Rigorous Learning Curve Bounds from Statistical Mechanics

David Haussler
U.C. Santa Cruz
Santa Cruz, California

Michael Kearns
AT&T Bell Laboratories
Murray Hill, New Jersey

H. Sebastian Seung
AT&T Bell Laboratories
Murray Hill, New Jersey

Naftali Tishby
Hebrew University
Jerusalem, Israel

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 (functional form) 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.

1 Introduction

According to the Vapnik-Chervonenkis (VC) theory of learning curves [31, 30], 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 [6, 8, 26].

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 [4]. 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) [20], 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 [23, 21].

Of course, the VC bounds might simply be inapplicable to these experiments, because backpropagation is not equivalent to empirical error minimization. Vapnik has 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) [30, 17, 3]. 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 [29].

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 [32]). 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 be-

[1] Here for simplicity we are using the $\tilde{O}(\cdot)$ notation, which hides logarithmic factors in the same way the $O(\cdot)$ notation hides constant factors.

[2] By a power law, we mean the functional form $(a/m)^b$, where $a, b > 0$ are constants.

havior, 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 [18, 24, 1, 27] and the thermodynamic limit [27]) 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: $\epsilon_{gen}(h) = \mathbf{Pr}_{x \in D}[h(x) \neq f(x)]$. Note that the training sample S depends on f and m and $\epsilon_{gen}(h)$ 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 $\epsilon_{gen}(h)$ 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 $\epsilon_{gen}(h)$ 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

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

Thus, $VS(S) \subseteq \mathcal{F}$ 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:

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

Thus, $VS(S)$ 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.

[3] Aside to the statistical physicist: the annealed approximation was previously used to approximate the learning curve of a Gibbs learner, which chooses a hypothesis from a Gibbs distribution with the empirical error as energy. Here we adopt a microcanonical rather than a canonical ensemble, enabling us to obtain rigorous upper bounds from the annealed theory, rather than approximations. These bounds hold for all empirical error minimization algorithms, including the zero temperature limit of the Gibbs algorithm. Because of our desire for rigor, we have not used the replica method [11] in this paper. Engel, van den Broeck, and Fink have used the replica method to calculate the maximum deviation between empirical and generalization error in the function class, and the maximum generalization error in the version space [10, 9]. Although the replica method produces exact results when used correctly, it rests upon an interchange of limits for which no rigorous justification has been found.

The goal of this subsection is to examine the relationship between $VS(S)$ and $B(\epsilon)$. More specifically, for a sample S of size m, we would like to calculate the probability that $VS(S)$ 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 $\mathbf{Pr}_S[VS(S) \subseteq B(\epsilon)]$, 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 $\mathbf{Pr}_S[VS(S) \subseteq B(\epsilon)]$, or equivalently, an upper bound on $\mathbf{Pr}_S[VS(S) \not\subseteq B(\epsilon)]$.

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

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

Since the rate of decay is slower for small $\epsilon_{gen}(h)$, 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

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

which proves the following theorem.

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

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

We will refer 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 $0 < \delta \leq 1$, 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 $\epsilon_{gen}(h) \leq \epsilon$, where ϵ is the smallest value satisfying $\sum_{h \in \overline{B(\epsilon)}} (1 - \epsilon_{gen}(h))^m \leq \delta$.*

2.2 The standard cardinality bound

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

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

By applying Theorem 1 to this bound, we obtain the standard result that with probability $1 - \delta$, any consistent hypothesis h obeys $\epsilon_{gen}(h) \leq (\ln(|\mathcal{F}|/\delta))/m$. 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 Equation (1) and those given in Equation (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 Equation (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 Equation (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 Equation (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 [30, 22, 7, 16] 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 $\epsilon_{gen}(h)$ 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 $\epsilon_{gen}(h_i) = \epsilon_i$. Then for each index $1 \leq j \leq r$ we can define the cardinality of the jth error shell $Q_j = |\{f' \in \mathcal{F} : \epsilon_{gen}(f') = \epsilon_j\}|$. 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:

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

Together with Theorem 1, we can obtain the following bound on $\epsilon_{gen}(h)$ 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

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

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 indexing 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 $N \to \infty$, 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 Equation (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:

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

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 $\epsilon_{i_{N,\tau}}^N \geq \epsilon^* + \tau$. 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

$$\sum_{j=i_{N,\tau}}^{r(N)} e^{t(N)[s(\epsilon_j^N) + \alpha\log(1-\epsilon_j^N)]} \tag{12}$$

$$\leq \sum_{j=i_{N,\tau}}^{r(N)} e^{-t(N)\Delta} \tag{13}$$

$$\leq (r(N) - i_{N,\tau})e^{-t(N)\Delta} \tag{14}$$

$$\leq r(N)e^{-t(N)\Delta} \tag{15}$$

where the first inequality follows from the fact that for all $i_{N,\tau} \leq j \leq r(N)$ we have $\epsilon_j^N \in [\epsilon^* + \tau, 1]$. 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 curves, since the sample size in the thermodynamic limit is always a constant times $t(N)$. Given the scaling function, the smaller the the entropy bound $s(\epsilon)$, the smaller the rightmost crossing ϵ^* will be, and consequently the better the bound obtained from Theorem 4.

2.5 Extracting scaled learning curves from the thermodynamic limit method

Theorem 4 gives a bound on the limiting generalization error of consistent algorithms on a sample size m that is a *fixed* constant α times the scaling function $t(N)$. However, the real value of the thermodynamic limit method emerges only when we now allow the value of α to vary, taking the thermodynamic limit by applying Theorem 4 to each value, and examine the learning curve as a function of increasing α. As we shall now see, it is in such *scaled learning curves* (we refer to them as scaled because they are expressed as a function of the multiple α of $t(N)$ rather than in the more traditional absolute number of examples) that interesting behavior such as phase transitions appears. We shall also see that the thermodynamic limit method permits an intuitive and highly visual derivation of scaled learning curves.

We first illustrate the derivation of scaled learning curves using several artificial examples. By artificial we mean that rather than defining natural function class, target function and distribution sequences $\mathcal{F}_N$, f_N and D_N, and then deriving an appropriate scaling function $t(N)$ and entropy bound $s(\epsilon)$, instead we will simply start with a given $s(\epsilon)$ and carry the analysis forward. However, the lower bound provided in Section 2.8 demonstrates that there do exist function class and distribution sequences whose true scaled learning curves match the bounds we will give in this section. In the following sections, we give examples of complete analyses (that is, beginning with given $\mathcal{F}_N$, f_N and D_N) for some natural function classes.

To start, suppose that for some scaling function $t(N)$ we have the permissible entropy bound $s(\epsilon) = 1$ (a rather weak entropy bound). Then in Figure 1, we have plotted both the constant entropy bound $s(\epsilon) = 1$, and the function $-\alpha \log(1-\epsilon)$ for three values $\alpha = \alpha_1, \alpha_2, \alpha_3$. The resulting rightmost intersections $\epsilon_1 = \epsilon^*(\alpha_1), \epsilon_2 = \epsilon^*(\alpha_2), \epsilon_3 = \epsilon^*(\alpha_3)$ are then identified on the ϵ-axis. Here we now adopt the convention of writing ϵ^* as a function of α, since we no longer regard α as a constant.

In Figure 2, we then plot the rightmost crossing $\epsilon^*(\alpha)$ as a continuous function of α (and identify the points (α_i, ϵ_i) for $i = 1, 2, 3$ from Figure 1). This plot is what we mean by the scaled learning curve, and Theorem 4 tells us that in the limit $N \to \infty$, this scaled learning curve bounds the generalization error of consistent algorithms given $\alpha t(N)$ examples.

Note from Figure 1 that $-\alpha \log(1-\epsilon)$ is essentially linear with slope α, and it is the rightmost intersection of this roughly linear function with $s(\epsilon)$ that gives the corresponding point on the scaled learning curve. Furthermore, the energy function is independent of the learning problem in Theorem 4, and thus in general, for any entropy bound $s(\epsilon)$, to get the scaled learning curve we will be looking at the leftward progress of the rightmost intersection $\epsilon^*(\alpha)$ between the nearly-linear energy and $s(\epsilon)$ as α grows. In the particular example $s(\epsilon) = 1$, this progress is quite uniform, resulting in the familiar power law scaled learning curve of Figure 2.

A less familiar and more interesting example occurs for the single-peak entropy bound $s(\epsilon)$ shown in Figure 3. [4] We shall shortly see in Section 2.6 that this entropy bound actually occurs for a natural and well-studied learning problem. In this example we see that for small α, the leftward progress of $\epsilon^*(\alpha)$ is rather slow, due to the large negative slope of $s(\epsilon)$ on the right side of its peak. This for instance is the case for α near the plotted value α_1. For some larger value of α, $\epsilon^*(\alpha)$ moves over the peak of $s(\epsilon)$ and thus begins decreasing more rapidly.

Then something interesting happens. There is a *critical value* α_2 that gives the intersection $\epsilon^*(\alpha_2) = \epsilon_2$. For this critical value, we see that the energy curve is barely intersecting the entropy curve. For $\alpha > \alpha_2$ (for example, for the plotted value α_3), we see from Figure 3 that the rightmost intersection is 0! Theorem 4 can be applied to obtain the scaled learning curve bound of Figure 4, which exhibits a *phase transition* from error ϵ_2 to perfect generalization (error 0) at $\alpha = \alpha_2$.

A similar but more subtle example is shown for another single-peak $s(\epsilon)$ in Figures 5 and 6. Here again, leftward progress of $\epsilon^*(\alpha)$ for smaller α is slow due to the large negative slope of $s(\epsilon)$ on the right-hand side of its peak (for instance, at $\alpha = \alpha_1$). Again, there is a critical value α_2 which results in an intersection at $\epsilon_2^+ = \epsilon^*(\alpha_2)$, slightly to the left of the peak of $s(\epsilon)$. However, for α just larger than α_2 we do *not* transition to perfect learning, but to error ϵ_2^-. The difference between this example and that of Figures 3 and 4 is that this time the entropy curve is sufficiently large near ϵ_2^- to "catch" $\epsilon^*(\alpha)$ for α above the critical value. Following the transition, the decrease of $\epsilon^*(\alpha)$ resumes rather gradual behavior (for instance, near α_3). This is all clearly seen in the scaled learning curve of Figure 6.

