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 [27, 26], 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 [5, 7, 22].

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) [18], 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.

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) [26, 15, 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 [25].

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 [28]). This has included learning curves exhibiting "phase transitions" (sudden drops in the generalization error) at small sample sizes, as well as asymptotic power law behavior [2] in which the power law exponent is neither 1 nor 1/2. Although these learning curves do not contradict the VC bounds, it seems fair to say that their behavior is qualitatively different. The theoretical revisions of the VC theory mentioned above cannot explain such behavior, because they conservatively modify only with the

[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.

constant factors of the same power laws.

In this paper, we show that ideas from statistical mechanics (namely, the annealed approximation [16, 20, 1, 23] and the thermodynamic limit [23]) 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 in the full paper). 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.

[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 [10] 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 [9, 8]. 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.

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.

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

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

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 [26, 19, 6, 14] 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 the full paper, 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 full paper).

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 note that the condition on the growth rate of $r(N)$ can be greatly relaxed, and we do so in the full paper.

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.

It is worth noting that although we have been devoting our attention to the rightmost intersection, since this upper bounds the generalization error, the leftmost intersection also has a meaning. With high probability, there are no hypotheses in the version space with error less than the leftmost intersection except for the target itself. So the version space minus the target is contained within an annulus [8] whose inner and outer limits are the leftmost and rightmost intersections.

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 the full paper.

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 [11, 13, 24]. 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

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

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

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 [11], and extended to the case of boolean inputs by Baum, Lyuu and Rivin [2, 17]. 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 [13, 24, 21, 8]. [6]

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 $\{0,1\}^N$.

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 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

[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 $N = 32$.

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 [12] 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 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. In this section we sketch the extension of our theory to the case of an *unrealizable* target, in which there exists no function in $\mathcal{F}$ with zero generalization error; details are given in the full paper. As in the realizable case, learning curve bounds are found using a thermodynamic limit method to extract scaled learning curves. Of course, now the learning curve approaches $\epsilon_{\min} = \epsilon_{gen}(h^*)$ rather than 0 as the number of examples is increased, where we define

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

Recall that in the realizable case, we focused on bounding the error of any consistent algorithm. In the unrealizable case, we analyze algorithms which choose a hypothesis with minimum empirical (or training) error, the frequency of disagreement with the target on a sample S. 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 $\epsilon_{trn}(h, S)$:

$$VS(S) = \{h \in \mathcal{F} : \epsilon_{trn}(h, S) = \min_{h' \in \mathcal{F}} \epsilon_{trn}(h', S)\}. \tag{25}$$

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, 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, and we will only be able to upper bound this survival probability. 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 $\mathbf{Pr}_S[h \in VS(S)] \leq e^{-u(\epsilon_{gen}(h))m}$. 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. In the full paper we show that $u(\epsilon) = -\ln(1 - (\sqrt{\epsilon} - \sqrt{\epsilon_{\min}})^2)$ is a universally permissible energy bound. However, for a specific learning model a larger permissible $u(\epsilon)$ may typically be derived, and this may result in better learning curve bounds.

Once a permissible energy bound is obtained, the bound on the generalization error is then found as the rightmost zero crossing of $s(\epsilon) - \alpha u(\epsilon)$, just as in the realizable case (details are given in the full paper). As an illustrative example we consider 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, then Figure 11 shows the resulting scaled learning curves for three 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. Furthermore, the value $\epsilon_{\min}(\gamma) = 0.01224$ is a critical value, in the sense that the learning curve phase transition disappears for

larger $\epsilon_{\min}(\gamma)$. Figure 12 shows a *phase diagram* that plots the critical value of α for which the learning curve experiences a phase transition as a function of $\epsilon_{\min}(\gamma)$. As the best achievable error $\epsilon_{\min}(\gamma)$ increases, the location of the phase transition becomes progressively larger, in an essentially linear fashion. At the critical value $\epsilon_{\min}(\gamma) = 0.01224$, the phase transition disappears entirely, and for larger $\epsilon_{\min}(\gamma)$ the learning curves look progressively closer to power laws, as in the topmost learning curve of Figure 11.

4 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.

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

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

[7] 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.

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

[9] 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.

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

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

[12] 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.

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

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

[15] 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.

[16] 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.

