In this paper, "reasonable" will essentially mean any algorithm that chooses a hypothesis function that is *consistent* with the training sample (or one that chooses a hypothesis with minimum empirical error on the sample in the unrealizable case). This notion is both natural and mathematically convenient, because it allows us to give an analysis of the behavior of $\epsilon_{\text{gen}}(h)$ that ignores the details of the learning algorithm, and to instead concentrate exclusively on the expected error of any consistent hypothesis.

2.1. *Relating the version space to the ϵ-ball*

For any sample S, we define the *version space* by

$$\text{VS}(S) = \{h \in \mathcal{F} : \forall \langle x, f(x)\rangle \in S, h(x) = f(x)\}.$$

Thus, $\text{VS}(S) \subseteq \mathcal{F}$ is simply the subclass of all functions h that are *consistent* with the target function f on the sample S. The *ϵ-ball* about the target function f is defined as the set of all functions with generalization error not exceeding ϵ:

$$B(\epsilon) = \{h \in \mathcal{F} : \epsilon_{\text{gen}}(h) \leq \epsilon\}.$$

Thus, $\text{VS}(S)$ is a sample-dependent subclass of $\mathcal{F}$, and $B(\epsilon)$ is a sample-independent subclass of $\mathcal{F}$, and both contain the target f.

The goal of this subsection is to examine the relationship between $\text{VS}(S)$ and $B(\epsilon)$. More specifically, for a sample S of size m, we would like to calculate the probability that $\text{VS}(S)$ is contained in $B(\epsilon)$. This probability is significant for learning, because it allows us to bound the error of any *consistent* learning algorithm: we can always assert that with probability at least $\mathbf{Pr}_S[\text{VS}(S) \subseteq B(\epsilon)]$, any consistent hypothesis has generalization error less than ϵ. Here the probability is taken over the m independent draws from D used to obtain S. We now derive a lower bound on $\mathbf{Pr}_S[\text{VS}(S) \subseteq B(\epsilon)]$, or equivalently, an upper bound on $\mathbf{Pr}_S[\text{VS}(S) \not\subseteq B(\epsilon)]$.

The probability that a function h of generalization error $\epsilon_{\text{gen}}(h)$ remains in the version space after m examples decays exponentially with m:

$$\mathbf{Pr}_S[h \in \text{VS}(S)] = (1 - \epsilon_{\text{gen}}(h))^m.$$

Since the rate of decay is slower for small $\epsilon_{\mathrm{gen}}(h)$, the version space should consist only of hypotheses with small generalization error. Let $\overline{B(\epsilon)} = \mathcal{F} - B(\epsilon)$, the functions in $\mathcal{F}$ with generalization error greater than ϵ. Since the probability of a disjunction of events is upper bounded by the sum of the probabilities of the events, we find that

$$\begin{aligned} \mathbf{Pr}_S[\mathrm{VS}(S) \not\subseteq B(\epsilon)] &= \mathbf{Pr}_S[\exists h \in \overline{B(\epsilon)} : h \in \mathrm{VS}(S)] && (1)\\ &\le \sum_{h \in \overline{B(\epsilon)}} \mathbf{Pr}_S[h \in \mathrm{VS}(S)] && (2)\\ &= \sum_{h \in \overline{B(\epsilon)}} (1 - \epsilon_{\mathrm{gen}}(h))^m && (3) \end{aligned}$$

which proves the following theorem.

Theorem 1. *$\mathbf{Pr}_S[VS(S) \subseteq B(\epsilon)] \ge 1 - \delta$, where*

$$\delta = \sum_{h \in \overline{B(\epsilon)}} (1 - \epsilon_{gen}(h))^m.$$

We will refer to Theorem 1 as the *union bound.* It is closely related to the annealed approximation, which has been used by physicists to study the performance of the Gibbs learning algorithm. Note that the sum in the union bound has a direct interpretation, being the average number of surviving hypotheses that lie outside $B(\epsilon)$.

We can restate Theorem 1 in the following alternate form, in which we regard δ as given and then bound the achievable ϵ.

Corollary 2. *Let $\mathcal{F}$ be any finite boolean function class. For any $0 < \delta \le 1$, with probability at least $1 - \delta$ any function $h \in \mathcal{F}$ consistent with m random examples of a target function in $\mathcal{F}$ obeys $\epsilon_{gen}(h) \le \epsilon$, where ϵ is the smallest value satisfying $\sum_{h \in \overline{B(\epsilon)}} (1 - \epsilon_{gen}(h))^m \le \delta$.*

2.2. *The standard cardinality bound*

Since $\epsilon_{\mathrm{gen}}(h) > \epsilon$ for all $h \in \overline{B(\epsilon)}$, the union bound can be further transformed by

$$\sum_{h \in \overline{B(\epsilon)}} (1 - \epsilon_{\mathrm{gen}}(h))^m \le \sum_{h \in \overline{B(\epsilon)}} (1 - \epsilon)^m \le |\mathcal{F}|(1 - \epsilon)^m. \tag{4}$$

By applying Theorem 1 to this bound, we obtain the standard result that with probability $1 - \delta$, any consistent hypothesis h obeys $\epsilon_{\mathrm{gen}}(h) \le (\ln(|\mathcal{F}|/\delta))/m$. Since the only dependence of this bound on the learning problem is through the cardinality of the function class $\mathcal{F}$, we will refer to it as the *cardinality bound.* In particular, it depends neither on the input distribution D nor on the target function f.

Although this bound is powerful because of its generality, there is no reason to believe that it is tight for specific distributions. Its tightness depends on the chain of inequalities beginning with Eq. (1) and those given in Eq. (4), and any link in this chain can be weak.

Most of the work of this paper will be directed toward finding tighter alternatives to Eq. (4). We will slice $\overline{B(\epsilon)}$ into many shells with different error levels rather than lump all of them together at ϵ, as was done in Eq. (4). Furthermore, our calculations will make use of all the shell cardinalities, not just the crude measure of total cardinality of the function class. This more refined bookkeeping can lead to learning curves that have radically different behavior than that predicted by the simple cardinality bound.

On the other hand, we will generally rely on the union bound as is. It is tight if the survivals of different hypotheses are mutually exclusive events. In fact, when hypotheses have small disagreement, their survivals are often positively correlated instead. Nevertheless, for the *finite* function classes examined here, the crudeness of Eq. (1) will not weaken our bounds too severely. In particular, we will exhibit examples of distribution-specific bounds that are much tighter than the distribution-free VC bounds.

It is only for *infinite* function classes that the union bound fails spectacularly, for here the bound diverges and becomes useless. The VC dimension, VC entropy, and random covering number (Dudley, 1978; Haussler, 1992; Pollard, 1984; Vapnik, 1982) are the known tools for dealing with the correlations neglected by the union bound. These tools have previously been applied to the function class as a whole. In our current research efforts, we are attempting to refine these tools by applying them to error shells. In Section 4 we discuss an alternative approach that reduces the infinite case to a sequence of finite problems.