As our next example we consider a double-peak entropy bound in Figures 7 and 8. Here we see there are two critical values, α_2 and α_4. Initial progress of $\epsilon^*(\alpha)$ occurs at a steady but controlled rate, for instance at α_1. As α becomes larger than α_2, there is a sudden burst of generalization (a phase transition), not to perfect generalization, but from error ϵ_2^+ to ϵ_2^- on the right side of the left peak of $s(\epsilon)$. Then progress is slow, for instance at α_3, until α becomes larger than α_4, at which point we have a transition to perfect generalization (so for α_5 the error is 0). One aspect of this example worth noting is the fact that although the energy may intersect $s(\epsilon)$ many times, we are interested only in the rightmost intersection.

As our final artificial example, we consider a three-peak entropy bound in Figures 9 and 10. This example demonstrates the interesting phenomenon of *shadowing* predicted by our theory, because despite the change in $s(\epsilon)$ from our last example, we see that the scaled learning curve of Figure 10 is quite similar in form to that of Figure 8. Figure 9 shows the reason for this: by the time α becomes larger than the first critical value α_2, the energy curve is already above the small middle peak of $s(\epsilon)$, and thus the phase transition is from ϵ_2^+ to ϵ_2^-, completely bypassing the middle peak. Thus, the small middle peak of $s(\epsilon)$ is in the "shadow" of the large rightmost peak. There is an intuitive explanation for this phenomenon. Despite the fact that (relative to the scaling function) there are a significant number of functions of generalization error approximately ϵ' (resulting in the middle peak of $s(\epsilon)$ centered at ϵ'), by the time the sample size is large enough to eliminate the considerably larger number of functions of generalization error approximately ϵ_2^+ from the version space, the functions at generalization error ϵ' are already eliminated from the version space. Note that if this middle peak were higher, there would be a brief transition from ϵ_2^+ to near ϵ', and then from there to a value

[4] Throughout this section, we will refrain from giving the explicit functions $s(\epsilon)$ used to generate the plots, since some of them are rather complicated, and it is their shape rather than their mathematical definitions that are of interest here.

on the right side of the left peak.

In all of these examples, we have concentrated on the qualitative behavior (including coarse phenomena such as phase transitions) of scaled learning curves at moderate values of α. Also of interest are the large α asymptotics of the scaled learning curve, that is, the asymptotic rate of approach to generalization error 0. In our theory this rate is obviously determined by the behavior of the entropy bound $s(\epsilon)$ for $\epsilon \approx 0$. It turns out that many natural examples of $s(\epsilon)$ fall into a few broad categories of behavior near 0, and this is discussed in Section 3.5.

2.6 Analysis of the Ising perceptron

We now tackle some real examples of the application of our theory, complete with determination of the appropriate scaling function and a permissible entropy bound.

We first consider the class of Ising perceptrons [12, 14, 28]. Suppose that the function class $\mathcal{F}_N$ consists of all homogeneous perceptrons in which the weights are constrained to be ± 1. [5] Let the distribution D_N be any spherically symmetric distribution on $\Re^N$, and let the target function $f_N \in \mathcal{F}_N$ be arbitrary. It will turn out that for this problem, the appropriate scaling function is simply $t(N) = N$. We now derive a permissible entropy bound for this scaling function, and then extract the associated scaled learning curve.

An Ising perceptron is parametrized by a weight vector $\mathbf{w}$ in the hypercube $\{-1, 1\}^N$, and maps $\mathbf{x} \in \Re^N$ to $\mathrm{sgn}(\mathbf{w} \cdot \mathbf{x})$. For a spherically symmetric distribution D_N, the probability of disagreement between two perceptrons is proportional to the angle between them. Hence if $\mathbf{w}_0$ is the weight vector of the target function,

$$\epsilon_{gen}(\mathbf{w}) = \frac{1}{\pi} \cos^{-1} \frac{\mathbf{w} \cdot \mathbf{w}_0}{N} = \frac{1}{\pi} \cos^{-1} \left(1 - \frac{2 d_H(\mathbf{w}, \mathbf{w}_0)}{N} \right) \tag{17}$$

where d_H denotes the Hamming distance. The Hamming distance layers the function class like an onion with N error shells surrounding the target at the center. The number of perceptrons at Hamming distance j from the target is $Q_j^N = \binom{N}{j}$, and they all have generalization error $\epsilon_j^N = (1/\pi) \cos^{-1}(1 - 2j/N)$. Since the binomial coefficients are bounded by

$$\frac{1}{N} \log Q_j^N \leq \mathcal{H}\left(\frac{j}{N} \right) = \mathcal{H}\left(\sin^2(\pi \epsilon_j^N / 2) \right) \tag{18}$$

where $\mathcal{H}(p) \equiv -p \log p - (1-p) \log(1-p)$, a permissible entropy bound for scaling function $t(N) = N$ is

$$s(\epsilon) = \mathcal{H}\left(\sin^2(\pi \epsilon / 2) \right). \tag{19}$$

We have acutally already discussed the resulting entropy-energy competition for this problem in Section 2.5. Recall that in Figure 3 we graph the competition, and in Figure 4 we graph the scaled learning curve obtained by applying Theorem 4. Thus for this problem our theory predicts slow initial learning, followed by a phase transition to perfect generalization at $\alpha_2 = 1.448$. We remind the reader that a sudden transition in our bound does not necessarily imply a sudden transition in the true behavior of any consistent learning algorithm. However, this bound does show that any consistent learning algorithm must have reached zero error with probability approaching 1 in the thermodynamic limit for scaled sample size greater than 1.448. This bound on the critical value was known from the work of Gardner and Derrida [12], and extended to the case of boolean inputs by Baum, Lyuu and Rivin [2, 19]. Here we are actually giving a bound on the entire learning curve, and the behavior of our bound is very similar in shape to learning curves obtained in both simulations and non-rigorous replica calculations from statistical physics [14, 28, 25, 9]. [6]

In Figure 11, we graph the *difference* of the entropy and energy curves shown in Figure 3, that is, we plot $s(\epsilon) + \alpha \log(1-\epsilon)$ for the three values of α. This plot is simply another way of visualizing the entropy-energy competition. The zero crossings of the graphs in Figure 11 correspond to the intersections of the entropy and energy curves in Figure 3, and thus it is now the leftward progress of the rightmost zero crossing of $s(\epsilon) + \alpha \log(1-\epsilon)$ that yields the scaled learning curve as α increases. The quantity $N[s(\epsilon) + \alpha \log(1-\epsilon)]$ is the logarithm of the average number of surviving hypotheses at distance ϵ from the target, and is the exponent in the sum of Equation (10). For $\alpha < \alpha_2$, there are two zero crossings. The right zero crossing yields the upper bound on generalization error of Theorem 4. The left zero crossing also has a meaning. With high probability, there are no hypotheses in the version space with error less than this left crossing except for the target itself. So the version space minus the target is contained within an annulus [9] whose inner and outer limits are the left and right zero crossings.

It is instructive to compare our bounds with the cardinality and VC bounds for this problem. Since both of these latter bounds go like N/m, and the lowest error shell is at $\epsilon_1 \sim 1/\sqrt{N}$, the critical m for perfect learning is $m \sim N^{3/2}$, rather than $m \sim N$.

2.7 Analysis of monotone boolean conjunctions

In this example, the input space X_N is the boolean hypercube $\{0,1\}^N$. The class $\mathcal{F}_N$ consists of the 2^N functions computed by the conjunction of a subset of the input variables $x_1, \ldots, x_N$, along with the empty (always 0) function $\emptyset$ and the universal (always 1) function $\{0,1\}^N$. The input distribution D_N is uniform over $\{0,1\}^N$. A similar scenario has also been analyzed in the machine learning literature [23, 21].

We will examine the thermodynamic limit for two different choices of target functions f_N. We begin with the target function $f = \{0,1\}^N$, in which every input is a positive example. Any conjunction h of exactly i variables from $x_1, \ldots, x_N$ has generalization error

$$\epsilon_{gen}(h) = Pr_{\vec{x} \in D_N}[h(\vec{x}) = 0] = 1 - 1/2^i.$$

Hence the error shells are $1/2 = \epsilon_1^N < \epsilon_2^N < \cdots < \epsilon_N^N = 1 - 1/2^N$, where $\epsilon_i^N = 1 - 1/2^i$. The number of conjunctions

[5] The designation "Ising" refers to the ± 1 constraint, which is present in the original Ising model of magnetism with N interacting spins.

[6] According to calculations using the replica method of statistical physics, for this problem the true scaled learning curve of the Gibbs learning algorithm (which chooses a random consistent hypothesis from the version space) exhibits a phase transition to perfect generalization at $\alpha = 1.245$. This picture is consistent with the results of exhaustive enumeration by computer for up to $N = 32$.

in the ith shell is $Q_i^N = \binom{N}{i} \leq N^i$. Since

$$\frac{\ln Q_i^N}{\log_2 N} \leq i \ln 2 = -\ln(1-\epsilon_i^N) \tag{20}$$

we choose the scaling function to be $t(N) = \log N$ and thus the sample size is written as $m = \alpha \log N$. A permissible entropy bound for $t(N)$ is $s(\epsilon) = -\ln(1-\epsilon)$.

The competition between $s(\epsilon)$ and $-\alpha \log(1-\epsilon)$ results in a scaled learning curve that exhibits a sudden transition: for any $0 \leq \alpha < 1$, the rightmost crossing $\epsilon^*(\alpha)$ does not exist and our bound on the generalization error is 1. But for $\alpha \geq 1$, $s(\epsilon)$ is dominated by $-\alpha \log(1-\epsilon)$, so $\epsilon^*(\alpha)$ makes a sudden transition to 0. In summary, our theory predicts that in the thermodynamic limit, for $\alpha < 1$ there is no generalization, but for $\alpha > 1$ there is perfect generalization.