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

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

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

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

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

[22] 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.

[23] 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.

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

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

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

[27] 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.

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

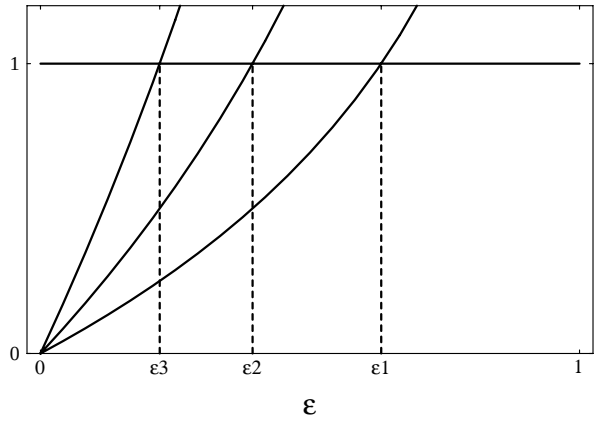

Figure 1: Rightmost intersections for a constant entropy bound $s(\epsilon) = 1$ and $-\alpha \log(1-\epsilon)$ for three values $\alpha = \alpha_1, \alpha_2, \alpha_3$.

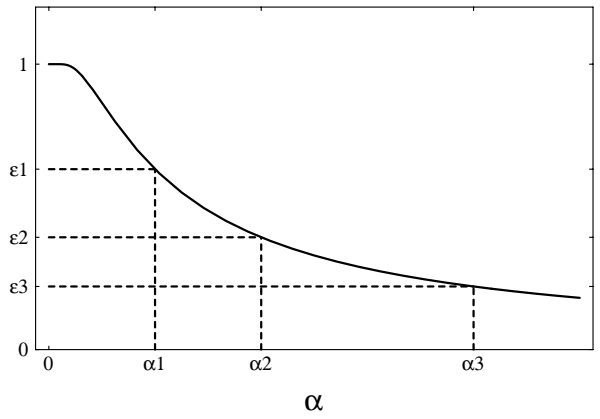

Figure 2: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 1.

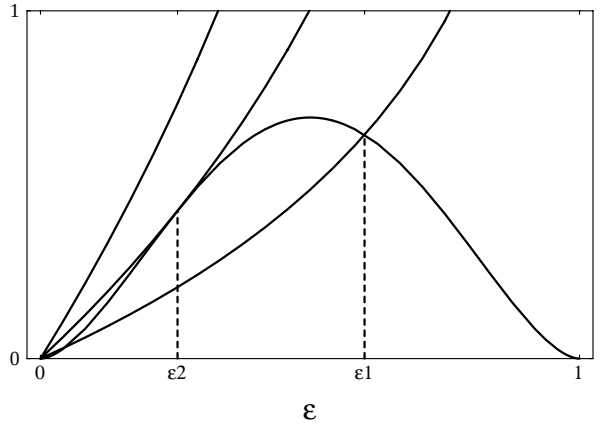

Figure 3: Rightmost intersections for a single-peak entropy bound (for the Ising perceptron of Section 2.6) and $-\alpha \log(1-\epsilon)$. The curves corresponding to the three values $\alpha_1 = 0.7$, $\alpha_2 = 1.448$ and $\alpha_3 = 2.5$ are plotted. The resulting three intersections are $\epsilon_1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

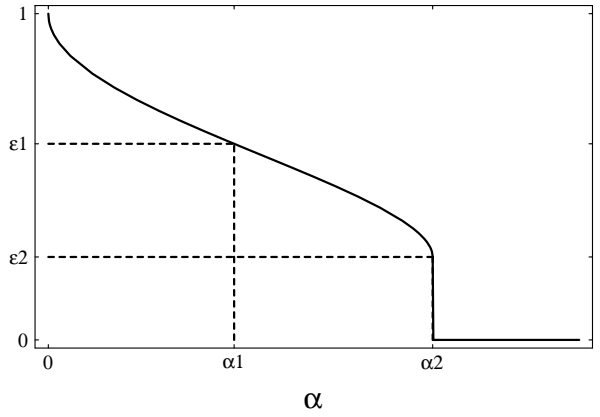

Figure 4: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 3 (Ising perceptron), showing a phase transition to zero error at the critical value $\alpha_2 = 1.448$.

