JOURNAL OF COMPUTER AND SYSTEM SCIENCES 55, 140–160 (1997)
ARTICLE NO. SS971505

Dale Schuurmans*

Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4, Canada

Received March 10, 1997

We consider the standard problem of learning a concept from random examples. Here a *learning curve* is defined to be the expected error of a learner's hypotheses as a function of training sample size. Haussler, Littlestone, and Warmuth have shown that, in the distribution-free setting, the smallest expected error a learner can achieve in the worst case over a class of concepts C converges rationally to zero error; i.e., $\Theta(t^{-1})$ in the training sample size t. However, Cohn and Tesauro have recently demonstrated that *exponential* convergence can often be observed in experimental settings (i.e., average error decreasing as $e^{\Theta(-t)}$). By addressing a simple non-uniformity in the original analysis this paper shows how the dichotomy between rational and exponential worst case learning curves can be recovered in the distribution-free theory. In particular, our results support the experimental findings of Cohn and Tesauro: for finite concept classes any consistent learner achieves exponential convergence, even in the worst case, whereas for continuous concept classes no learner can exhibit sub-rational convergence for every target concept and domain distribution. We also draw a precise boundary between rational and exponential convergence for simple concept chains—showing that somewhere-dense chains always force rational convergence in the worst case, while exponential convergence can always be achieved for nowhere-dense chains.

1. INTRODUCTION

When learning a concept from random examples, we naturally expect the accuracy of a learner's hypotheses to improve as the learner sees more training examples. In some sense, it is the rate of this improvement that best describes the learning performance of the system. Given i.i.d. random examples, a hypothesizer's *learning curve* is defined by the expected error of its hypotheses as a function of training sample size. Intuitively, this is what one measures by repeatedly training a learning system on a fixed problem at different sample sizes and plotting the average hypothesis error that results as a function of training sample size. Since we anticipate that a learner's hypotheses will improve with increased training sample size, we measure the quality of a learning curve by the rate at which the average hypothesis error converges to zero.

* E-mail: dale@cs.toronto.edu, daes@linc.cis.upenn.edu. Current address: Institute for Research in Cognitive Science, University of Pennsylvania, 3401 Walnut Street, Suite 400A, Philadelphia, PA, 19104–6228.

Obviously, the quality of a hypothesizer's learning curve is determined by any prior knowledge it has about the underlying target concept and domain distribution. For example, if the exact target concept were known *a priori* then zero error is trivially achieved. Obtaining rapid convergence to zero error is more interesting if we know less about the target concept and domain distribution beforehand. Here we consider the model of prior knowledge popularized by Valiant [25]: we assume that the target concept is known to belong to some class C but nothing is known about the distribution of domain objects P, which could be arbitrary. Given this model we naturally consider what can be achieved in the "worst case, distribution-free" sense. Specifically, for a concept class C we are interested in determining the best learning curve that can be obtained in the worst case over all target concepts in C and all possible domain distributions P.

An analysis of this form has been carried out by Haussler *et al.* [14] who develop a special learning strategy 1IGPS (for "1-inclusion graph prediction strategy") that always achieves rational worst case convergence for any concept class C. Specifically, they show that, given t training examples, 1IGPS always attains an expected error of at most $O(t^{-1})$ for any target concept $c \in C$ and any domain distribution P (provided C has finite VC dimension [26]). Moreover, Haussler *et al.* [14] show that *no* learner can do better than this: given t training examples, any learner L must obtain an expected error of at least $\Omega(t^{-1})$ for some target concept $c_t \in C$ and some domain distribution P_t (provided C is non-trivial). This shows that the best achievable worst case expected error always behaves as a "rational" function of t (i.e., $\Theta(t^{-1})$) for any reasonable concept class.

1.1. *Issue*

These results would seem to suggest that we should always expect to observe rational learning curves, at least in the worst case (for any reasonable class C). However, one does not always observe rational learning curves in practice! This is clearly demonstrated in a recent study by Cohn and Tesauro which shows that *exponential* learning curves can be

0022-0000/97 $25.00

obtained in many experimental settings [6, 7]. Specifically, Cohn and Tesauro tested a backpropagation learning procedure, BP, on pairs of concept classes defined by identical neural network architectures on $\{0, 1\}^n$ and $[0, 1]^n$ respectively. Although defined on different input domains, these paired concept classes have identical VC dimension (in at least one case considered), and hence are *isomorphic* under the previous theory of [14]. However, in spite of this, Cohn and Tesauro observe dramatically different learning curves in each case: By repeatedly training BP at various sample sizes t and plotting the average error of its hypotheses, they invariably observe that

1. BP's average hypothesis error decreases as $\Theta(t^{-1})$ for *continuous* concept classes (defined by networks with $[0, 1]$ inputs), with a specific rate of decrease that closely matches the quantitative predictions of [14]; whereas,

2. BP's average hypothesis error decreases as $e^{\Theta(-t)}$ for *finite* concept classes (defined by networks with $\{0, 1\}$ inputs).