Our bound can be checked by deriving the exact learning behavior. In the problem described, every random example is positive for f_N, and every positive example $\vec{x}$ eliminates from the version space any conjunction containing a variable that is set to 0 in $\vec{x}$. Since half of the remaining variables is eliminated by each example, it should take roughly $\log_2 N$ examples to eliminate all N variables and hence all conjunctions, leaving only the target function.

A more precise calculation goes as follows. Since each variable has probability 2^{-m} of surviving m examples, the number j of surviving variables obeys a binomial distribution:

$$P(j) = \binom{N}{j} \left(\frac{1}{2^m}\right)^j \left(1 - \frac{1}{2^m}\right)^{N-j} \tag{21}$$

The function with maximum generalization error in the version space is a conjunction of all j surviving variables, so that $\max_{h \in VS(S)} \epsilon_{gen}(h) = \epsilon_j^N$. Then Chernoff bounds on the fluctuations in j yield

$$1 - 2^{-N2^{-m}(1-\tau)} \leq \max_{h \in VS(S)} \epsilon_{gen}(h) \leq 1 - 2^{-N2^{-m}(1+\tau)} \tag{22}$$

with confidence greater than $1 - 2e^{-N\tau^2/3}$. Taking the thermodynamic limit with $m = \alpha \log_2 N$, then $\epsilon \to 1$ for any $\alpha > 1$, and $\epsilon \to 0$ for any $\alpha < 1$ with confidence approaching 1.

For this model, the cardinality and VC bounds give a learning curve of order N/m, which drops below the lowest error level $\epsilon_1^N = 1/2$ for m of order N. Hence these bounds also predict perfect generalization, but with a bound on the critical m of order N rather than $\log N$.

Now let the target function be the empty function $f_N = \emptyset$. Since a conjunction h of i variables has $\epsilon_{gen}(h) = 1/2^i$, the error shells are $1/2^N = \epsilon_1^N < \epsilon_2^N < \cdots < \epsilon_N^N = 1/2$, where $\epsilon_i^N = 1/2^{N-i+1}$. The number of conjunctions in the ith shell is $Q_i^N = \binom{N}{N-i} \leq N^{N-i}$. We again choose $t(N) = \log N$ as the scaling function. Then

$$\frac{\ln Q_i^N}{\log_2 N} \leq (N-i) \ln 2 = -\ln 2\epsilon_i^N \tag{23}$$

so that $s(\epsilon) = -\ln 2\epsilon$ is a permissible entropy bound for $t(N)$. The rightmost zero crossing of $s(\epsilon)$ and $-\alpha \log(1-\epsilon)$ gives the scaled learning curve $\epsilon \sim O(\log \alpha / \alpha)$.

One interesting aspect of this learning problem is that the scaled learning curve is highly dependent on the target function. Whereas learning the target functions $f_N = \{0,1\}^N$ led to a sudden transition in generalization, learning the empty function $f_N = \emptyset$ led to a slow power law decrease. This is in marked contrast to the Ising perceptron problem, where the learning curve is independent of which weight vector is the target function.

2.8 The thermodynamic limit lower bound

In this section, we give a theorem demonstrating that Theorem 4 is tight in a fairly general sense (modulo the given entropy bound). More precisely, for any function $s(\epsilon)$ meeting certain mild conditions, we construct a family of function classes $\mathcal{F} = \{\mathcal{F}_N\}$ such that $s(\epsilon)$ is a permissible entropy bound for the scaling function $t(N) = N$, and in the thermodynamic limit the rightmost crossing of the functions $s(\epsilon)$ and $2\alpha\epsilon$ is a lower bound on the generalization error of worst hypothesis in the version space. Note that although this does not exactly match Theorem 4, which gives as an upper bound the rightmost crossing of $s(\epsilon)$ and $-\alpha \log(1-\epsilon)$, the qualitative behavior of the scaled learning curves obtained by intersecting with $2\alpha\epsilon$ and $-\alpha \log(1-\epsilon)$ is essentially the same. In particular, our lower bound shows that the various scaled learning curve phenomena examined in Section 2.5 (such as phase transitions and shadowing) can actually occur for certain function classes and distributions.

In the same way that lower bounds for the VC theory show that if the only parameter of the learning problem we consider is the VC dimension, then the existing learning curve upper bounds based on the VC dimension are essentially the best possible, Theorem 5 shows that if the only parameter of the learning problem we use is a given entropy bound $s(\epsilon)$, then Theorem 4 gives essentially the best possible learning curve upper bound. Thus, in the absence of further information about the function class, distribution and target function sequences, the scaled learning curves derived in Section 2.5 are essentially the best possible. Similarly, the lower bound shows that better learning curves for the Ising perceptron and boolean conjunction problems that depend only on the entropy bound cannot be obtained.

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 $\epsilon_{gen}(h) = j/N$ 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 [13] 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 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 $h \in \mathcal{F}_N$, $\epsilon_{gen}(h) = \mathbf{Pr}_{\langle x,y \rangle \in D_N}[h(x) \neq y]$. For simplicity we will assume that there is a unique best hypothesis in $\mathcal{F}_N$

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

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 $\epsilon_{\min} = \epsilon_{gen}(h^*)$ rather than 0 as the 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:

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

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 $\epsilon_{trn}(h, S)$:

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

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 $(1 - \epsilon_{gen}(h))^m$ 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 $(1 - \epsilon_{gen}(h))^m$ 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

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

In other words, we imagine that $u(\epsilon_{gen}(h))$ assesses a penalty to $\epsilon_{gen}(h)$ that increases with larger $\epsilon_{gen}(h)$, 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:

$$\begin{aligned} &\mathbf{Pr}_S[h \in VS(S)] && (28)\\ &\leq \mathbf{Pr}_S[\epsilon_{trn}(h,S) \leq \epsilon_{trn}(h^*,S)] && (29)\\ &\leq \left[1 - \epsilon(h,h^*) + \sqrt{\epsilon(h,h^*)^2 - (\epsilon_{gen}(h) - \epsilon_{\min})^2}\right]^m && (30)\end{aligned}$$

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 [5] (see Section A.2 of the Appendix).

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

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

then from Equation (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 Equation (32) gives a universally permissible energy bound. After a little algebra, this bound reduces to

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

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 $VS(S)$ is contained in $B(\epsilon)$. This enables us to bound the error of all empirical error minimization algorithms. For with confidence $\mathbf{Pr}_S[VS(S) \subseteq B(\epsilon)]$, 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

$$\begin{aligned} \sum_{h \in \overline{B(\epsilon)}} e^{-u(\epsilon_{gen}(h))m} &\leq |\mathcal{F}|(1 - (\sqrt{\epsilon} - \sqrt{\epsilon_{\min}})^2)^m && (35)\\ &\leq |\mathcal{F}| e^{-(\sqrt{\epsilon} - \sqrt{\epsilon_{\min}})^2 m} && (36)\end{aligned}$$

because $\epsilon_{gen}(h) > \epsilon$ 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

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

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 Equation (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 [30].

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 Equation (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 $\epsilon_{gen}(h) = \epsilon_i$. 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

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

with probability at least $1-\delta$. Thus we have a bound on $\epsilon_{gen}(h)$ 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 is 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 Equation (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 Equation (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 Equation (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 Equation (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 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 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,

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

The distribution of inputs $\mathbf{x}$ 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 [15].

In this case, one can show that

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

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

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

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 Equation (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

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

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 Equations (32) and (33), corresponding to the choices of $v(\epsilon)$ as above and $v(\epsilon) = \epsilon + \epsilon_{\min}$. These four possibilities give the bounds exhibited below:

	cardinality	entropy
$v(\epsilon) = \epsilon + \epsilon_{\min}$	$\alpha^{-1/2}$	$(\log \alpha)/\alpha$
$v(\epsilon) \sim \sqrt{\Delta\epsilon}$	$\alpha^{-2/3}$	$((\log \alpha)/\alpha)^2$

(53)

Note how much weaker some of the bounds are than others.

3.5 Large-α asymptotics of scaled learning curves

Our formalism can be used to give a classification of the large-α asymptotics of scaled learning curves, [7] thus completing a classification program that has been suggested by several researchers [24, 25, 1]. From Equation (32) and Lemma 9, the weaker form

$$u(\epsilon) = \frac{(\epsilon - \epsilon_{\min})^2}{2v(\epsilon)} \tag{54}$$

is derived as a permissible energy bound in the Appendix in Section A.2. The entropy-energy competition then takes the form

$$s(\Delta\epsilon) = \alpha u(\Delta\epsilon) = \alpha \frac{(\Delta\epsilon)^2}{2v(\Delta\epsilon)} \tag{55}$$

where we have rewritten all functions of ϵ as functions of the difference $\Delta\epsilon = \epsilon - \epsilon_{\min}$.

Since the only model-dependent quantities are $s(\Delta\epsilon)$ and $v(\Delta\epsilon)$, we can classify the large α asymptotics of scaled learning curves. In fact, the only model-dependent quantity that need enter is a single exponent x, defined by

$$s(\Delta\epsilon) v(\Delta\epsilon) \sim (\Delta\epsilon)^x \tag{56}$$

near $\Delta\epsilon = 0$. This yields the following cases:

- If $x > 2$, there is a first-order (sudden) phase transition to perfect learning. This is assuming that $s(0) = 0$, so that $\Delta\epsilon = 0$ is always a solution of Equation (55), if not the rightmost solution. This is the generic case, unless there are exponentially many functions with $\epsilon = \epsilon_{\min}$.
- If $1 < x < 2$, the error decays as a power law, $1/\alpha^{2-x}$.
- In the marginal case $x = 2$, the behavior can be affected by logarithmic corrections to the power law of Equation (56). In the absence of such corrections, there is a second-order (continuous) transition to perfect learning in which the error drops to zero like $\epsilon \sim \alpha_c - \alpha$. In the presence of a logarithmic correction, $s(\Delta\epsilon) v(\Delta\epsilon) \sim -(\Delta\epsilon)^2 \log \Delta\epsilon$, the error bound decays exponentially with α.