2.3. *Decomposition into error shells*

Since we are assuming $\mathcal{F}$ to be a finite class of functions, there are only a finite number of possible values that $\epsilon_{\text{gen}}(h)$ can assume. Let us name and order these possible *error values* $0 = \epsilon_1 < \epsilon_2 < \cdots < \epsilon_r \leq 1$. Thus, $r \leq |\mathcal{F}|$, and for each $1 \leq i \leq r$ there exists an $h_i \in \mathcal{F}$ such that $\epsilon_{\text{gen}}(h_i) = \epsilon_i$. Then for each index $1 \leq j \leq r$ we can define the cardinality of the jth error shell $Q_j = |\{f' \in \mathcal{F} : \epsilon_{\text{gen}}(f') = \epsilon_j\}|$. Thus Q_j is the *number* of functions in $\mathcal{F}$ whose generalization error is exactly ϵ_j, and $\sum_{j=1}^{r} Q_j = |\mathcal{F}|$. Hence we arrive at the *shell decomposition* of the union bound:

$$\sum_{h \in \overline{B(\epsilon_i)}} (1 - \epsilon_{\text{gen}}(h))^m = \sum_{j=i}^{r} Q_j (1 - \epsilon_j)^m \tag{5}$$

Together with Theorem 1, we can obtain the following bound on $\epsilon_{\text{gen}}(h)$ for consistent learning algorithms.

Theorem 3. *For any fixed sample size m and confidence value δ, with probability at least $1 - \delta$ any $h \in VS(S)$ obeys $\epsilon_{gen}(h) \leq \epsilon_i$, where ϵ_i is the smallest error value satisfying $\sum_{j=i}^{r} Q_j (1 - \epsilon_j)^m \leq \delta$.*

In other words, if we fix the confidence δ then Theorem 3 provides the bound

$$\epsilon_{\text{gen}}(h) \leq \min \left\{ \epsilon_i : \sum_{j=i}^{r} Q_j (1 - \epsilon_j)^m \leq \delta \right\} \tag{6}$$

with probability at least $1 - \delta$ for any consistent h. While this bound is clearly a function of m, its behavior is not especially easy to understand in its current form. For this we rely on a particular limit popular in the statistical mechanics literature known as the *thermodynamic limit.*

2.4. *The thermodynamic limit method*

There are two basic ideas or assumptions behind the thermodynamic limit method as we formalize it. The first idea is that we are often interested in the learning curve of a parametric class of functions, and in such cases the number of functions in the class at any given error value may have a limiting asymptotic behavior as the number of parameters becomes large. The second idea is to exploit this limiting behavior in order to describe learning curves as a competition between the logarithm of the number of functions at a given error value (an *entropy* term) and the error value itself (an *energy* term).

As we shall see, the most important step in applying the thermodynamic limit method, both technically and conceptually, is to find the right *scaling* with which to analyze the learning curve, and to find the best entropy bound for this scaling. The thermodynamic limit method assumes that an appropriate scaling and entropy bound are given, and then provides a learning curve analysis for them, much in the same way that VC theory assumes that the VC dimension is known and then provides learning curve upper bounds. Thus the real work of the user in applying the thermodynamic limit method (which may be considerable) lies in finding the best scaling and entropy bound.

In order to properly define and use the thermodynamic limit method, we cannot limit our attention to a fixed finite class $\mathcal{F}$ of functions, but must instead assume an infinite *sequence* of finite function classes (of presumably increasing but always finite cardinality). As we have already suggested, it will be convenient to think of this sequence as being obtained in some uniform manner by increasing the number of parameters in a parametric class of functions. Thus, let $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N, \ldots,$ be any infinite sequence of classes of functions, where each $\mathcal{F}_N$ is a class of boolean functions over an input space X_N and obeys $|\mathcal{F}_N| \leq 2^N$. We may think of N as just an abstract index obeying $N \geq \log|\mathcal{F}_N|$, and thus representing the number of bits or parameters required to encode functions in $\mathcal{F}_N$. Let D_N be a fixed probability distribution over X_N. A typical example of these objects is where we let X_N be N-dimensional Euclidean space, D_N be the uniform distribution over the unit sphere in X_N, and $\mathcal{F}_N$ be the class of all N-dimensional perceptrons in which each weight is constrained to be either 1 or -1.

Now suppose that for each class $\mathcal{F}_N$ we also choose a fixed target function $f_N \in \mathcal{F}_N$, thus yielding an infinite sequence of target functions $f_1, f_2, \ldots, f_N, \ldots$. Our goal now is to provide a framework in which we can analyze the limiting generalization error, as $N \rightarrow \infty$, of any algorithm that always chooses a hypothesis consistent with m random examples of f_N drawn according to D_N.

There are a number of problems with this proposal. Foremost among these is the question of whether there actually exists any interesting limiting behavior. For instance, in our discussion so far we have been suggesting that all the classes $\mathcal{F}_N$ are "similar" in the sense of being obtained through some nice uniform parametric process, with only the number of parameters varying. If this assumption is grossly violated, and each $\mathcal{F}_N$ looks radically different than the last, it may be nonsensical to analyze the limiting behavior of a consistent algorithm's error. Similarly, even if the $\mathcal{F}_N$ are generated in a uniform fashion, a highly nonuniform sequence of target functions f_N may render the limit meaningless.

There is no definitive solution to such obstacles: there do exist function class, distribution and target function sequences for which there is no limiting generalization error for consistent algorithms, and obviously no theory can assign a tight asymptotic limit in such cases. The thermodynamic limit method survives these problems by only providing an upper bound on the asymptotic generalization error. In those cases where the limit does not exist, this upper bound may be weak or even vacuous. However, we hope to show through examples that in many natural cases the limiting behavior is both well-defined and captured by our theory, and that the resulting upper bound correctly predicts learning curve behavior that is radically different from that predicted by more standard methods.

A second and more technical objection to our proposal is that if we *fix* a sample size m and let $N \to \infty$, we should not expect to obtain any nontrivial bound on the generalization error, since the function classes are becoming larger but the sample size remains fixed. This is exactly right, and for this reason the thermodynamic limit method examines the learning curve behavior as both $m \to \infty$ and $N \to \infty$, but at some *fixed rate*. This allows us to meaningfully investigate, for instance, the asymptotic generalization error when the number of examples is $1/2$ the number of parameters, twice the number of parameters, 10 times the number of parameters, and so on. This is frequently the language in which experimentalists discuss learning curves.