Interestingly, these results were obtained by considering the same set of target weights and uniform domain distributions in each case.

The exponential rate of decrease observed by Cohn and Tesauro is clearly contrary to the predictions of the previous HLW theory [14]. This raises the important question of understanding why the HLW theory fails to predict exponential learning curves even when they are readily observable in practice. One possible explanation is that the theory is worst case, whereas the experimental results depend on a particular setup. That is, since the experimental results only demonstrate exponential convergence for specific target concepts and domain distributions, it could still be the case that rational convergence is obtained in the absolute worst case. However, this does not explain why rational learning curves *are* typical in some circumstances (e.g., continuous concept classes) but not in others (e.g., finite concept classes). Simply claiming that the worst case theory fails to capture the typical situations encountered in practice is not accurate; sometimes it does. An adequate explanation of learning curve behavior must account for the robust dichotomy between rational and exponential learning curves that Cohn and Tesauro observe between continuous and finite concept classes.

It turns out that this discrepancy can be resolved by making a simple observation about the previous theory: A close inspection of the results of [14] reveals that the analysis is *non-uniform* in training sample size t. In particular, the lower bound result presented in [14] chooses a different domain distribution and possibly a different target concept for each training sample size t. Clearly, this does not reflect the situation normally encountered in practice, nor that investigated by Cohn and Tesauro, where these are held fixed. This raises the question of whether, in a model where the domain distribution and target concept are held fixed, there are situations where the best achievable worst case learning curve is exponential and other situations where it is rational. It turns out that the answer to both questions is *yes*.

1.2. *Results*

By performing an analysis of worst case learning curves where the domain distribution and target concept are held fixed, this paper shows how the dichotomy between rational and exponential convergence can be recovered in the worst case distribution-free setting. In particular, we show that there are concept classes that permit exponential learning curves, even in the worst case over all target concepts and domain distributions; but there are other concept classes that force any learner to produce rational learning curves for some fixed distributions and target concepts.

Our first results, presented in Section 3 below, establish the basic dichotomy between rational and exponential worst case learning curves. We first show (Proposition 1) that for any finite concept class C, any learner that guesses consistent concepts from C will obtain exponential convergence, even in the worst case. We then show (Proposition 2) that it is impossible to achieve better than exponential convergence in the worst case for any non-trivial concept class, so this must be the optimal form of worst case learning curve. However, exponential convergence cannot be guaranteed for every possible concept class: Next we show that *continuous* concept classes force any learner to obtain rational learning curves even for fixed target concepts and domain distributions. In particular, we show (Theorem 4) that for any continuous concept chain C there is a fixed domain distribution P that forces any learner to exhibit a rational convergence for a significant portion of targets in C. Note that this is a stronger result than the lower bound in [14], in the sense that it is easier to force bad behavior by choosing a different domain distribution and target concept for each training sample size t than it is to show bad behavior results even when these are held fixed. Note also how these results corroborate the experimental findings of Cohn and Tesauro: observing exponential learning curves for finite concept classes is no accident since this is achieved by any consistent learner. On the other hand, any continuous concept class forces rational convergence in the worst case.

Of course, there is still a significant gap between finite and continuous concept classes. For example, the results say nothing about what happens for countably infinite classes. This leaves open the question of identifying the exact conditions that dictate between rational and exponential worst case learning curves, and determining whether any other intermediate forms of worst case convergence are possible

under the distribution-free model (e.g., $\Theta(t^{-n})$ or some other form). It turns out that precise answers to these questions can be obtained for the special case of concept *chains*; i.e., classes which are linearly ordered under set inclusion.

In Section 4 below we draw an exact boundary between rational and exponential worst case learning curves for the special case of concept chains. Here we establish that the determining condition is the presence or absence of any *dense* subchains in the original concept chain (i.e., a subchain where between any two concepts there is a third). Specifically, we show (Theorem 5) that for any somewhere-dense concept chain C there is a domain distribution P that forces any learner to exhibit rational convergence for some targets in C. On the other hand, we also show (Theorem 6) that for any nowhere-dense chain C there is a special learning strategy CHOLC that always obtains exponential convergence, even in the worst case over all possible target concepts in C and domain distributions P. Together, these two results characterize the exact boundary between rational and exponential worst case convergence for concept chains. This also shows that no other form of worst case convergence is possible (for concept chains) in the distribution-free setting.