This classification scheme is a generalization of that of Sompolinsky and his colleagues to include unrealizable rules [25].

[7] Note that the large-α asymptotics, which by definition invoke a thermodynamic limit, may be different from the large m asymptotics for a fixed function class.

4 The Infinite Case

The final generalization of our theory that needs to be discussed is to the frequent case in which the function class $\mathcal{F}$ (whether it realizes the target function or not) has infinite cardinality. Unfortunately, while there are certainly several plausible directions we can take to adapt our theory to this case, none of these has emerged as definitively the best choice for handling the infinite case. This is partially due to the lack of known natural examples of infinite classes that lead to learning curve behavior other than a power law (thus suggesting that the extremely general VC dimension-based approach is sufficient for analyzing most classes), and partially due to the difficulty of the calculations required by the various approaches. Thus, by necessity our examination of the infinite case will be considerably more open-ended than for the finite case.

We begin by noting that practically every step of our analysis for the finite case was based on computing the (finite) cardinality of some subclass of $\mathcal{F}$. This began with the shell decomposition of $\mathcal{F}$ to obtain the subclass cardinalities Q_j, whose logarithms were eventually bounded by the entropy function $s(\epsilon)$ in the thermodynamic limit method. Obviously, new ideas will be required in order to carry out a similar analysis in the infinite case. Our eventual goal should be to preserve the essentials of our theory: namely, to again describe learning curves as a competition between "entropy" and "energy", with the largest value for which energy dominates entropy being a bound on the generalization error of empirical minimization algorithms. However, there are now several distinct candidates for our entropic measure. We now discuss in some detail just one of these candidates, which essentially attempts to reduce the infinite case to a series of finite problems. In Section 6, we briefly mention alternative approaches that are the focus of our current research.

4.1 The covering approach

In the covering approach, we reduce an infinite cardinality function class to a series of finite classes, and perform our analysis for the finite case on each of these classes in order to obtain a bound on the learning curve.

For any fixed function class $\mathcal{F}$ (of possibly infinite cardinality), any distribution D, and any value $\gamma \in [0,1]$, a subclass $\mathcal{F}[\gamma] \subseteq \mathcal{F}$ is called a *γ-cover* of $\mathcal{F}$ with respect to D if for every $f \in \mathcal{F}$ there exists an $f' \in \mathcal{F}[\gamma]$ such that $\epsilon(f, f') \leq \gamma$. In other words, while there may be functions in $\mathcal{F}$ that are not realizable in $\mathcal{F}[\gamma]$, the extent of this unrealizability is bounded by the parameter γ.

There is a canonical greedy construction of γ-covers that will be particularly helpful to keep in mind. Thus, throughout this section, for any fixed value γ, we assume that $\mathcal{F}[\gamma]$ is a γ-cover of $\mathcal{F}$ with respect to D obtained by initially choosing any function in $\mathcal{F}$, then inductively adding to $\mathcal{F}[\gamma]$ at each step any $f \in \mathcal{F}$ that is distance at least γ (with respect to D) from all $h \in \mathcal{F}[\gamma]$. This process is repeated until no more functions can be added. It is easy to see that the resulting set $\mathcal{F}[\gamma]$ does indeed form a γ-cover, and it is known that this γ-cover is in fact at most twice the cardinality of the *smallest* possible γ-cover. Furthermore, suppose $\gamma' < \gamma$. Then we can extend $\mathcal{F}[\gamma]$ to obtain a γ'-cover $\mathcal{F}[\gamma'] \supseteq \mathcal{F}[\gamma]$ by again greedily adding to $\mathcal{F}[\gamma]$ functions that are at distance at least γ' until no such function exists. The resulting cover $\mathcal{F}[\gamma']$ will again have cardinality at most twice the smallest γ'-cover. In this way we can obtain for any sequence $\gamma_1 > \gamma_2 > \gamma_3 > \cdots$ a sequence of *nested* covers $\mathcal{F}[\gamma_1] \subseteq \mathcal{F}[\gamma_2] \subseteq \mathcal{F}[\gamma_3] \subseteq \cdots$.

Let us fix $\gamma \in [0,1]$, and assume that $\mathcal{F}$ has a finite γ-cover with respect to D. This is not as severe an assumption as it might initially seem. For instance, it is well-known that any class of VC dimension d has a γ-cover of cardinality at most $O(1/\gamma^d)$ with respect to any distribution and for every γ. Furthermore, if a class is not finitely γ-coverable with respect to D, then the generalization error cannot be made less than γ in *any* finite number of examples. Thus, we see that finite coverability is really a minimal assumption for attaining small generalization error.

With a fixed γ-cover $\mathcal{F}[\gamma]$ of $\mathcal{F}$ with respect to D in mind, it is a straightforward application of our theory for the finite unrealizable case to analyze the algorithm that performs empirical error minimization with respect to $\mathcal{F}[\gamma]$. Given m examples, this algorithm outputs any $h \in \mathcal{F}[\gamma]$ with minimum empirical error on the sample. Note that this algorithm explicitly does *not* choose from the full class $\mathcal{F}$, but limits its search to the fixed finite subclass $\mathcal{F}[\gamma]$. For a fixed target function (contained in $\mathcal{F}$ or not), the thermodynamic limit method applied to $\mathcal{F}[\gamma]$ results in a bound on the error of ϵ^*_γ, where ϵ^*_γ is the rightmost crossing function of a permissible entropy bound $s_\gamma(\epsilon)$ for $\mathcal{F}[\gamma]$ and an energy function $\alpha u_\gamma(\epsilon)$, where as before $\epsilon_{\min}(\gamma) \leq \gamma$ is the smallest possible generalization error achievable in $\mathcal{F}[\gamma]$. The idea of using empirical minimization over a finite cover for an infinite class has also been investigated by Benedek and Itai [3] in their investigation of distribution-specific sample complexity, and also by Vapnik [30].

Things become more interesting when we take the natural step of analyzing the algorithm that first chooses an advantageous value for the realizability parameter γ and then performs empirical minimization using $\mathcal{F}[\gamma]$. More precisely, if we assume that the algorithm has knowledge of $s_\gamma(\epsilon)$ for each γ, [8] and is given $m = \alpha t(N)$ examples of the target function, then the algorithm will explicitly choose γ to minimize the resulting rightmost crossing ϵ^*_γ.

It is worth mentioning at this point that while such an algorithm may be difficult or impossible to implement (requiring the possibly difficult choice of γ and knowledge of the finite covers $\mathcal{F}[\gamma]$), it is worth study for at least two reasons. First, the algorithm is of some theoretical interest since it explicitly considers the potential trade-off between the best *error* achievable in the chosen cover $\mathcal{F}[\gamma]$ (which improves as $\gamma \to 0$), and the *size* of $\mathcal{F}[\gamma]$ (which increases as $\gamma \to 0$). Second, although one might not implement such an algorithm in practice, any bound we can provide on its generalization error can provide bounds on the generalization error of optimal algorithms (such as the Bayes or Gibbs algorithms in a Bayesian framework [17]).

In the thermodynamic limit, we may upper bound the generalization error of this algorithm by

$$\epsilon^* = \min_{\gamma \in [0,1]} \epsilon^*_\gamma. \tag{57}$$

Let us interpret this bound. For each fixed γ, we are computing the rightmost crossing ϵ^*_γ of $s_\gamma(\epsilon)$ and $\alpha u_\gamma(\epsilon)$. What

[8] This is a nontrivial assumption, since in many of the examples we have examined, the entropy bound depends strongly on the target function, which we of course assume is unknown. Thus, we are really assuming here that either $s_\gamma(\epsilon)$ is invariant to the target function (as in the realizable Ising perceptron), or that is a worst-case entropy bound over all target functions.

is the expected behavior of this crossing as $\gamma \to 0$? Well, as $\gamma \to 0$ the covers $\mathcal{F}[\gamma]$ are becoming larger (since we require more functions to achieve the greater realizability), and we thus expect $s_\gamma(\epsilon)$ to increase. Indeed, if we use the nested cover construction suggested at the beginning of this section, then for any $\gamma' \leq \gamma$ we will have $s_{\gamma'}(\epsilon) \geq s_\gamma(\epsilon)$ for every ϵ. Thus, decreasing γ has the effect of "lifting" $s_\gamma(\epsilon)$ (although perhaps in a very nonuniform and complex manner). If $u_\gamma(\epsilon)$ remained unchanged as γ decreased, then the lift to $s_\gamma(\epsilon)$ could only cause the crossing ϵ^*_γ to increase, thus predicting that decreasing γ could never help.

However, $u_\gamma(\epsilon)$ does *not* remain unchanged as γ decreases. Rather, smaller γ results in a smaller value for the optimal error $\epsilon_{\min}(\gamma) \leq \gamma$, thus shifting the energy curve $u_\gamma(\epsilon)$ to the left. If $s_\gamma(\epsilon)$ remained unchanged as $\gamma \to 0$, we would predict that decreasing γ could never hurt, and would choose $\gamma = 0$.

Thus in general, the covering analysis predicts that while for each fixed γ, the best error for resolution γ is determined by the competition between $s_\gamma(\epsilon)$ and $\alpha u_\gamma(\epsilon)$, the overall best error is governed by the competition between the lift to $s_\gamma(\epsilon)$ and the leftward shift to $u_\gamma(\epsilon)$ as $\gamma \to 0$.