Returning to the development, once we fix target function sequence $f_N \in \mathcal{F}_N$, we can again define the error levels $0 = \epsilon_1^N < \epsilon_2^N < \cdots < \epsilon_{r(N)}^N \leq 1$ for $\mathcal{F}_N$ with respect to D_N, where $r(N) \leq |\mathcal{F}_N|$ is the number of error levels for this $\mathcal{F}_N$, D_N and f_N, and for clarity we have included a superscript on the error levels indicating N. Recall that by Theorem 3, we can reduce the problem of bounding the error of a hypothesis from $\mathcal{F}_N$ consistent with m examples of f_N drawn according to D_N to the problem of finding the smallest error level ϵ_i^N such that the right-hand sum in Eq. (6) is bounded by δ (where, in the thermodynamic limit, δ will go to 0). The first step of the thermodynamic limit method is to simply rewrite this sum in a more convenient but entirely equivalent exponential form:

$$\sum_{j=i}^{r(N)} Q_j^N \left(1 - \epsilon_j^N\right)^m = \sum_{j=i}^{r(N)} e^{\log Q_j^N + m \log(1 - \epsilon_j^N)}. \tag{7}$$

Notice that in each term of this sum, the exponent term $\log Q_j^N$ is positive, and the exponent term $m \log(1 - \epsilon_j^N)$ is negative. Thus, informally speaking, the contribution of the jth term in the sum is largely determined by the competition between these two quantities: if $\log Q_j^N \gg -m \log(1 - \epsilon_j^N)$ then the contribution of the jth term is large (and thus, to make the overall sum smaller than δ, we must eliminate terms by increasing i and consequently weakening our bound on the error), and if $\log Q_j^N \ll -m \log(1 - \epsilon_j^N)$ then the contribution of the jth term is negligible.

In particular, if the sample size m is such that $\log Q_j^N \gg -m \log(1 - \epsilon_j^N)$ for *all* j then we cannot give a nontrivial bound on the error, and if $\log Q_j^N \ll -m \log(1 - \epsilon_j^N)$ for all j, and $r(N)$ is not too large, then the error should be close to 0. Such cases are uninteresting. In general, the values of the sample size m for which it will be most interesting to analyze the learning curve are those for which there is some real competition between the $\log Q_j^N$ and the $-m\log(1-\epsilon_j^N)$. Thus we need to find the right *scale* at which to examine the learning curve. At the same time, we would like to replace the competition between these two discrete quantities by the competition between two continuous functions of a single real parameter ϵ. The obvious choice for a continuous approximation to the $-m\log(1-\epsilon_j^N)$ is simply $m\log(1-\epsilon)$. The choice of a continuous approximation to the $\log Q_j^N$ depends on their behavior, which may be quite complex, and which we now try to capture.

Thus the next and crucial step of the thermodynamic limit method is to choose the appropriate *scaling function* and to provide an associated *entropy bound.* As mentioned already, these are functions that are assumed to be given in the thermodynamic limit method. Let $t(N)$ be any mapping from the natural numbers to the natural numbers such that $t(N) \to \infty$ as $N \to \infty$, and let $s : [0,1] \to \Re^+$ be any continuous function. Then we say that $s(\epsilon)$ is *a permissible entropy bound with respect to* $t(N)$ if there exists a natural number N_0 such that for all $N \geq N_0$ and for all $1 \leq j \leq r(N)$, $(1/t(N))\log Q_j^N \leq s(\epsilon_j^N)$.

We refer to $t(N)$ as a *scaling function.* The intention is that when $t(N)$ is properly chosen it captures the scale at which the learning curve is most interesting, and that the entropy bound $s(\epsilon)$ tightly captures the behavior of the $(1/t(N))\log Q_j^N$. We will see that we obtain our best upper bounds on generalization error for a given scaling function when the thermodynamic limit method is used with the smallest possible permissible entropy bound for this scaling function.

Given a scaling function $t(N)$ and a permissible entropy bound $s(\epsilon)$, for $N \geq N_0$ we may now rewrite and bound our sum:

$$\sum_{j=i}^{r(N)} e^{\log Q_j^N + m\log(1-\epsilon_j^N)} \tag{8}$$

$$= \sum_{j=i}^{r(N)} e^{t(N)[(1/t(N))\log Q_j^N + (m/t(N))\log(1-\epsilon_j^N)]} \tag{9}$$

$$\leq \sum_{j=i}^{r(N)} e^{t(N)[s(\epsilon_j^N) + \alpha\log(1-\epsilon_j^N)]} \tag{10}$$

where we define $\alpha = m/t(N)$, and in taking our limit $m, N \to \infty$, α will remain constant. Before doing so, however, let us pause to notice the benefits of our definitions in the final summation: each exponent's dependence on N has been isolated in the factor $t(N)$, and the remaining factor is the continuous function $s(\epsilon) + \alpha\log(1-\epsilon)$, evaluated at only the discrete points ϵ_j^N.

Let us now let $m, N \to \infty$ (and thus $t(N) \to \infty$) but let $m/t(N) = \alpha > 0$ remain constant. Define $\epsilon^* \in [0,1]$ to be the largest $\epsilon \in [0,1]$ such that $s(\epsilon) \geq -\alpha\log(1-\epsilon)$. Note that both $s(\epsilon)$ and $-\alpha\log(1-\epsilon)$ are non-negative functions, and $0 = -\alpha\log(1-\epsilon) \leq s(\epsilon)$ for $\epsilon = 0$. Thus ϵ^* is simply the rightmost crossing point of these functions (we define $\epsilon^* = 1$ if $s(\epsilon)$ stays above $-\alpha\log(1-\epsilon)$ for all $0 \leq \epsilon < 1$). We wish to argue that provided we examine our sum only for terms in which $\epsilon > \epsilon^*$, then under certain conditions the thermodynamic limit of the sum is 0. In other words, in the thermodynamic limit we can bound the generalization error of any consistent hypothesis by ϵ^*. Intuitively, the reason for this is that if $s(\epsilon) < -\alpha \log(1-\epsilon)$ then $e^{t(N)[s(\epsilon)+\alpha\log(1-\epsilon)]} \to 0$ as $t(N) \to \infty$.

More precisely, let $\tau \in (0, 1]$ be an arbitrarily small quantity, and for each N, define the index $i_{N,\tau}$ to be the smallest satisfying $\epsilon^N_{i_{N,\tau}} \geq \epsilon^* + \tau$. Let us define Δ by

$$\Delta = \min\{-\alpha \log(1-\epsilon) - s(\epsilon) : \epsilon \in [\epsilon^* + \tau, 1]\}. \tag{11}$$

Note that Δ is well-defined since the quantify

$$-\alpha \log(1-\epsilon) - s(\epsilon)$$

is strictly positive for all $\epsilon \in [\epsilon^* + \tau, 1]$. We can now write

$$\begin{aligned}
&\sum_{j=i_{N,\tau}}^{r(N)} e^{t(N)[s(\epsilon^N_j)+\alpha\log(1-\epsilon^N_j)]} && (12)\\
&\quad \leq \sum_{j=i_{N,\tau}}^{r(N)} e^{-t(N)\Delta} && (13)\\
&\quad \leq (r(N) - i_{N,\tau}) e^{-t(N)\Delta} && (14)\\
&\quad \leq r(N) e^{-t(N)\Delta} && (15)
\end{aligned}$$

where the first inequality follows from the fact that for all $i_{N,\tau} \leq j \leq r(N)$ we have $\epsilon^N_j \in [\epsilon^* + \tau, 1]$. The expression $r(N)e^{-t(N)\Delta}$ will go to 0 in the thermodynamic limit, as desired, provided $r(N)$ is $o(e^{t(N)\Delta})$ (this condition is easily met by all of the examples we shall analyze, but for completeness its relaxation is discussed in the Appendix in Section A.1).