Finally, in Section 5 we address the point that the concept classes considered by Cohn and Tesauro in their computer simulations were represented with limited precision and hence fundamentally finite. This implies that all of the learning curves they observed must have been exponential. However, we demonstrate that, at the scale of training sample sizes considered, convergence can appear rational even when it is fundamentally exponential.

1.3. *Significance and Related Work*

These results show how the dichotomy between rational and exponential learning curves can be recovered in the distribution-free setting. Previous results on distribution-free learning curves [14] suggested that rational convergence was the only possible worst case form; however, this was based on a non-uniform analysis. By pursuing a uniform analysis, we are able to distinguish the conditions under which rational and exponential worst case convergence takes place based solely on the structure of the concept class C. This contradicts the common suggestion [13, 24] that such a characterization must take into account special properties of the domain distribution P.

Most theoretical studies of learning curve behavior adopt a distribution-*specific* model that assumes the learner knows the domain distribution P *a priori*. Given these stronger assumptions, many researchers have demonstrated that exponential learning curves are possible. For example, several researchers have analyzed the behavior of particular learning procedures for specific concept classes and domain distributions and shown that exponential convergence results [10, 11, 16, 20]. Others have shown that even general learning procedures can obtain exponential learning curves for specific concept classes and domain distributions [3, 4, 13, 18]. This paper shows how, in a general way, exponential convergence can still be revealed under much weaker assumptions.

The largest body of work concerning the analysis of learning curves is the Bayesian statistical mechanical approach, which not only assumes that the learner knows the domain distribution *a priori*, but also has access to a *prior* distribution of possible target concepts. These analyses consider the *average case* learning curves obtained by the Bayes and Gibbs learning procedures. Here too rational convergence appears to be a natural form, as suggested in general by [1], proved more rigorously by [12], and demonstrated in a specific case study by [19]. Given the much stronger assumptions of this model, it is not too surprising that many researchers have indicated a dichotomy between rational and exponential average case learning curves [3, 23, 24]. The common suggestion is that this dichotomy is determined by the existence of "gaps" between target concepts. However, it is easy to refute this suggestion in general [9] (see also the counterexample in Section 6), so this really must be an informal observation. Our purpose here is to rigorously establish that this dichotomy already exists under the much weaker distribution-free model.

We also draw a clean boundary between the two types of convergence in terms of a simple structural property of concept classes; namely, the presence of a dense subchain. As most distribution-specific analyses address particular case studies, they do not provide general characterizations of the concept space properties that permit or prohibit exponential convergence. Interestingly, only two possible modes of worst case convergence appear possible in the distribution-free setting: rational and exponential (proved for concept chains, but only conjecture in general). This is unlike the distribution-specific case where all intermediate forms of convergence are evidently possible [13]. Clearly the distribution-specific analyses give tighter characterizations of the learning curves one might observe in practice, but require more problem-specific information [13], in fact, more than is generally available in practice. The benefits of the distribution-free theory are its wider range of applicability in practical situations.

Finally, note that from a practical perspective it is important to predict the specific *rates* of convergence not just the functional forms of learning curves. A reasonable characterization of convergence rates has obvious applications for choosing the "complexity" of a hypothesis class relative to the amount of available training data, or determining whether sufficient data are available to achieve desired error levels, etc. Another interesting application, considered by Vapnik *et al.* [27], is to estimate the

"effective" VC dimension of a learning machine by fitting empirical learning curves to a theoretically derived (rational) form. Of course, a necessary prerequisite for any practical characterization of empirical learning curves is predicting whether rational versus exponential convergence will take place: obviously one cannot accurately predict specific rates of convergence without being able to predict the underlying functional form of the learning curve.

We now turn to the development of the main results of this paper.

2. FORMAL MODEL

We are considering the standard problem of learning a concept from examples. Formally, we have a domain of objects X on which a target *concept* $c \subset X$ is defined. An *example* is a pair $\langle x, 1_c(x)\rangle$ specifying a domain object $x \in X$ and giving the value of c's indicator function 1_c at x. (For simplicity, we denote this training example $cx = \langle x, 1_c(x)\rangle$, and for a sequence of domain objects $\mathbf{x}^t = \langle x_1, ..., x_r\rangle$, we denote the corresponding sequence of training examples by $c\mathbf{x}^t = \langle\langle x_1, 1_c(x_1)\rangle, ..., \langle x_t, 1_c(x_t)\rangle\rangle$.) Formally, a *learner* L is just a mapping from training sequences $c\mathbf{x}^t$ to hypotheses $h \subset X$, i.e., $L: (X \times \{0, 1\})^* \to 2^X$. So we denote L's hypothesis for a training sequence $c\mathbf{x}^t$ by $L(c\mathbf{x}^t) = h \subset X$. Here we consider a *batch* training protocol, where after a fixed training period t, any hypothesis h produced by L is then tested *ad infinitum* on subsequent test examples. A hypothesis h makes classification error on any test example $cx = \langle x, c(x)\rangle$ for which $h(x) \neq c(x)$. We denote the entire set of such objects by $h \triangle c$ (the symmetric difference between h and c) and use $h \equiv c$ to denote $(h \triangle c)^c$.