5 Generalization of the Theory to Distribution Learning

We believe that the basic components of the theory outlined here — namely, the identification of the appropriate entropy and energy bounds, and the resulting bound on the learning curve in terms of their competition — should generalize considerably beyond the basic model of supervised learning of boolean functions examined in this paper. By this we mean the theory should generalize to cover many different models of learning from random independent observations, using a variety of loss functions. To demonstrate this, we now informally work out a simple example in which we calculate learning curve bounds, in the thermodynamic limit, for a certain class of probability distributions with respect to the well-known Kullback-Leibler divergence.

Let the target distribution D over $\{0,1\}^N$ be defined as follows: for each $1 \leq i \leq N$, we let the ith bit of the output vector be 0 with probability $(1-p)$ and 1 with probability p. Here p is a parameter in $[0, 1/2]$ that will remain fixed for the ensuing discussion. Thus, the distribution D can be regarded as outputting a random vector obtained by corrupting each bit of the vector $\vec{0} = 00\cdots0$ with independent probability p.

Let the class of hypothesis distributions be similarly defined by all the possible "center" vectors $\vec{v} \in \{0,1\}^N$. Thus, the vector $\vec{v}$ represents the distribution $D_{\vec{v}}$ obtained by corrupting each bit of $\vec{v}$ with independent probability p, and the target $D \equiv D_{\vec{0}}$. It should be clear that the Kullback-Leibler divergence of $D_{\vec{v}}$ from the target D depends only on the Hamming distance between $\vec{v}$ and $\vec{0}$, which is just the number of 1's appearing in the vector $\vec{v}$.

We now undertake an analysis of the Kullback-Leibler divergence, as a function of the sample size m, of the hypothesis $D_{\vec{v}}$ minimizing the empirical log-loss

$$loss(D_{\vec{v}}, S) = \sum_{\vec{y} \in S} \log(1/D_{\vec{v}}[\vec{y}]). \tag{58}$$

Here S consists of m independent random draws from the target distribution D. Thus, we are simply analyzing in our theory the learning curve of the maximum-likelihood approach to this problem.

Now it is not hard to show that if $\vec{v}$ is a vector with exactly r 1's in it, then the Kullback-Leibler divergence of $D_{\vec{v}}$ to D is

$$r\left(p\log\frac{1}{1-p} + (1-p)\log\frac{1}{p} - H(p)\right) \tag{59}$$

where $H(p)$ is the usual binary entropy of p. Note that the divergence is 0 when $r = 0$ (the divergence of the target from itself is 0), and it is also 0 when $p = 1/2$ (since then every $\vec{v}$ generates the uniform distribution on $\{0,1\}^N$). Since p is fixed , let us use $C_p = p\log(1/(1-p)) + (1-p)\log(1/p) - H(p)$ to denote the constant inside the parentheses above. For convenience, we also divide the Kullback-Leibler divergence by N just to make our measure of generalization error an order 1 quantity. Then we see that our error levels are just $\epsilon^N_r = r(C_p/N)$ for $0 \leq r \leq N$, and the number of distributions in the class that are at divergence ϵ^N_r from the target is $Q^N_r = \binom{N}{r}$.

We now turn to the problem of finding a suitable energy function. In other words, suppose that $\vec{v}$ is a fixed vector with exactly r 1's, and suppose we draw a sample S of m vectors from the target distribution D. Then what is $\mathbf{Pr}_{S\in D^m}[loss(D_{\vec{v}}, S) \leq loss(D, S)]$?

To bound this probability, note that the difference in the log-loss incurred by the two distributions on any fixed vector $\vec{y}$ depends only on the setting in $\vec{y}$ of the r bits where $\vec{v}$ and $\vec{0}$ disagree (which we may assume without loss of generality are the first r bits). On a 0 in bits 1 through r, the target pays $\log(1/(1-p))$ and $D_{\vec{v}}$ pays $\log(1/p)$, and on a 1, the costs are reversed. Thus our problem simply reduces to the following: we have $m \cdot r$ Bernoulli trials, each with probability p of tails. What is the probability that we have a majority of tails? Now we can just use standard Chernoff bounds to obtain the following bound:

$$\mathbf{Pr}_{S\in D^m}[loss(D_{\vec{v}}, S) \leq loss(D, S)] \leq e^{-(mr/3)(1-2p)^2/(4p)}. \tag{60}$$

Thus when we write out our summation of entropy times energy (corresponding to Equation 7 in the boolean function learning setting), the rth term is $\binom{N}{r}e^{-(mr/3)(1-2p)^2/(4p)}$. Using the bound $\binom{N}{r} \leq N^r$ we can bound the rth term by $e^{r\log N-(mr/3)(1-2p)^2/(4p)}$. Factoring out the scaling factor $t(N) = \log N$, we rewrite this $e^{\log N(r-(\alpha r/3)(1-2p)^2/(4p))}$ where we define $\alpha = m/\log N$. In the thermodynamic limit, this predicts a phase transition to perfect generalization for α proportional to $p/(1-2p)^2$. This makes some sense, in that the critical α goes to infinity as p approaches $1/2$.

6 Conclusion

Two questions have often been raised in the computational learning theory community regarding the statistical physics approach to learning curves. Can it be made rigorous? Does it give any results that can not be derived from the VC theory? In this paper, we shown that for finite function classes and excluding replica calculations, the answer to both questions is affirmative. Under certain circumstances, our theory provides much tighter bounds than the VC theory, best illustrated in our examples exhibiting phase transitions.

Our theory gives tighter bounds than the VC theory at the expense of increasing the number of problem-dependent quantities. Since the computation of the entropy bound $s(\epsilon)$

requires knowledge of the input distribution, it is considerably more difficult than the computation of the VC dimension, which requires knowledge of only the function class. For this reason, applications of our theory to real problems may be difficult. Thus, our theory is descriptive rather than prescriptive at this point: it should be regarded more as an attempt to come to a theoretical understanding of the true behavior of learning curves, rather than as a tool for application.

There is obviously still much work to do in our theory, and we now list some of the research directions we are pursuing.

- **The infinite case.** The most glaring weakness of our theory, especially in comparison to the VC theory, is that we have developed and analyzed it only for finite cardinality concept classes. We are currently investigating extensions to the infinite case that are more refined than the covering approach discussed in Section 4.1, and are based on combining the shell decomposition with the VC dimension, VC entropy and random covering numbers [30, 22, 7, 16].

- **Expressing our bounds as penalty functions.** One of the most interesting aspects of the VC theory is Vapnik's explicit prescription in the unrealizable setting for trading off hypothesis class complexity (and therefore, ability to realize the target function) against empirical error [30]. This prescription is known as *structural risk minimization*, and the form it takes can be directly traced to the form of the VC bounds on learning curves. The fact that we now have learning curve bounds whose functional form can differ radically from the VC bounds opens the possibility for structural risk minimization prescriptions that are different from Vapnik's. Although possibly difficult to apply, such prescriptions may have interesting theoretical interpretations and consequences.

- **Alternatives to the computation of $s(\epsilon)$.** We mentioned above that at this point our theory is descriptive rather than prescriptive. It would be nice to at least partially remedy this situation. The main barrier is our assumption that $s(\epsilon)$ is known to the designer of a learning algorithm, which in turn implies knowledge of the input distribution. Might it be possible to estimate $s(\epsilon)$ from data, even for special function classes of interest? If one has only partial information about the input distribution, can this be translated into useful partial information about $s(\epsilon)$. Note that such considerations must be central to any attempt to apply our theory in a practical manner, for instance to structural risk minimization.

Acknowledgements

We are grateful to Haim Sompolinsky and Vladimir Vapnik for enlightening conversations and helpful comments. We would also like to thank Chris van den Broeck for organizing the Workshop on Statistical Mechanics of Generalization at Alden Biesen. We are grateful for the support of NSF grant IRI-9123692 and the U.S.–Israel BSF grant 90-0189.

References

[1] S. Amari, N. Fujita, and S. Shinomoto. Four types of learning curves. *Neural Computation*, 4(4):605–618, 1992.

[2] E. B. Baum and Y.-D. Lyuu. The transition to perfect generalization in perceptrons. *Neural Comput.*, 3:386–401, 1991.

[3] G. Benedek and A. Itai. Learnability with respect to fixed distributions. *Theoret. Comput. Sci.*, 86(2):377–389, 1991.

[4] D. Cohn and G. Tesauro. How tight are the Vapnik-Chervonenkis bounds. *Neural Comput.*, 4:249–269, 1992.

[5] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[6] L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. 1994. Preprint.

[7] R. M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6(6):899–929, 1978.

[8] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–251, 1989.

[9] A. Engel and W. Fink. Statistical mechanics calculation of Vapnik Chervonenkis bounds for perceptrons. *J. Phys.*, 26:6893–6914, 1993.

[10] A. Engel and C. van den Broeck. Systems that can learn from examples: replica calculation of uniform convergence bounds for the perceptron. *Phys. Rev. Lett.*, 71:1772–1775, 1993.

[11] E. Gardner. The space of interactions in neural network models. *J. Phys.*, A21:257–270, 1988.

[12] E. Gardner and B. Derrida. Three unfinished works on the optimal storage capacity of networks. *J. Phys.*, A22:1983–1994, 1989.

[13] S. A. Goldman, M. J. Kearns, and R. E. Schapire. On the sample complexity of weak learning. In *Proceedings of the 3rd Workshop on Computational Learning Theory*, pages 217–231. Morgan Kaufmann, San Mateo, CA, 1990.

[14] G. Györgyi. First-order transition to perfect generalization in a neural network with binary synapses. *Phys. Rev.*, A41:7097–7100, 1990.

[15] G. Gyorgyi and N. Tishby. Statistical theory of learning a rule. In K. Thuemann and R. Koeberle, editors, *Neural Networks and Spin Glasses*. World Scientific, 1990.

[16] D. Haussler. Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[17] D. Haussler, M. Kearns, and R. E. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. In *Proceedings of the 4th Workshop on Computational Learning Theory*, pages 61–74. Morgan Kaufmann, San Mateo, CA, 1991.