We have shown:

Theorem 4. *Let $s(\epsilon)$ be any continuous function that is a permissible entropy bound with respect to the scaling function $t(N)$, and suppose that $r(N) = o(e^{t(N)\Delta})$ for any positive constant Δ. Then as $m, N \to \infty$ but $\alpha = m/t(N)$ remains constant, for any positive τ we have*

$$\mathbf{Pr}_S[VS(S) \subseteq B(\epsilon^* + \tau)] \to 1. \tag{16}$$

Here the probability is taken over all samples S of size $m = \alpha t(N)$ for the target function in $f \in \mathcal{F}_N$. and ϵ^* is the rightmost crossing point of $s(\epsilon)$ and $-\alpha\log(1-\epsilon)$. In other words, in the thermodynamic limit any hypothesis h consistent with $\alpha t(N)$ examples will have generalization error $\epsilon_{gen}(h) \leq \epsilon^* + \tau$ with probability 1.

We can finally see in Theorem 4 the roles of the scaling function $t(N)$ and the entropy bound $s(\epsilon)$. The scaling function $t(N)$ defines the units by which we shall measure learning

Theorem 5. *Let* $s : [0, 1/2] \to [0, 1]$ *be any continuous function bounded away from* 1 *and such that* $s(0) = s(1) = 0$. *Then there exists a function class sequence* $\mathcal{F}_N$ *over* X_N *(where* $|\mathcal{F}_N| = 2^N$*), a distribution sequence* D_N *over* X_N*, and a target function sequence* $f_N \in \mathcal{F}_N$ *such that:* (1) $s(\epsilon)$ *is a permissible entropy bound with respect to the scaling function* $t(N) = N$, *and* (2) *For any* $\alpha > 0$, *if* $\epsilon^* \in [0, 1/2]$ *is the largest value satisfying* $2\alpha\epsilon^* \geq s(\epsilon^*)$, *then as* $N \to \infty$ *there is constant probability that there exists a function* $h \in \mathcal{F}_N$ *consistent with* $m = \alpha N$ *random examples satisfying* $\epsilon_{gen}(h) \geq \epsilon^*$.

Proof: (Sketch) For every N, the class $\mathcal{F}_N$ will contain the function f_N which is identically 0 on all inputs. For the lower bound argument, for every value of N, f_N will always be the target function against which we measure generalization error. The distribution D_N will always be uniform over the domain X_N, which will always consist of 2^N discrete points, so $X_N = \{1, 2, \ldots, 2^N\}$.

A high-level sketch of the main ideas follows. For any N, the class $\mathcal{F}_N$ will be constructed so that there are exactly $N/2$ error levels, namely $\epsilon_j^N = j/N$ for $1 \leq j \leq N/2$. Now let $s : [0, 1/2] \to [0, 1]$ be *any* continuous function bounded away from 1 and satisfying $s(0) = s(1/2) = 0$. The idea is that for any N and any $1 \leq j \leq N/2$, $\mathcal{F}_N$ will contain exactly $2^{s(j/N)\cdot N}$ functions whose error with respect to f_N is j/N. Thus, for any ϵ, as $N \to \infty$, there will eventually be arbitrarily close to $2^{s(\epsilon)\cdot N}$ functions of error arbitrarily close to ϵ. This ensures that $s(\epsilon)$ will be a permissible entropy bound with respect to the scaling function $t(N) = N$. Furthermore, these functions will be specially chosen to force the claimed lower bound.

In more detail, for every N and every $1 \leq j \leq N/2$, $\mathcal{F}_N$ will contain a subclass of functions $\mathcal{F}_N^j$, where $|\mathcal{F}_N^j| = 2^{s(j/N)\cdot N}$. Note that this implies $|\mathcal{F}_N| < (N/2)2^N$ since $s(\epsilon) < 1$. For every $h \in \mathcal{F}_N^j$ and every $(2j/N)2^N < x \leq 2^N$, $h(x) = 0$. In other words, on a fraction $1 - (2j/N)$ of the input space, all the $h \in \mathcal{F}_N^j$ agree with the target function f_N.

However, on the points $\{1, 2, \ldots, (2j/N)2^N\}$ each $h \in \mathcal{F}_N^j$ will behave as a unique parity function on a domain of size $(2j/N)2^N$. More precisely, we can define an isomorphism between $\{1, 2, \ldots, (2i/N)2^N\}$ and the hypercube of the same size, and let each function in $\mathcal{F}_N^j$ (when restricted to $\{1, 2, \ldots, (2j/N)2^N\}$) be isomorphic to a unique parity function on this hypercube. (Note that $s(\epsilon)$ must obey $2^{s(\epsilon)\cdot N} \leq 2\epsilon \cdot 2^N$ in order to ensure there are enough unique parity functions. The condition $s(\epsilon) < 1$ is sufficient to give this asymptotically.) Thus, each $h \in \mathcal{F}_N^j$ has $\epsilon_{\text{gen}}(h) = j/N$ since each parity function outputs 1 on half of the hypercube inputs and f_N is identically 0.

Now let us analyze, in the thermodynamic limit, the largest generalization error of any function in the version space of the constructed family $\mathcal{F}_N$ (for target functions f_N and uniform distributions D_N). By our construction, for any ϵ, as $N \to \infty$ there are eventually $2^{s(\epsilon)\cdot N}$ functions in $\mathcal{F}_N$ of generalization error arbitrarily close to ϵ (namely, $\epsilon \pm 1/N$). Let the sample size $m = \alpha N$. As $N \to \infty$, the number of sample points falling in the set $\{1, 2, \ldots, 2\epsilon \cdot 2^N\}$ becomes sharply peaked at $(2\epsilon)\alpha N$. The remaining sample points fail to eliminate any of the functions of generalization error ϵ since they all agree with the target function f_N on the remaining points.

Now it is known (Goldman, Kearns, & Schapire, 1990) that in order to eliminate $2^{s(\epsilon)\cdot N}$ parity functions over a uniform distribution, the sample size m must obey $m \geq s(\epsilon) \cdot N$;

for smaller m, there is a constant probability that at least one parity function remains in the version space. Thus, we obtain that if $(2\epsilon)\alpha N \leq s(\epsilon)N$ then there is constant probability that the version space contains a function of generalization error at least ϵ. In other words, $2\alpha\epsilon \geq s(\epsilon)$ is a condition for eliminating all functions of generalization error ϵ from the version space, thus proving the theorem. □

3. The finite and unrealizable case

One highly restrictive aspect of all of our analysis so far is the assumption that the labels of the examples are generated by some target function in $\mathcal{F}$, and hence it is always possible to obtain zero generalization error. We now consider the relaxation of this restriction to the case where there may exist no function in $\mathcal{F}$ with zero generalization error. We call this case the *unrealizable* target case. This actually covers two cases. In the first, the labels of the examples are generated by some target function that is not in $\mathcal{F}$. In the second, and more general case, each labeled example $\langle x_i, y_i\rangle$ in S, $1 \leq i \leq m$ is generated independently according to a distribution D_N on $X_N \times \{0, 1\}$, which plays the role that was played jointly by the distribution D_N and the target function in the realizable case. Here D_N can model noise in the examples as well. We pursue this second, more general case here.