As in most theoretical analyses of concept learning, we adopt the i.i.d. random example model, which assumes domain objects are independently generated by a fixed domain distribution P and labelled according to a fixed target concept c. This is a natural model of many practical learning situations where the sequence of training objects is unpredictable and there is no correlation between successive objects. Under this model, the *error* of a hypothesis h with respect to target concept c and domain distribution P is given by $\mathrm{P}(h \triangle c)$. Note that this defines a natural (pseudo)metric on the space of concepts $d_{\mathrm{P}}(h, c) = \mathrm{P}(h \triangle c)$, which gives a natural measure of distance between concepts.

Given this model, note that a learner L maps training sequences $c\mathbf{x}^t$ to hypotheses, and that each such hypothesis $L(c\mathbf{x}^t)$ will have an error with respect to the target concept c and domain distribution P. We will denote this error by $\mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) = d_{\mathrm{P}}(L(c\mathbf{x}^t), c)$. Now consider L's behavior for a fixed training sample size t: Since the training examples are i.i.d., a fixed domain distribution P induces a corresponding product distribution P^t on X^t, and a fixed target concept c then induces a fixed distribution over training sequences $c\mathbf{x}^t$. Thus, from this distribution over training sequences, L induces a distribution over hypotheses, and in turn a distribution over hypothesis *errors*. Therefore, for fixed c and P, we can determine the *expected* error of L's hypotheses given t training examples by

$$\mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) = \int_{X^t} \mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t)\, d\mathrm{P}^t(\mathbf{x}^t). \quad (1)$$

Given these definitions we define L's *learning curve* with respect to c and P by its expected error, $\mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(L, \mathrm{P}, \mathbf{x}^t)$, as a function of the training sample size t.

As mentioned, the quality of a hypothesizer's learning curve depends strongly on its prior knowledge about the domain distribution and target concept. Here we adopt the simple model of prior knowledge introduced by Valiant [25], which assumes the extent of a learner's prior knowledge can be captured solely by a class C to which the target concept is known to belong. Given this model, we are interested in determining the best learning curve that can be achieved in the worst case over a class of concepts C, given arbitrary domain distributions P. Below we pursue a *uniform* analysis of this question, where the domain distribution and target concept are held fixed for all training sample sizes t. Since we are primarily concerned with the distinction between rational and exponential learning curves, we also focus on the worst case *asymptotic* form of a learner's curve. Thus, for a concept class C, we investigate the worst case asymptotic form of learning curve a learner obtains for *fixed* target concepts $c \in C$ and domain distributions P:

DEFINITION 1 (Worst Case Learning Curve). We say that a concept class C has a $\Theta'(g(t))$ worst case learning curve, written $\mathrm{LC}(C) = \Theta'(g(t))$, if

1. there *exists* a learner L that achieves $\mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) = O(g(t))$ for every target concept $c \in C$ and domain distribution P; and

2. for *every* learner L, there is a target concept $c' \in C$ and a domain distribution P that forces L to obtain $\mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(L, \mathrm{P}, c', \mathbf{x}^t) = \Omega'(g(t))$.

Thus we say $\mathrm{LC}(C) = \Theta'(g(t))$ if some learner achieves $O(g(t))$ worst case convergence, but every learner can be forced to have an expected error of at least $\Omega'(g(t))$ for some fixed domain distribution P and target concept $c' \in C$. Here the notation $f(t) = \Omega'(g(t))$ means there exists a constant α such that $f(t) \geqslant \alpha g(t)$ for *infinitely many* $t > 0$. We use this weaker definition instead of the standard "for all but finitely many $t > 0$" because there is no way to prevent a learner from periodically guessing right on large training sample sizes. That is, we want to rule out the case where a learner systematically cycles through a finite (or countable) concept class, ignores the training data, and yet periodically achieves zero error for any target in the class.

3. BASIC DICHOTOMY

We first establish the basic dichotomy between rational and exponential worst case learning curves in the distribution-free model. Here we show that any finite concept class has an exponential worst case learning curve, whereas any continuous concept class forces rational convergence in the worst case.