[18] E. Levin, N. Tishby, and S. Solla. A statistical approach to learning and generalization in neural networks. In R. Rivest, editor, *Proc. 3rd Annu. Workshop on Comput. Learning Theory*. Morgan Kaufmann, 1989.

[19] Y.-D. Lyuu and I. Rivin. Tight bounds on transition to perfect generalization in perceptrons. *Neural Comput.*, 4:854–862, 1992.

[20] G. L. Martin and J. A. Pittman. Recognizing hand-printed letters and digits using backpropagation learning. *Neural Comput.*, 3:258–267, 1991.

[21] E. Oblow. Implementing Valiant's learnability theory using random sets. *Machine Learning*, 8(1):45–74, 1992.

[22] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[23] W. Sarrett and M. Pazzani. Average case analysis of empirical and explanation-based learning algorithms. *Machine Learning*, 9(4):349–372, 1992.

[24] D. B. Schwartz, V. K. Samalam, J. S. Denker, and S. A. Solla. Exhaustive learning. *Neural Comput.*, 2:374–385, 1990.

[25] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review*, A45:6056–6091, 1992.

[26] H. U. Simon. General bounds on the number of examples needed for learning probabilistic concepts. In *Proceedings of the 6th Annual ACM Conference on Computational Learning Theory*, pages 402–411. ACM Press, New York, NY, 1993.

[27] H. Sompolinsky, H. S. Seung, and N. Tishby. Learning curves in large neural networks. In *Proc. 4th Annu. Workshop on Comput. Learning Theory*, pages 112–127. Morgan Kaufmann, San Mateo, CA, 1991.

[28] H. Sompolinsky, N. Tishby, and H. S. Seung. Learning from examples in large neural networks. *Phys. Rev. Lett.*, 65(13):1683–1686, 1990.

[29] V. Vapnik, E. Levin, and Y. LeCun. Measuring the VC dimension of a learning machine. *Neural Comput.*, 1994. To appear.

[30] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[31] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[32] T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65:499–556, 1993.

A Technical Appendix

A.1 Relaxing the bound on the number of error levels

One undesirable aspect of the statement of Theorem 4 is the demand that $r(N) = o(e^{t(N)\Delta})$ for all values $\Delta > 0$, that is, the insistence that the number of error levels $r(N)$ be a strictly subexponential function of chosen scaling function $t(N)$. In this section we briefly show how this condition can be sidestepped without changing the essential character of the thermodynamic limit method. The basic idea is this: if the true number of error levels $r(N)$ is too large to apply Theorem 4, we can instead apply the theorem using a smaller number of error levels of our own choosing.

More precisely, rather than using the error levels ϵ_j^N, $1 \leq j \leq r(N)$, that are determined by the definition of the $\mathcal{F}_N$, f_N and D_N, let us instead let $r(N)$ be *any* function meeting the condition $r(N) = o(e^{t(N)\Delta})$ for all values $\Delta > 0$, and let the ϵ_j^N be *any* sequence of error values that we choose. Thus, now there may in fact be *no* functions in $\mathcal{F}$ at generalization error ϵ_j^N. We now redefine Q_j^N to be all those functions in $\mathcal{F}_N$ whose generalization error falls in the interval $[\epsilon_j^N, \epsilon_{j+1}^N)$. The intuition is that we are first putting functions of nearby generalization error in the same "bin", and assuming (pessimistically) that all functions in the same bin have the smallest possible generalization error for this bin.

The definition of a permissible entropy bound $s(\epsilon)$ with respect to the scaling function $t(N)$ remains unaltered, and it can be verified that under the new definitions, Theorem 4 still holds. Given a scaling function $t(N)$, the number and spacing of the error levels we should choose to obtain the best analysis depends on the problem. A natural choice is to space the error levels evenly over $[0,1]$, but this is not the only possibility and may not be the best one for certain problems.

A.2 Derivation of general energy bound form

Here we show how Equations (30) and (54) can be derived.

Lemma 9 *(Sanov) Let $Z_1, \ldots, Z_m$ be i.i.d. random variables taking on the values $\{-1, 0, 1\}$ with probabilities $\{p_{-1}, p_0, p_1\}$, resp. If the mean $p_1 - p_{-1}$ of Z_i is positive, then the probability that the empirical mean is nonpositive is bounded by*

$$\mathbf{Pr}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i \leq 0\right] \leq \left(1 - (\sqrt{p_1} - \sqrt{p_{-1}})^2\right)^m \quad (61)$$

$$\leq \exp\left(-\frac{m(p_1 - p_{-1})^2}{2(p_1 + p_{-1})}\right) \quad (62)$$

Proof: Let $T = \frac{1}{m}\sum_{i=1}^{m} Z_i$ be the empirical mean. Then from Markov's inequality it follows that

$$\mathbf{Pr}[T \leq 0] = \mathbf{Pr}[e^{-m\lambda T} \geq 1] \quad (63)$$

$$\leq \mathbf{E}[e^{-m\lambda T}] \quad (64)$$

$$= \prod_{i=1}^{m} \mathbf{E}[e^{-\lambda Z_i}] \quad (65)$$

$$= (p_1 e^{-\lambda} + p_0 + p_{-1} e^{\lambda})^m , \quad (66)$$

for any positive λ. In particular, it is true for the λ^* satisfying $e^{-\lambda^*} = \sqrt{p_{-1}/p_1}$. Making this substitution and using

$p_0 = 1 - p_1 - p_{-1}$, we find the first inequality of the lemma. The second inequality follows from

$$\begin{aligned} \frac{(p_1 - p_{-1})^2}{p_1 + p_{-1}} &= (\sqrt{p_1} - \sqrt{p_{-1}})^2 \frac{(\sqrt{p_1} + \sqrt{p_{-1}})^2}{p_1 + p_{-1}} && (67) \\ &\leq 2(\sqrt{p_1} - \sqrt{p_{-1}})^2 && (68) \\ &\leq -2\log\left(1 - (\sqrt{p_1} - \sqrt{p_{-1}})^2\right) && (69) \end{aligned}$$

□

To prove Equation (30) using this lemma, we note that the random variable $\epsilon_{trn}(h, S) - \epsilon_{trn}(h^*, S)$ is precisely the empirical mean of the random variables

$$Z_i = \chi[h(x_i) \neq y_i] - \chi[h^*(x_i) \neq y_i], \tag{70}$$

where each $\langle x_i, y_i \rangle$ is an example drawn independently from D_N. Each Z_i takes on the values $\{-1, 0, 1\}$ with probabilities

$$\begin{aligned} p_1 &= \mathbf{Pr}[(h(x) \neq y) \wedge (h^*(x) = y)] && (71) \\ p_0 &= \mathbf{Pr}[(h(x) \neq y) \wedge (h^*(x) \neq y)] \\ &\quad + \mathbf{Pr}[(h(x) = y) \wedge (h^*(x) = y)] && (72) \\ p_{-1} &= \mathbf{Pr}[(h(x) = y) \wedge (h^*(x) \neq y)] && (73) \end{aligned}$$

where $\langle x, y \rangle$ is an example drawn randomly from D_N. These are related to probabilities of disagreement via

$$\begin{aligned} \epsilon(h, h^*) &= p_1 + p_{-1} && (74) \\ \epsilon(h) - \epsilon(h^*) &= p_1 - p_{-1} && (75) \end{aligned}$$

Making the appropriate substitutions in Equation (62) yields the desired result.

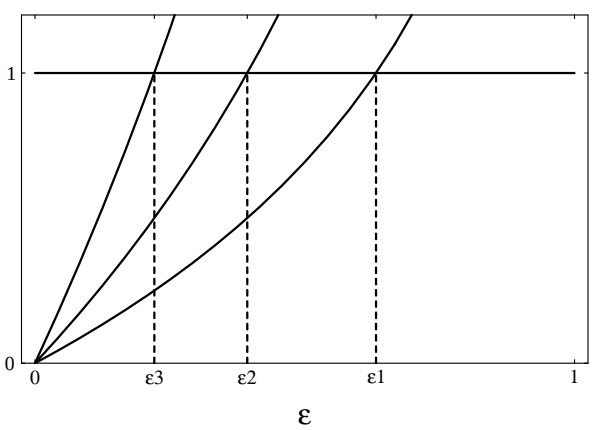

Figure 1: Rightmost intersections for a constant entropy bound $s(\epsilon) = 1$ and $-\alpha \log(1 - \epsilon)$ for three values $\alpha = \alpha_1, \alpha_2, \alpha_3$.

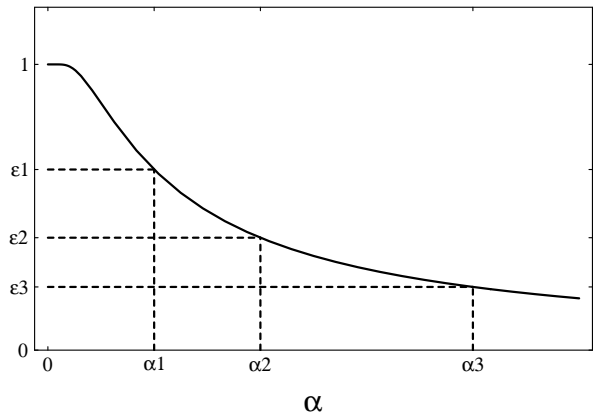

Figure 2: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 1.