In analogy with the realizable case, for any function $h \in \mathcal{F}_N$, $\epsilon_{\text{gen}}(h) = \mathbf{Pr}_{\langle x,y\rangle \in D_N} [h(x) \neq y]$. For simplicity we will assume that there is a unique best hypothesis in $\mathcal{F}_N$

$$h^* = \operatorname*{argmin}_{h\in\mathcal{F}} \epsilon_{\text{gen}}(h), \tag{24}$$

although it is easy to generalize the arguments to handle cases where there is a tie. (Since $\mathcal{F}_N$ is finite, we need not worry about there being an infinite sequence of better and better hypothesis, with no best hypothesis in $\mathcal{F}_N$.) Our goal in this section is to analyze the learning curve for this unrealizable case in the same manner as for the realizable case, providing a thermodynamic limit method and extracting scaled learning curves. Of course, now the learning curve approaches $\epsilon_{\min} = \epsilon_{\text{gen}}(h^*)$ rather than 0 as the number of examples is increased. We shall see that interesting technical differences from the realizable case are also forced upon us in the analysis.

Recall that in the realizable case, we focused on bounding the error of any consistent algorithm. In the unrealizable case, we analyze an empirical error minimization algorithm. We define the *training error* or *empirical error* of a hypothesis h to be the frequency of disagreement on a sample S:

$$\epsilon_{\text{trn}}(h, S) = \frac{1}{m}\sum_{i=1}^{m} \chi[h(x_i) \neq y_i] \tag{25}$$

where the indicator function χ is 1 when its argument is true and zero otherwise. An empirical error minimization algorithm chooses a hypothesis from the version space, which we now redefine to be the set of all functions that minimize the training error $\epsilon_{\text{trn}}(h, S)$:

$$\text{VS}(S) = \left\{h \in \mathcal{F} : \epsilon_{\text{trn}}(h, S) = \min_{h'\in\mathcal{F}} \epsilon_{\text{trn}}(h', S)\right\}. \tag{26}$$

3.1. Energy functions

One of the main differences between the unrealizable and realizable cases is the form of the bound we can obtain on the probability that a fixed function $h \in \mathcal{F}$ "survives" m random examples, that is, remains in the version space and hence is eligible to be chosen by an empirical error minimization algorithm. Recall that in the realizable case, this probability was exactly $(1 - \epsilon_{\text{gen}}(h))^m$ since $\epsilon_{\min} = 0$ and minimum empirical error is equivalent to consistency. In the unrealizable case, the situation is more complicated: we will only be able to upper bound this survival probability. Unlike the realizable case, where the exact expression $(1 - \epsilon_{\text{gen}}(h))^m$ for the survival probability was eventually translated in the thermodynamic limit method to a function $-\alpha \log(1 - \epsilon)$ in the exponent that was *universal* for all problems (the specifics of the problem affecting only the scaling function and entropy bound), in the unrealizable case we may sometimes need to use energy bounds that depend on the problem specifics. Furthermore, the quality of bound we use can have significant effects on the behavior of the resulting scaled learning curve, especially in the large α limit.

We will treat this bound on the survival probability as a parameter of the analysis. More precisely, let us refer to a function $u(\epsilon)$ as a *permissible energy bound* (with respect to $\mathcal{F}$, D and the target function) if for any $h \in \mathcal{F}$ and any sample size m we may write

$$\mathbf{Pr}_S[h \in \text{VS}(S)] \leq e^{-u(\epsilon_{\text{gen}}(h))m}. \tag{27}$$

In other words, we imagine that $u(\epsilon_{\text{gen}}(h))$ assesses a penalty to $\epsilon_{\text{gen}}(h)$ that increases with larger $\epsilon_{\text{gen}}(h)$, and the probability that h survives to be in the version space (and thus the probability that an empirical minimization algorithm may choose h) decreases exponentially in m times this penalty.

Permissible energy bounds will all be derived from the following chain of inequalities:

$$\mathbf{Pr}_S[h \in \text{VS}(S)] \tag{28}$$

$$\leq \mathbf{Pr}_S[\epsilon_{\text{trn}}(h, S) \leq \epsilon_{\text{trn}}(h^*, S)] \tag{29}$$

$$\leq \left[1 - \epsilon(h, h^*) + \sqrt{\epsilon(h, h^*)^2 - (\epsilon_{\text{gen}}(h) - \epsilon_{\min})^2}\right]^m \tag{30}$$

where $\epsilon(h_1, h_2)$ is the probability of disagreement between h_1 and h_2 on the label of a random example drawn according to D_N. The first inequality follows from the fact that the training error of any hypothesis h in the version space must be no greater than the training error of any other hypothesis in the class, including h^* in particular. The second follows from Sanov's theorem on large deviations (Cover & Thomas, 1991) (see Section A.2 of the Appendix).

For the realizable case we have $\epsilon_{\min} = 0$ and $\epsilon(h, h^*) = \epsilon_{\text{gen}}(h)$, so $\mathbf{Pr}_S[h \in \text{VS}(S)] \leq (1 - \epsilon_{\text{gen}}(h))^m$ already follows from the second inequality. To obtain an energy bound in the unrealizable case, we must somehow relate $\epsilon(h, h^*)$ to $\epsilon_{\text{gen}}(h)$. If $v(\epsilon)$ is a function that satisfies

$$\epsilon(h, h^*) \leq v(\epsilon_{\text{gen}}(h)) \tag{31}$$

then from Eq. (30)

$$u(\epsilon) = -\ln\left(1 - v(\epsilon) + \sqrt{v^2(\epsilon) - (\epsilon - \epsilon_{\min})^2}\right) \tag{32}$$

is a permissible energy bound. In our theory, learning curves are determined by the competition between energy and entropy, with the best bounds being obtained for the largest energy bound (which corresponds to the most rapidly decaying bound on the survival probability as a function of m). For this reason, we see that smaller $v(\epsilon)$ is, the better the resulting energy bound. Now by the triangle inequality, we can always find $v(\epsilon)$ such that $\epsilon - \epsilon_{\min} \leq v(\epsilon) \leq \min\{\epsilon + \epsilon_{\min}, 1\}$, and cannot find a smaller $v(\epsilon)$. Since the choice $v(\epsilon) = \epsilon + \epsilon_{\min}$ is always possible, plugging this into Eq. (32) gives a universally permissible energy bound. After a little algebra, this bound reduces to

$$u(\epsilon) = -\ln\left(1 - (\sqrt{\epsilon} - \sqrt{\epsilon_{\min}})^2\right) \tag{33}$$

However, better $v(\epsilon)$ may be obtained in certain cases. For instance, if we are fortunate enough to have $v(\epsilon) = \epsilon - \epsilon_{\min}$ for some problem, then $u(\epsilon) = -\ln(1 - \epsilon + \epsilon_{\min})$ is a permissible energy bound, which is essentially linear in ϵ and thus nearly the same as for the realizable case. We now sketch the technical development for the unrealizable case using a generic permissible energy bound $u(\epsilon)$, occasionally pointing out the effects of specific energy bounds on learning curves. We examine these effects more closely in Section 3.5.