3.1. *Finite Concept Classes*

First observe that for any finite concept class C, it is reasonably obvious that any consistent learner always obtains an exponential learning curve for any fixed target concept and domain distribution. (We say that a learner L is *consistent for* a class C if, given any sequence of training examples $c\mathbf{x}^t$ generated by some $c \in C$, L produces a hypothesis $L(c\mathbf{x}^t) \in C$ that correctly classifies every training example in $c\mathbf{x}^t$.)

PROPOSITION 1 (Finite UB). *For any finite concept class C: any consistent learner L for C obtains an exponential learning curve (i.e., $\mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) = e^{O(-t)}$) for every target concept $c \in C$, regardless of the domain distribution P.*

Proof. Fix an arbitrary target concept $c \in C$ and an arbitrary domain distribution P. There will be at most $N = |C| - 1$ non-zero difference sets $D_0 = \{c \triangle c_i : \mathrm{P}(c \triangle c_i) > 0\}$. Let p_0 be the minimum such probability. Then the probability that some difference set remains unobserved after t training examples is at most $N(1 - p_0)^t$. Note that observing a domain object from each difference set implies that a consistent learner L for C will produce a hypothesis with zero error. Therefore, $\mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) \leqslant N(1 - p_0)^t = e^{O(-t)}$. ∎

Of course this is not a practically useful bound in that it does not accurately reflect the learning curves observed in practice. However, it does establish that exponential convergence can indeed take place in the distribution-free model. (The question of how to achieve more accurate bounds on the specific rate of convergence is raised below and discussed in detail in Section 6.)

Given that we have demonstrated the possibility of achieving exponential worst case convergence for finite concept classes, we now show that it is impossible to achieve better than exponential learning curves for any non-trivial concept class C. (A class C is said to be *non-trivial* if it contains at least two concepts $c, d \in C$ such that $c \triangle d \neq \varnothing$ and $c \equiv d \neq \varnothing$.) That is, for a non-trivial class C we can always find a domain distribution P that forces any learner to obtain an exponential learning curve for some fixed target concept in C.

PROPOSITION 2 (Universal LB). *For any non-trivial concept class C: there is a domain distribution* P *that forces any learner L to obtain an exponential learning curve (i.e., $\mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(L, \mathrm{P}, c', \mathbf{x}^t) = e^{\Omega'(-t)}$) for some target concept $c' \in C$.*

Proof. Since C is non-trivial there must be two concepts $c_1, c_2 \in C$ such that $(c_1 \triangle c_2) \neq \varnothing$ and $(c_1 \equiv c_2) \neq \varnothing$. Therefore we can fix a domain distribution P such that $d_{\mathrm{P}}(c_1, c_2) = p$ for some $0 < p < 1$. (E.g., choose $x_1 \in (c_1 \triangle c_2)$ and $x_2 \in (c_1 \equiv c_2)$, and set $\mathrm{P}(x_1) = p$ and $\mathrm{P}(x_2) = 1 - p$). For this distribution, given any training sample size t, any learner L must obtain

$$
\begin{aligned}
&\operatorname*{avg}_{c_i \in \{c_1, c_2\}} \mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(L, \mathrm{P}, c_i, \mathbf{x}^t) \\
&\quad \geqslant \tfrac{1}{2}\mathrm{E}_{\mathbf{x}^t}[\mathrm{err}(L, \mathrm{P}, c_1, \mathbf{x}^t) \\
&\qquad + \mathrm{err}(L, \mathrm{P}, c_2, \mathbf{x}^t) \,|\, c_1\mathbf{x}^t = c_2\mathbf{x}^t]\ \mathrm{P}^t(c_1\mathbf{x}^t = c_2\mathbf{x}^t) \\
&\quad = \tfrac{1}{2}\mathrm{E}_{\mathbf{x}^t}[\mathrm{err}(L, \mathrm{P}, c_1, \mathbf{x}^t) \\
&\qquad + \mathrm{err}(L, \mathrm{P}, c_2, \mathbf{x}^t) \,|\, c_1\mathbf{x}^t = c_2\mathbf{x}^t](1 - p)^t \\
&\quad \geqslant \tfrac{1}{2}\, p(1 - p)^t = e^{\Omega(-t)}.
\end{aligned}
$$

The last inequality holds since for any $\mathbf{x}^t$ such that $c_1\mathbf{x}^t = c_2\mathbf{x}^t$ we get $L(c_1\mathbf{x}^t) = L(c_2\mathbf{x}^t) = h$, and hence $d_{\mathrm{P}}(h, c_1) + d_{\mathrm{P}}(h, c_2) \geqslant d_{\mathrm{P}}(c_1, c_2) = p$ by the triangle inequality. Finally, note that obtaining an average expected error of at least $e^{\Omega(-t)}$ for every t implies that L must obtain at least this expected error on one of c_1 or c_2 for infinitely many t. ∎

This shows that exponential convergence is in fact the best achievable form of worst case learning curve in the distribution-free model for any non-trivial concept class. So, from Propositions 1 and 2, we have that any non-trivial, finite concept class C has exactly an exponential worst case learning curve, and this is achieved by any consistent learner L for C.