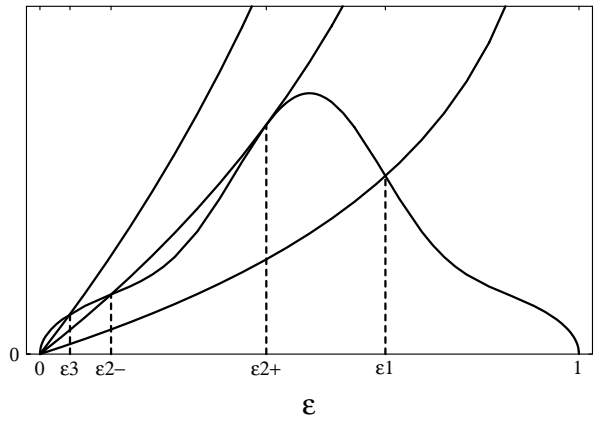

Figure 5: Rightmost intersections for a single-peak entropy bound and $-\alpha \log(1-\epsilon)$, showing a critical value α_2.

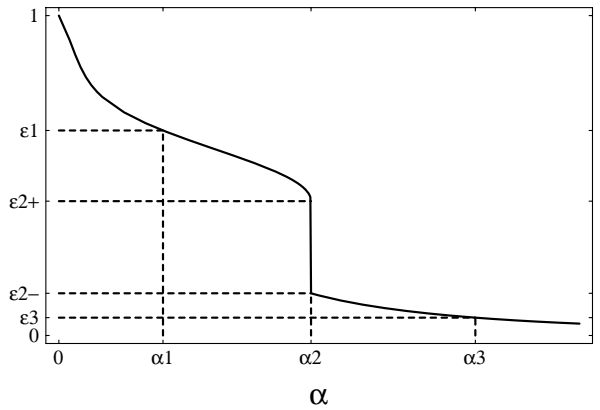

Figure 6: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 5, showing a phase transition to nonzero error at the critical value α_2.

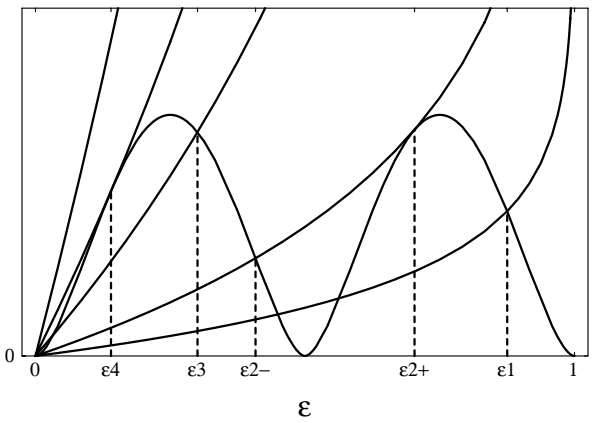

Figure 7: Rightmost intersections for a double-peak entropy bound and $-\alpha \log(1-\epsilon)$, showing critical values α_2 and α_4.

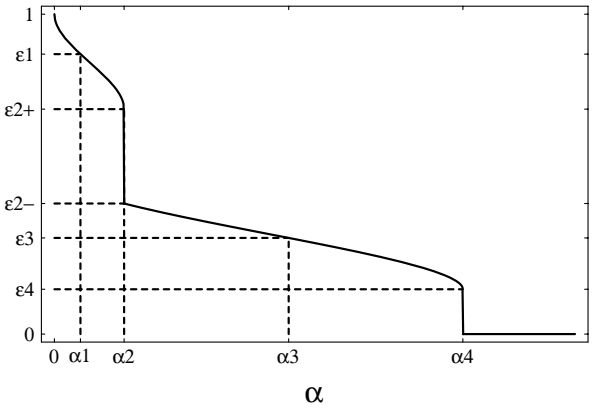

Figure 8: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 7, showing a phase transition to nonzero error at the critical value α_2, and a phase transition to 0 error at the critical value α_4.