3.2. *Technical development for the unrealizable case*

As was done for the realizable case in Section 2.1, we can write a union bound on the probability that $\mathrm{VS}(S)$ is contained in $B(\epsilon)$. This enables us to bound the error of all empirical error minimization algorithms. For with confidence $\mathbf{Pr}_S[\mathrm{VS}(S) \subseteq B(\epsilon)]$, we can assert that the hypothesis with minimal training error has generalization error less than ϵ.

Let $\epsilon > \epsilon_{\min}$ be given. Then any permissible energy bound $u(\epsilon)$ can be used to lower bound the probability that every function outside $B(\epsilon)$ has training error larger than the training error of h^*:

Theorem 6. *Let $u(\epsilon)$ be a permissible energy bound. Then $\mathbf{Pr}_S[VS(S) \subseteq B(\epsilon)] \geq 1 - \delta$, where*

$$\delta = \sum_{h \in \overline{B(\epsilon)}} e^{-u(\epsilon_{gen}(h))m} \tag{34}$$

Theorem 1 is a special case with $u(\epsilon) = -\log(1 - \epsilon)$.

With the universally permissible energy function $u(\epsilon) = -\ln(1 - (\sqrt{\epsilon} - \sqrt{\epsilon_{\min}})^2)$, the standard cardinality bound becomes

$$\sum_{h \in \overline{B(\epsilon)}} e^{-u(\epsilon_{\text{gen}}(h))m} \leq |\mathcal{F}|\left(1 - (\sqrt{\epsilon} - \sqrt{\epsilon_{\min}})^2\right)^m \tag{35}$$

$$\leq |\mathcal{F}| e^{-(\sqrt{\epsilon} - \sqrt{\epsilon_{\min}})^2 m} \tag{36}$$

because $\epsilon_{\text{gen}}(h) > \epsilon$ for all $h \in \overline{B(\epsilon)}$. Setting the latter quantity to δ and solving for ϵ yields

$$\epsilon = \epsilon_{\min} + 2\sqrt{\frac{\epsilon_{\min}\ln(|\mathcal{F}|/\delta)}{m}} + \frac{\ln(|\mathcal{F}|/\delta)}{m}. \tag{37}$$

Hence in analogy with Section 2.2 for the realizable case, it follows that for any empirical error minimization algorithm, with confidence $1-\delta$ the hypothesis h it produces satisfies

$$\epsilon_{\text{gen}}(h) \leq \epsilon_{\min} + 2\sqrt{\frac{\epsilon_{\min}\ln(|\mathcal{F}|/\delta)}{m}} + \frac{\ln(|\mathcal{F}|/\delta)}{m}, \tag{38}$$

giving the same bound we obtained in the realizable case when $\epsilon_{\min} = 0$.

This worst case bound already has some interesting behavior in the thermodynamic limit. To see this, let assume that $\mathcal{F}_N = 2^N$, as large as we allow, and further that the best entropy function that we can obtain is the trivial function $s(\epsilon) = 1$. Let $t(N) = N$. Then $\ln|\mathcal{F}_N|/m = 1/\alpha$. Hence, from Eq. (38), in the thermodynamic limit we obtain the scaled learning curve

$$\epsilon - \epsilon_{\min} \leq 2\sqrt{\frac{\epsilon_{\min}}{\alpha}} + \frac{1}{\alpha}. \tag{39}$$

This curve exhibits a faster learning rate, scaling roughly like $1/\alpha$ in the early stages of learning, until $\alpha \approx 1/4\epsilon_{\min}$, the point at which both terms in the bound are equal, then it begins to scale more like $2\sqrt{\epsilon_{\min}/\alpha}$ as α gets larger and the first term in the bound begins to dominate. This behavior has also been noted by Vapnik (1982).

Returning to the general development, just as in the realizable case we can refine the union bound of Theorem 6 via a shell decomposition. Still more improvement may come from finding a better energy function of the form in Eq. (32). Addressing the first improvement, just as in the realizable case in Section 2.3, we proceed to slice the function class into error shells. Let $\epsilon_{\min} = \epsilon_1 < \epsilon_2 < \cdots < \epsilon_r$ be all of the possible values for the generalization error for functions in $\mathcal{F}$, and let Q_i be the number of functions $h \in \mathcal{F}$ satisfying $\epsilon_{\text{gen}}(h) = \epsilon_i$. The analog of Theorem 3 in the unrealizable case is:

Theorem 7. *Let $u(\epsilon)$ be a permissible energy bound. Then for any fixed sample size m and confidence value δ, with probability at least $1-\delta$ any $h \in VS(S)$ obeys $\epsilon_{gen}(h) \leq \epsilon_i$, where $\epsilon_i \geq \epsilon_{\min}$ is the smallest error level satisfying*

$$\sum_{j=i}^{r} Q_j e^{-u(\epsilon_j)m} \leq \delta. \tag{40}$$

In other words, for any δ we may write

$$\epsilon_{\text{gen}}(h) \leq \min\left\{\epsilon_i : \sum_{j=i}^{r} Q_j e^{-u(\epsilon_j)m} \leq \delta\right\} \tag{41}$$

with probability at least $1 - \delta$. Thus we have a bound on $\epsilon_{\text{gen}}(h)$ that implicitly depends on m, but as in the realizable case, this bound is more easily understood in a thermodynamic limit.

Towards this goal, in analogy with Section 2.4 for the realizable case, we again can rewrite the summation obtained by shell decomposition in a convenient exponential form.

$$\sum_{j=i}^{r} Q_j e^{-u(\epsilon_j)m} \tag{42}$$

$$= \sum_{j=i}^{r} e^{\log Q_j - u(\epsilon_j)m} \tag{43}$$

$$= \sum_{j=i}^{r} e^{t(N)[(1/t(N)) \log Q_j - (m/t(N))u(\epsilon_j)]} \tag{44}$$

where $t(N)$ is a scaling function of our choice. Thus we see that in the unrealizable case, the bound on generalization error again involves a competition between the entropic expression $(1/t(N)) \log Q_j$ and the energetic expression $(m/t(N))u(\epsilon_j)$. Using the same definition of the permissible entropy function $s(\epsilon)$ as in the realizable case, we obtain the following theorem, whose proof is entirely analogous to the realizable setting.

Theorem 8. *Let $u(\epsilon)$ be a permissible energy bound. Let $s(\epsilon)$ be any continuous function that is a permissible entropy bound with respect to the scaling function $t(N)$, and suppose that $r(N) = o(e^{t(N)\Delta})$ for any positive constant Δ. Then as $m, N \to \infty$ but $\alpha = m/t(N)$ remains constant, for any positive τ we have*

$$\mathbf{Pr}_S[VS(S) \subseteq B(\epsilon^* + \tau)] \to 1. \tag{45}$$

Here the probability is taken over all samples S of size $m = \alpha t(N)$, where each example is drawn independently according to D_N, and ϵ^* is the rightmost crossing point of $s(\epsilon)$ and $\alpha u(\epsilon)$. In other words, in the thermodynamic limit any hypothesis h with the minimum number (over $\mathcal{F}$) of observed disagreements on the $\alpha t(N)$ examples will have generalization error $\epsilon_{gen}(h) \leq \epsilon^* + \tau$ with probability 1.