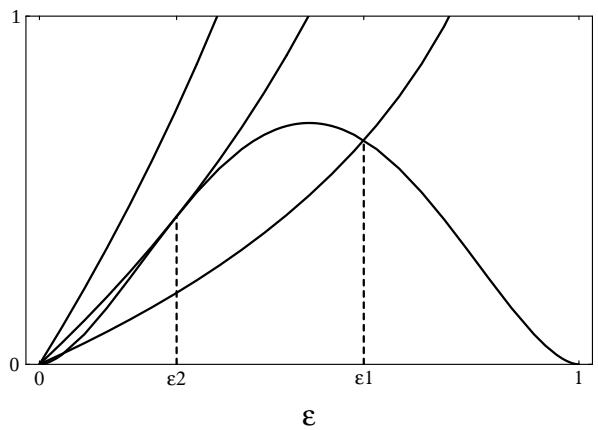

Figure 3: Rightmost intersections for a single-peak entropy bound (for the Ising perceptron of Section 2.6) and $-\alpha \log(1 - \epsilon)$. The curves corresponding to the three values $\alpha_1 = 0.7$, $\alpha_2 = 1.448$ and $\alpha_3 = 2.5$ are plotted. The resulting three intersections are $\epsilon_1 = 0.6011$, $\epsilon_2 = 0.2543$ and 0. The value $\alpha_2 = 1.448$ is a critical value, resulting in the phase transition seen in Figure 4.

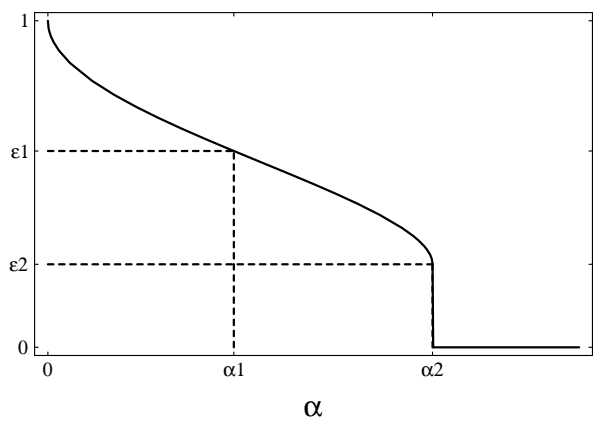

Figure 4: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 3 (Ising perceptron), showing a phase transition to zero error at the critical value $\alpha_2 = 1.448$.

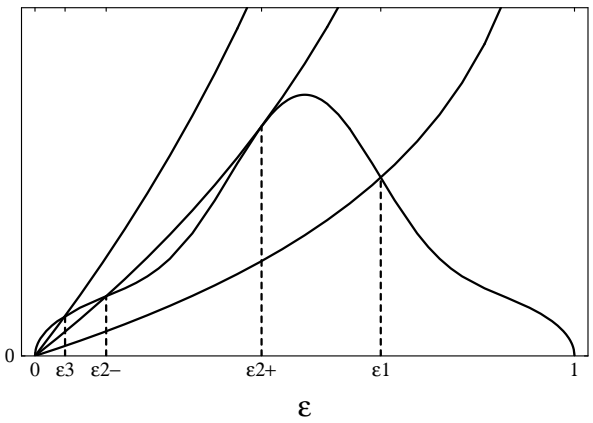

Figure 5: Rightmost intersections for a single-peak entropy bound and $-\alpha \log(1-\epsilon)$, showing a critical value α_2.

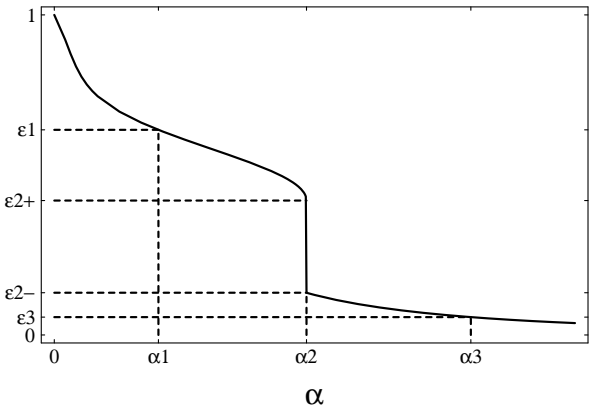

Figure 6: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 5, showing a phase transition to nonzero error at the critical value α_2.

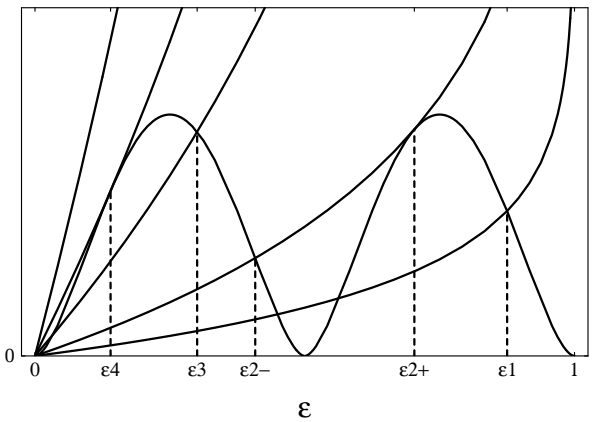

Figure 7: Rightmost intersections for a double-peak entropy bound and $-\alpha \log(1-\epsilon)$, showing critical values α_2 and α_4.

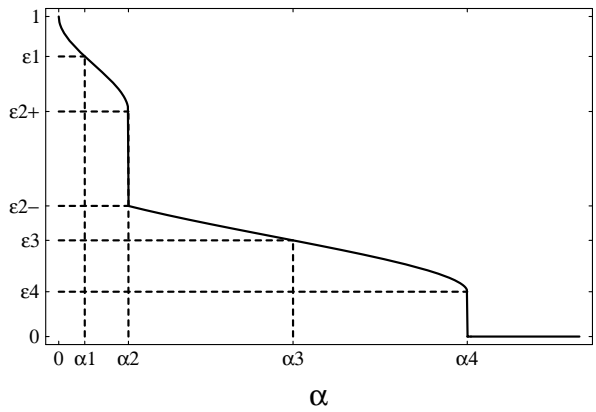

Figure 8: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 7, showing a phase transition to nonzero error at the critical value α_2, and a phase transition to 0 error at the critical value α_4.

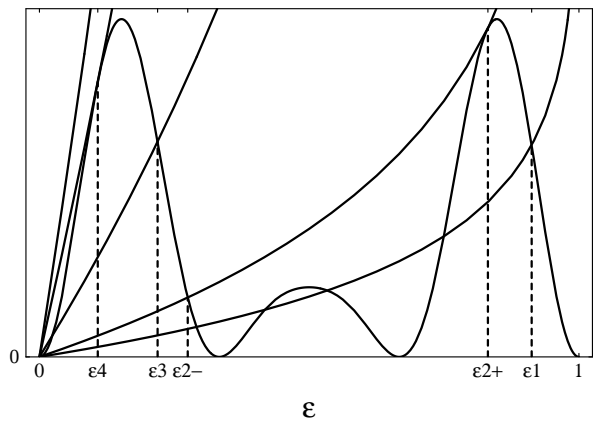

Figure 9: Rightmost intersections for a triple-peak entropy bound and $-\alpha \log(1-\epsilon)$, showing critical values at α_2 and α_4 and demonstrating the phenomenon of shadowing.

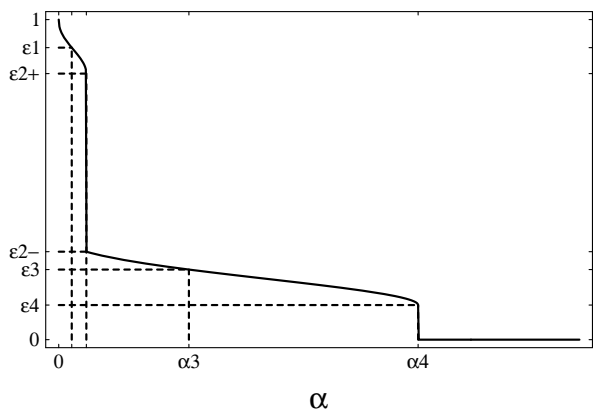

Figure 10: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 9, showing a phase transition to nonzero error at the critical value α_2, and a phase transition to 0 error at the critical value α_4.

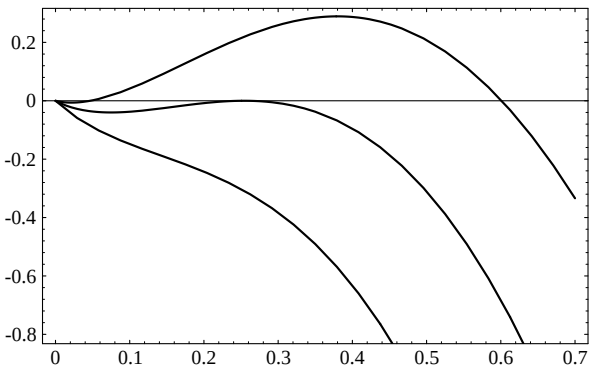

Figure 11: The function $s(\epsilon) + \alpha \log(1-\epsilon)$ for the Ising perceptron, plotted for the same values of $\alpha_1, \alpha_2, \alpha_3$ as in Figure 3.

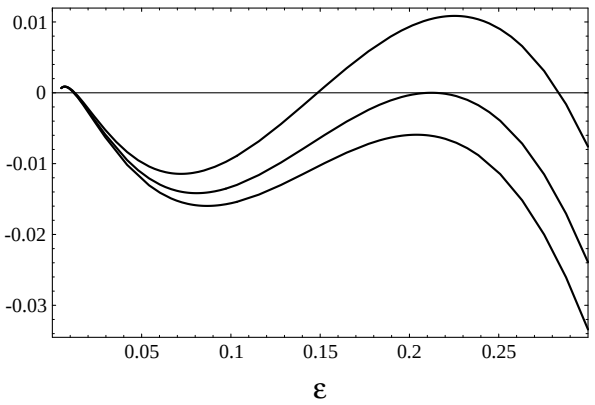

Figure 12: The function $s(\epsilon) - \alpha u(\epsilon)$ for the unrealizable Ising perceptron discussed in Section 3.3, with $\epsilon_{\min}(\gamma) = 0.005$. The function is plotted for the values $\alpha = 2.0, 2.063, 2.1$ (top to bottom).