COROLLARY 1. *Any non-trivial, finite concept class C has an exponential worst case learning curve: $\mathrm{LC}(C) = e^{\Theta'(-t)}$.*

It is interesting to observe now these results compare to the non-uniform theory of [14]. Although we obtain exponential learning curves for any fixed domain distribution, it turns out that there is no single "worst case" distribution that maximizes the expected error for all training sample sizes t. That is, we obtain a different worst case domain distribution for each training sample size t. So although each individual curve is exponential, any universal upper bound over all curves happens to be rational. Figure 1 illustrates this discrepancy between the worst case bounds of [14], which consider a different domain distribution P_t for each t, and Proposition 1, which considers a single P for all t. This figure also illustrates how the precise learning curve obtained depends strongly on the specific domain distribution and target concept under consideration. This means that for exponential learning curves we

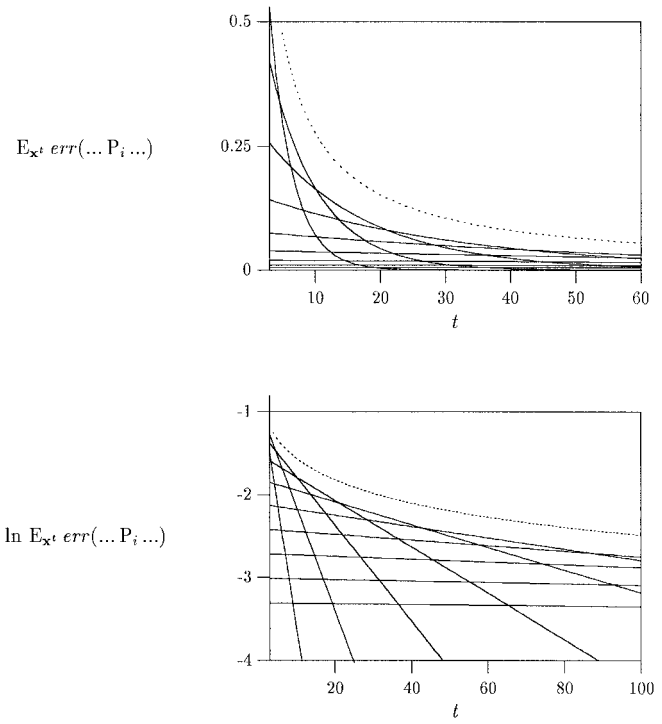

FIG. 1. Comparing uniform versus non-uniform learning curve bounds. This illustrates how a series of exponential learning curves determined by distributions P_1, P_2, ..., *etc.*, actually has a rational upper envelope.

cannot expect to obtain a tight characterization of specific convergence rates in the distribution-free model. Instead we know that such an analysis would require us to take additional distribution and target information into account beyond just the structure of the concept class C (this point is revisited in Section 6 below).

3.2. *Continuous Concept Chains*

We have shown that a uniform analysis can yield exponential learning curves for certain concept classes, even though the non-uniform HLW theory [14] predicts only rational forms. This raises the question of whether every concept class permits exponential convergence under a uniform analysis, or whether there are concept classes that can still force rational learning curves even for fixed domain distributions and target concepts. Here we show that there are indeed concept classes that force rational convergence, even in a model where the target concept and domain distribution are held fixed. In particular, this is true of any *continuous* concept class.

DEFINITION 2 (Chains and Continuity). A *concept chain* is a class C that is totally-ordered under set-inclusion; i.e., for every distinct c_1 and c_2 in C, either $c_1 \subset c_2$ or $c_1 \supset c_2$. A *continuous chain* is a concept chain that is order-isomorphic to $\mathbb{R}$; i.e., a chain that can be indexed $C = \{c_y : y \in \mathbb{R}\}$ such that $c_y \subset c_z$ for $y \leqslant z$. (A simple example of a continuous chain is the class of initial segment concepts on $[0, 1]$.)

Below we establish that any continuous concept class forces at least rational convergence in the worst case, even under a uniform analysis. However, before proving this, we first observe that any consistent learner already achieves rational worst case convergence for any concept chain.