Just as in the realizable case, Theorem 8 allows us to extract scaled learning curves that express generalization error as a function of α. It is also easily verified that the thermodynamic limit lower bound of Theorem 5 translates unchanged to the unrealizable setting.

In summary, for the unrealizable case in the thermodynamic limit, the generalization error can be upper bounded by the rightmost crossing of $s(\epsilon)$ and a competing energy function of the form in Eq. (32) times α. Thus the basic theory derived for the realizable case survives relatively nicely. Furthermore, we will shortly see that while the overall picture is described by this competition, slight changes to simple models of unrealizability can yield important changes to $s(\epsilon)$ and the energy function, and thus to the resulting learning curve.

3.3. Analysis of an unrealizable Ising perceptron

We now illustrate the use of the thermodynamic limit method in the unrealizable case by considering an unrealizable variant of the Ising perceptron problem considered in Section 2.6. Let the target function f_N be the perceptron in which every weight is $+1$, and let the function class $\mathcal{F}_N$ consist of all Ising perceptrons which have *at least* γN weights ($\gamma \in [0, 1]$) that are -1. (Note that unlike the realizable Ising perceptron case, here the choice of target function matters.) Again let the distribution D_N be any spherically symmetric distribution on $\Re^N$. Thus, the target function is not contained in $\mathcal{F}_N$, and the minimum error $\epsilon_{\min}(\gamma)$ is given by applying Eq. (17), so $\epsilon_{\min}(\gamma) = (1/\pi)\cos^{-1}(1-2\gamma)$. This minimum error is achieved by all of those functions in $\mathcal{F}_N$ with the minimum allowed number γN of -1 weights, of which there are exactly $\binom{N}{\gamma N}$. We shall regard γ as a parameter measuring the extent of the unrealizability.

The correct scaling function for this problem is again $t(N) = N$, and it is easy to see the effects of the unrealizability parameter γ on this problem. The resulting permissible entropy bound $s_\gamma(\epsilon)$ is identically 0 in the range $[0, \epsilon_{\min}(\gamma)]$, as there are no functions in $\mathcal{F}_N$ at these generalization errors. In the range $[0, \epsilon_{\min}(\gamma)]$, however, $s_\gamma(\epsilon) = s(\epsilon)$, where $s(\epsilon)$ is simply the entropy bound for the realizable Ising perceptron given by Eq. (19). Thus our entropy bound in the unrealizable case is simply that of the realizable case, but truncated to the left of $\epsilon_{\min}(\gamma)$.

The effects of this truncation on the predicted scaled learning as a function of γ turn out to be quite interesting. If we use the universally permissible energy bound given by Eq. (32) then figures 12, 13 and 14 show the resulting entropy-energy competition for three different degrees of unrealizability (that is, three values of $\epsilon_{\min}(\gamma)$) by plotting $s(\epsilon) - \alpha u(\epsilon)$. In each case of $\epsilon_{\min}(\gamma)$, we plot $s(\epsilon) - \alpha u(\epsilon)$ for three different values of α. When $\epsilon_{\min}(\gamma)$ is small (thus, the target function is nearly realized by the function class), the behavior is quite similar to that of the realizable case in figure 11. By the time $\epsilon_{\min}(\gamma)$ is as large as

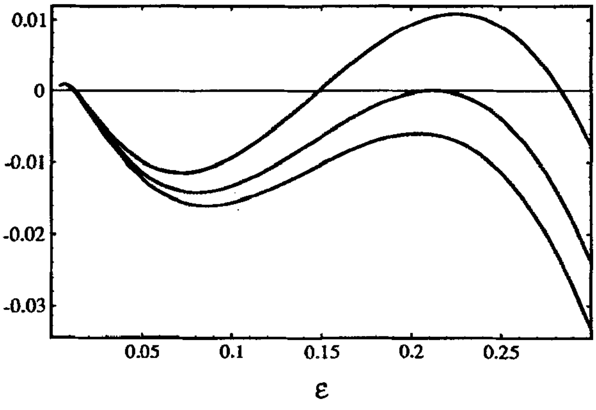

Figure 12. The function $s(\epsilon) - \alpha u(\epsilon)$ for the unrealizable Ising perceptron discussed in Section 3.3, with $\epsilon_{\min}(\gamma) = 0.005$. The function is plotted for the values $\alpha = 2.0, 2.063, 2.1$ (top to bottom).

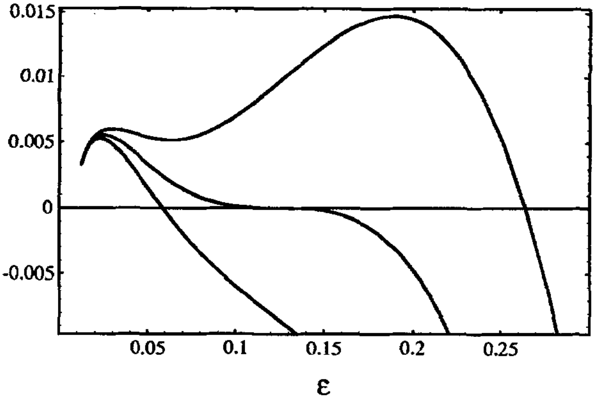

Figure 13. The function $s(\epsilon) - \alpha u(\epsilon)$ for the unrealizable Ising perceptron discussed in Section 3.3, with $\epsilon_{\min}(\gamma) = 0.01224$. This value for $\epsilon_{\min}(\gamma)$ is a critical value, in the sense that the learning curve phase transition disappears for larger $\epsilon_{\min}(\gamma)$. The function is plotted for the values $\alpha = 2.5, 2.659, 2.8$ (top to bottom).

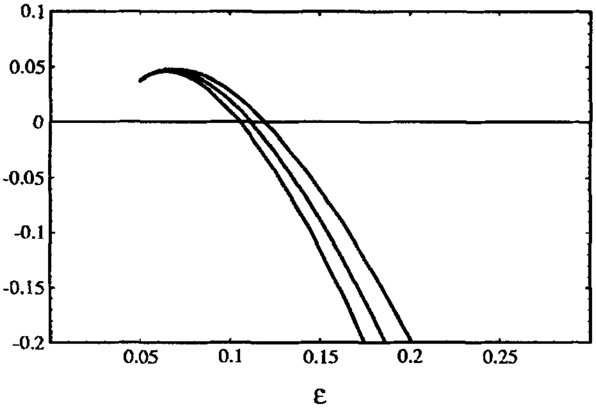

Figure 14. The function $s(\epsilon) - \alpha u(\epsilon)$ for the unrealizable Ising perceptron discussed in Section 3.3, with $\epsilon_{\min}(\gamma) = 0.05$. The function is plotted for the values $\alpha = 10, 11, 12$ (top to bottom).