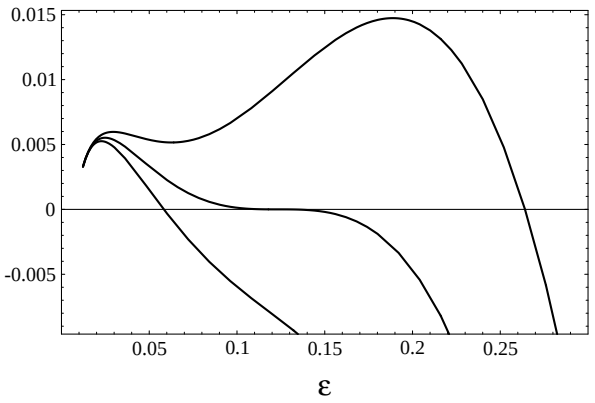

Figure 13: The function $s(\epsilon) - \alpha u(\epsilon)$ for the unrealizable Ising perceptron discussed in Section 3.3, with $\epsilon_{\min}(\gamma) = 0.01224$. This value for $\epsilon_{\min}(\gamma)$ is a critical value, in the sense that the learning curve phase transition disappears for larger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

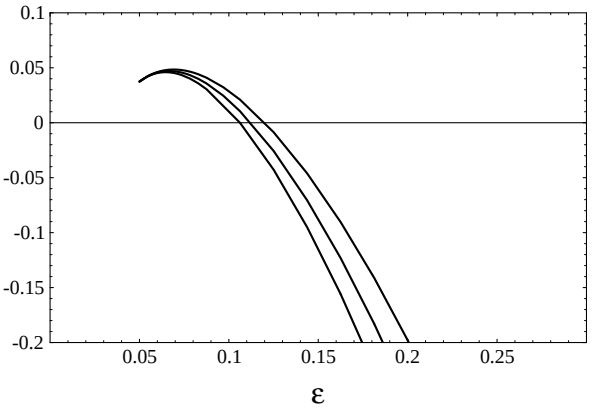

Figure 14: The function $s(\epsilon) - \alpha u(\epsilon)$ for the unrealizable Ising perceptron discussed in Section 3.3, with $\epsilon_{\min}(\gamma) = 0.05$. The function is plotted for the values $\alpha = 10, 11, 12$ (top to bottom).

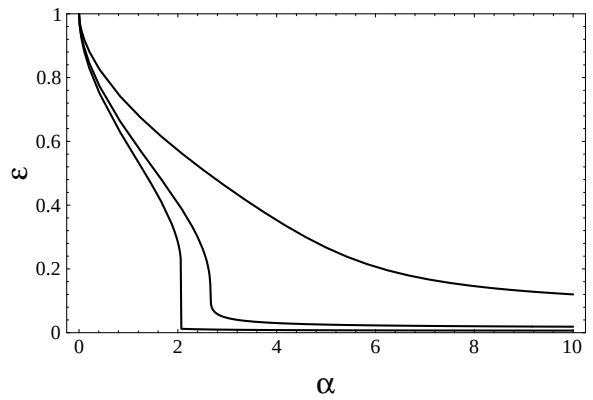

Figure 15: The scaled learning curves $\epsilon^*_\gamma(\alpha)$ for the unrealizable Ising perceptron discussed in Section 3.3, for the three values $\epsilon_{\min}(\gamma) = 0.005, 0.01224, 0.05$ (bottom to top).

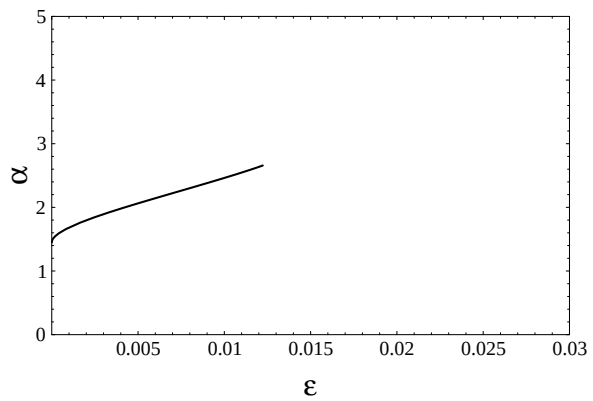

Figure 16: Phase diagram showing line of first-order transitions beginning at $\alpha = 1.448$ for $\epsilon_{\min}(\gamma) = 0$ and terminating at $\alpha = 2.659$ for $\epsilon_{\min}(\gamma) = 0.01224$.

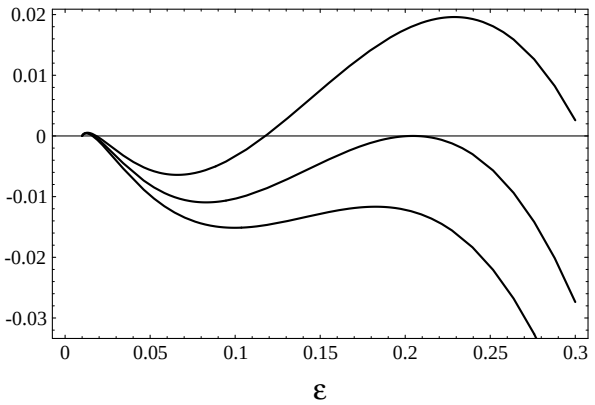

Figure 17: The function $s(\epsilon) - \alpha u(\epsilon)$ for the unrealizable Ising perceptron discussed in Section 3.4, with $\epsilon_{\min}(\gamma) = 0.01$. The function is plotted for the values $\alpha = 2.0, 2.1184, 2.2$ (top to bottom).

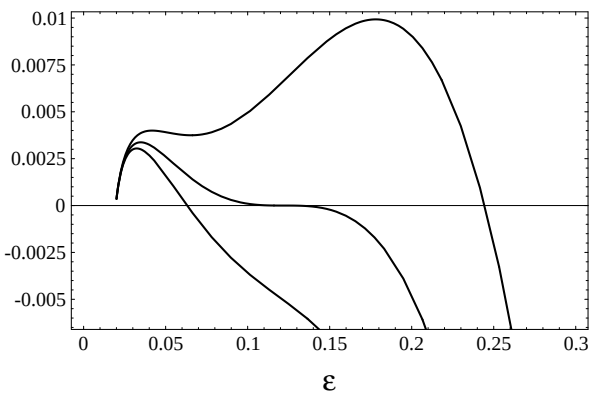

Figure 18: The function $s(\epsilon) - \alpha u(\epsilon)$ for the unrealizable Ising perceptron discussed in Section 3.4, with $\epsilon_{\min}(\gamma) = 0.01969$. This value for $\epsilon_{\min}(\gamma)$ is a critical value, in the sense that the learning curve phase transition disappears for larger $\epsilon_{\min}(\gamma)$. The function is plotted for the values $\alpha = 2.5, 2.6136, 2.7$ (top to bottom).

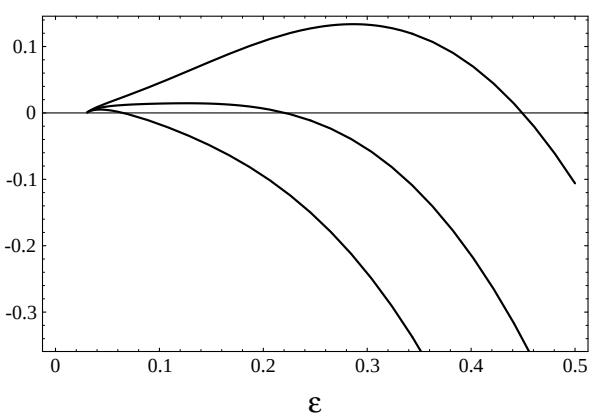

Figure 19: The function $s(\epsilon) - \alpha u(\epsilon)$ for the unrealizable Ising perceptron discussed in Section 3.4, with $\epsilon_{\min}(\gamma) = 0.03$. The function is plotted for the values $\alpha = 2, 3, 4$ (top to bottom).

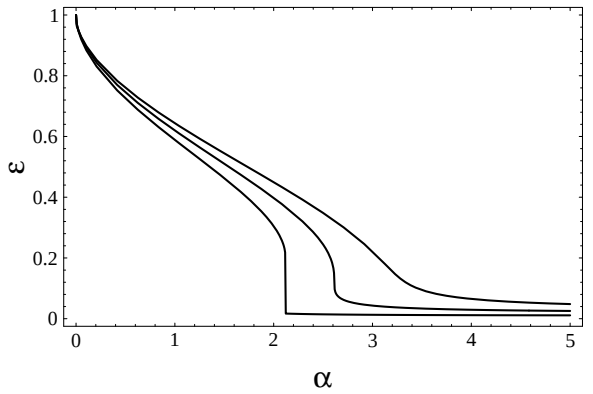

Figure 20: The scaled learning curves $\epsilon^*_\gamma(\alpha)$ for the unrealizable Ising perceptron discussed in Section 3.4, for the three values $\epsilon_{\min}(\gamma) = 0.01, 0.01969, 0.03$ (bottom to top).

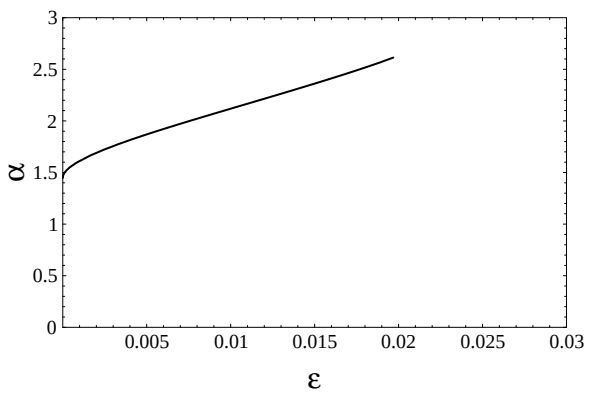

Figure 21: Phase diagram showing line of first-order transitions beginning at $\alpha = 1.448$ for $\epsilon_{\min}(\gamma) = 0$ and terminating at $\alpha = 2.6136$ for $\epsilon_{\min}(\gamma) = 0.01969$.