PROPOSITION 3 (Chain UB). *For any concept chain C: any consistent learner L for C obtains a rational learning curve (i.e., $E_{\mathbf{x}^t} \mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) = O(t^{-1})$ for every target concept $c \in C$, regardless of the domain distribution* P.

Proof. Since any non-trivial chain obviously has VC dimension 1, the results of Haussler *et al.* [14] show that the special learning strategy 1IGPS obtains $\mathrm{E}_{\mathbf{x}^t} \mathrm{err}(1\mathsf{IGPS}, \mathrm{P}, c, \mathbf{x}^t) = O(t^{-1})$ in this case. Also, by [14, Theorem 6.1] we know that any consistent learner L for C must obtain $E_{\mathbf{x}^t} \mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) = O((\ln t)/t)$. Here, we strengthen these results slightly by showing that any consistent learner L for C actually obtains $\mathrm{E}_{\mathbf{x}^t} \mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) = O(t^{-1})$. Proving this also allows us to introduce some definitions and notation that will be useful later.

DEFINITION 3 (Uncertainty Interval). For a concept chain C, notice that any training sequence $c\mathbf{x}^t$ determines an *uncertainty interval* about the target concept c given by $[s(c\mathbf{x}^t), \ell(c\mathbf{x}^t)] = \{h \in C : s(c\mathbf{x}^t) \subset h \subset \ell(c\mathbf{x}^t)\}$, where $s(c\mathbf{x}^t)$ and $\ell(c\mathbf{x}^t)$ are the smallest and largest concepts consistent with $c\mathbf{x}^t$ respectively. Formally, we define the *smallest* concept consistent with a training example cx by $s(cx) = \varnothing$ if $x \notin c$, $s(cx) = \bigcap \{h \in C : h \subset c \text{ and } x \in h\}$ if $x \in c$; and the *largest* consistent concept by $\ell(cx) = X$ if $x \in c$, $\ell(cx) = \bigcup \{h \in C : h \supset c \text{ and } x \notin h\}$ if $x \notin c$. Then, for a sequence $c\mathbf{x}^t$ we define $s(c\mathbf{x}^t) = \bigcup_{x \in \mathbf{x}^t} s(cx)$ and $\ell(c\mathbf{x}^t) = \bigcap_{x \in \mathbf{x}^t} \ell(cx)$. Any uncertainty interval $[s(c\mathbf{x}^t), \ell(c\mathbf{x}^t)]$ has a *width* with respect to a domain distribution P given by $\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t) = \mathrm{P}(\ell(c\mathbf{x}^t) - s(c\mathbf{x}^t))$.

Thus for a concept chain, we can think of the training examples as monotonically reducing the interval of uncertainty about an unknown target concept. Clearly, any consistent learner L for C must guess a hypothesis from this interval, so the error of L's hypothesis must be bounded by the interval width: $\mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) \leqslant \mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t)$. Therefore, all that remains to show is that $\mathrm{E}_{\mathbf{x}^t} \mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t) = O(t^{-1})$. To this, we make the observation that the worst case situation is represented by the uniform chain.

DEFINITION 4 (Uniform Chain). For a domain $X = [0, 1]$, let $I = \{\mathbf{i} = [0, i] : i \in [0, 1]\}$ be the class of initial segment concepts $\mathbf{i}$ on $[0, 1]$, and let U denote the uniform distribution over $[0, 1]$. Then the *uniform chain* is the concept space (I, U) formed from I and U. (Note that we can think of a class C and distribution P as comprising a metric *space* (C, P) with inter-concept distances given by the metric d_{P}.)

Note that the uniform chain (I, U) satisfies the identity $d_{\mathrm{U}}(\mathbf{i}, \mathbf{j}) = |i - j|$.

LEMMA 1. *For any* $c \in (C, \mathrm{P})$ *there is an* $\mathbf{i} \in (I, \mathrm{U})$ *such that*

$$\mathrm{E}_{\mathbf{x}^t} \operatorname{wid}(C, \mathrm{P}, c, \mathbf{x}^t) \leqslant \mathrm{E}_{\mathbf{x}^t} \operatorname{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t).$$

So it suffices to determine how quickly uncertainty intervals shrink in a uniform chain. Here it turns out that we can determine an exact rate of decrease for this simple space.