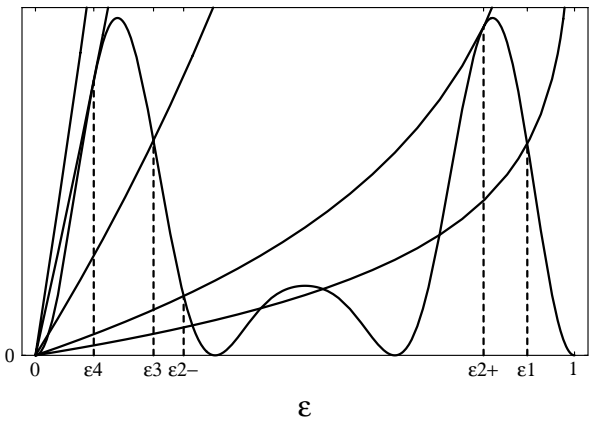

Figure 9: Rightmost intersections for a triple-peak entropy bound and $-\alpha \log(1-\epsilon)$, showing critical values at α_2 and α_4 and demonstrating the phenomenon of shadowing.

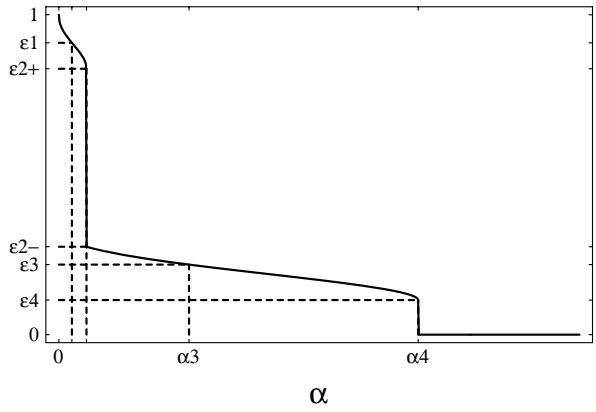

Figure 10: Scaled learning curve $\epsilon^*(\alpha)$ corresponding to the entropy-energy competition of Figure 9, showing a phase transition to nonzero error at the critical value α_2, and a phase transition to 0 error at the critical value α_4.

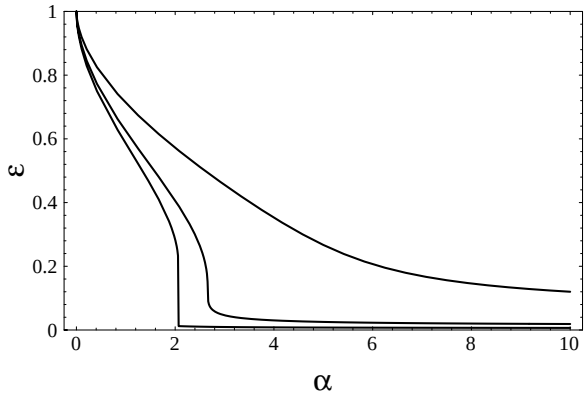

Figure 11: The scaled learning curves $\epsilon^*_\gamma(\alpha)$ for the unrealizable Ising perceptron discussed in Section 3, for the three values $\epsilon_{\min}(\gamma) = 0.005, 0.01224, 0.05$ (bottom to top). The value 0.01224 for $\epsilon_{\min}(\gamma)$ is a critical value, in the sense that the learning curve phase transition disappears for larger $\epsilon_{\min}(\gamma)$.

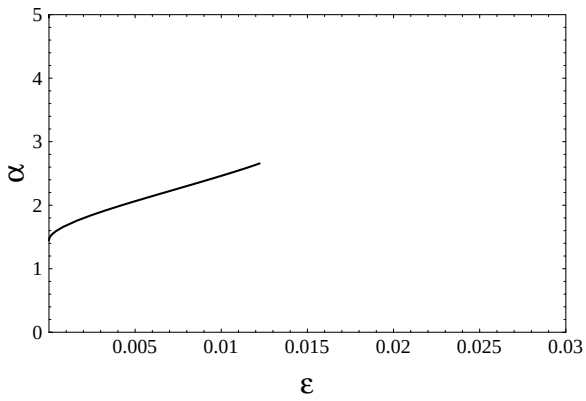

Figure 12: Phase diagram showing line of first-order transitions beginning at $\alpha = 1.448$ for $\epsilon_{\min}(\gamma) = 0$ and terminating at $\alpha = 2.659$ for $\epsilon_{\min}(\gamma) = 0.01224$.