0.05 in figure 14, we can see that the leftward progress of the zero crossing as α increases is quite uniform—the unrealizability has thus erased all traces of a phase transition. The intermediate value $\epsilon_{\min}(\gamma) = 0.01224$ is the boundary between these two behaviors: for smaller $\epsilon_{\min}(\gamma)$, the resulting learning curve will still exhibit some phase transition, while for larger $\epsilon_{\min}(\gamma)$, the transition is erased (although there may still be some trace of a phase transition in the form of accelerated generalization). This can all be clearly seen in figure 15, which shows the resulting scaled learning curves for these values of $\epsilon_{\min}(\gamma)$. Thus we see that the increase of γ not only increases the best error $\epsilon_{\min}(\gamma)$, it affects the very form of the learning curve. In particular, as γ increases the asymptotic rate of approach to $\epsilon_{\min}(\gamma)$ becomes slower. Figure 16 shows a *phase diagram* that plots the critical value of α for

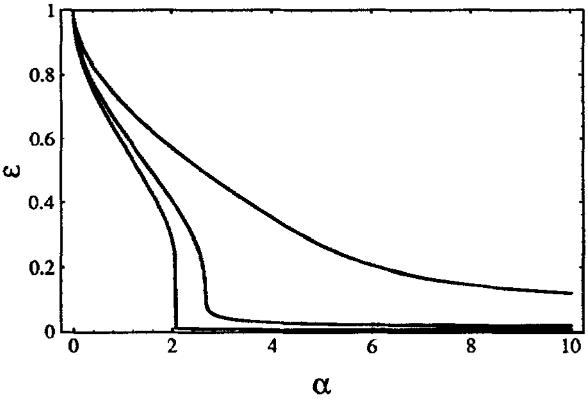

Figure 15. The scaled learning curves $\epsilon_\gamma^*(\alpha)$ for the unrealizable Ising perceptron discussed in Section 3.3, for the three values $\epsilon_{\min}(\gamma) = 0.005, 0.01224, 0.05$ (bottom to top).

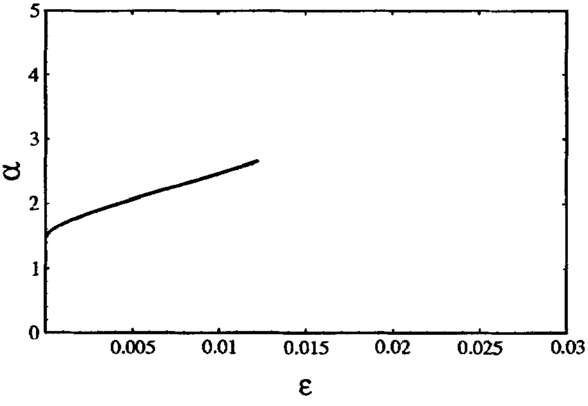

Figure 16. Phase diagram showing line of first-order transitions beginning at $\alpha = 1.448$ for $\epsilon_{\min}(\gamma) = 0$ and terminating at $\alpha = 2.659$ for $\epsilon_{\min}(\gamma) = 0.01224$.

which the learning curve experiences a phase transition as a function of $\epsilon_{\min}(\gamma)$—thus, as we have already mentioned, no value is plotted for $\epsilon_{\min}(\gamma) > 0.01224$ since no phase transition occurs in this case.

3.4. Analysis of the Ising perceptron with input noise

Here we consider the case when D_N is obtained by applying a target function consisting of an Ising perceptron $\mathbf{w}^*$ to inputs corrupted by additive Gaussian noise ξ. Thus in a random training example $\langle \mathbf{x}, y\rangle$ from D_N,

$$y = f(\mathbf{x}, \xi) = \text{sgn}(\mathbf{w}^* \cdot (\mathbf{x} + \xi)). \tag{46}$$

The distribution of inputs **x** is Gaussian, with unit variance on each component. The distribution of noise ξ is also Gaussian, with variance $\gamma^2 - 1$ on each component. A similar problem was examined by Györgyi and Tishby (1990).

In this case, one can show that

$$\epsilon_{\text{gen}}(\mathbf{w}) = \frac{1}{\pi}\cos^{-1}(R/\gamma) \tag{47}$$

$$\epsilon_{\min}(\gamma) = \epsilon_{\text{gen}}(\mathbf{w}^*) = \frac{1}{\pi}\cos^{-1}(1/\gamma) \tag{48}$$

$$\epsilon_{\text{gen}}(\mathbf{w}, \mathbf{w}^*) = \frac{1}{\pi}\cos^{-1} R \tag{49}$$

where $R = \mathbf{w} \cdot \mathbf{w}^*/N$.

The entropy function takes the form

$$s_\gamma(\epsilon) = \mathcal{H}((1 - \cos\pi\epsilon/\cos\pi\epsilon_{\min}(\gamma))/2). \tag{50}$$

To derive the energy function, we use

$$v_\gamma(\epsilon) = \frac{1}{\pi}\cos^{-1}(\cos\pi\epsilon/\cos\pi\epsilon_{\min}(\gamma)) \tag{51}$$

and plug into Eq. (32) to obtain $u_\gamma(\epsilon)$. Our error bound is then the rightmost solution of $s_\gamma(\epsilon) = \alpha u_\gamma(\epsilon)$. The entropy $s_\gamma(\epsilon)$ is a single hump, as in the zero noise case. However, the edges of the hump are at $\epsilon = \epsilon_{\min}(\gamma)$ and $\epsilon = 1 - \epsilon_{\min}(\gamma)$, outside of which the entropy is zero. At the edges, the entropy rises like $\Delta\epsilon \log \Delta\epsilon$ (where $\Delta\epsilon = \epsilon - \epsilon_{\min}(\gamma)$), and thus has infinite slope. In contrast the energy has zero slope, since it behaves like $(\Delta\epsilon)^{3/2}$. Hence the asymptotic behavior must be

$$\epsilon - \epsilon_{\min}(\gamma) = O\left(\frac{\log\alpha}{\alpha}\right)^2 \tag{52}$$

However, the large α asymptotics are not the whole story. For $\epsilon_{\min}(\gamma) < 0.01969$, the error bound undergoes a first order transition to nonzero error. In other words, although the input noise prevents a transition to perfect learning, when it is small it does not erase all traces of the transition.

Plots of $s(\epsilon) - \alpha u(\epsilon)$ for three different values of $\epsilon_{\min}(\gamma)$ are given in figures 17, 18 and 19, and the corresponding learning curves in figure 20. The phase diagram indicating the critical value of α for each value of $\epsilon_{\min}(\gamma)$ is plotted in figure 21.

As an illuminating exercise, we note that four different bounds can be written using the tools of this paper. For the entropy there are two choices, the simple cardinality bound $s(\epsilon) = 1$ and the tighter bound above. For the energy there are two choices, given by Eqs. (32) and (33), corresponding to the choices of $v(\epsilon)$ as above and $v(\epsilon) = \epsilon + \epsilon_{\min}$.