LEMMA 2. *For any initial segment concept* $\mathbf{i} = [0, i] \in I$,

$$\mathrm{E}_{\mathbf{x}^t} \operatorname{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) = \frac{2 - (1 - i)^{t+1} - i^{t+1}}{t + 1}.$$

This rate of decrease is explicitly rational in t, so combining the fact that $\operatorname{err}(L, \mathrm{P}, c, \mathbf{x}^t) \leqslant \operatorname{wid}(C, \mathrm{P}, c, \mathbf{x}^t)$ with Lemmas 1 and 2 we have shown that any consistent learner L for C must obtain $\mathrm{E}_{\mathbf{x}^t} \operatorname{err}(L, \mathrm{P}, c, \mathbf{x}^t) = O(t^{-1})$ for any $c \in C$. This ends the proof of Proposition 3. ∎

Thus, the worst case learning curve for any concept chain can never be worse than rational. It remains to show that rational convergence is the best that any learner can achieve in this case.

THEOREM 4 (Continuous LB). *For any continuous concept chain* C: *there is a domain distribution* P *that forces any learner* L *to obtain a rational learning curve* (*i.e.*, $\mathrm{E}_{\mathbf{x}^t} \operatorname{err}(L, \mathrm{P}, c', \mathbf{x}^t) = \Omega'(t^{-1})$) *for some target concept* $c' \in C$.

Proof. Since C is continuous it can be indexed $C = \{c_y : y \in [0, 1]\}$ such that $c_y \subset c_z$ for $y < z$. Given this indexing, we can fix a domain distribution P such that $d_{\mathrm{P}}(c_y, c_z) = |y - z|$. (One can always construct this distribution by the same procedure used to construct the Lebesgue measure on $[0, 1]$; see e.g., [2, Chapter 1]). Note that the resulting concept space (C, P) is isomorphic to the uniform chain (I, U) in Definition 4. Therefore, it suffices to establish the lower bound for (I, U).

For this space, we already know by Lemma 2 that the width of any uncertainty interval only decreases rationally to zero (i.e., $\mathrm{E}_{\mathbf{x}^t} \operatorname{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) = \Omega(t^{-1})$ for any $\mathbf{i} = [0, i] \in I$), so it would be surprising if a learner could do significantly better than this for every $\mathbf{i} \in I$. To prove that any learner will be forced to exhibit rational convergence for some fixed target $\mathbf{i}' \in I$, we employ the same averaging argument used in Proposition 2. In particular, we fix a prior distribution Q on the collection of initial segment concepts I and argue that any learner L must obtain a large expected error *on average* over targets in I.

Let Q be the *uniform* prior on I. For this prior it turns out that the simple "midpoint" guessing strategy, MP, is Bayes optimal.

Strategy MP. For a training sequence $\mathbf{i}\mathbf{x}^t$, which yields the uncertainty interval $[\mathbf{s}(\mathbf{i}\mathbf{x}^t), \mathbf{l}(\mathbf{i}\mathbf{x}^t)]$, guess the midpoint concept $\mathbf{m} = [0, m]$ defined by the endpoint $m = (s + \ell)/2$.

It is easy to show that no learner L can obtain a smaller average expected error than MP, when target concepts $\mathbf{i}$ are chosen randomly according to Q.

LEMMA 3. *For any learner* L,

$$\mathrm{E}_{\mathbf{i}} E_{\mathbf{x}^t} \operatorname{err}(L, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) \geqslant \mathrm{E}_{\mathbf{i}} E_{\mathbf{x}^t} \operatorname{err}(\mathsf{MP}, \mathrm{U}, \mathbf{i}, \mathbf{x}^t).$$

Therefore, it suffices to establish a rational lower bound on MP's average expected error. Here we see that, not surprisingly, MP achieves an average expected error that is a fixed fraction of the expected width of the uncertainty interval.

LEMMA 4. *For the learning strategy* MP,

$$\mathrm{E}_{\mathbf{i}} \mathrm{E}_{\mathbf{x}^t} \operatorname{err}(\mathsf{MP}, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) = \tfrac{1}{4} \mathrm{E}_{\mathbf{i}} \mathrm{E}_{\mathbf{x}^t} \operatorname{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t).$$

Therefore, combining the results of Lemmas 3, 4, and finally 2, we see that any learner L must obtain

$$\begin{aligned} \mathrm{E}_{\mathbf{i}} \mathrm{E}_{\mathbf{x}^t} \operatorname{err}(L, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) &\geqslant \frac{1}{4} \int_0^1 \frac{2 - (1 - i)^{t+i} - i^{t+1}}{t + 1} \, di \\ &= \frac{1}{2(t + 2)}. \end{aligned} \tag{2}$$

Clearly, since this lower bound holds on average over all concepts in I, it must hold for some $\mathbf{i}_t \in I$ at each training sample size t. However, we need to establish the stronger claim that there is a single $\mathbf{i}'$ in I that forces $\mathrm{E}_{\mathbf{x}^t} \operatorname{err}(L, \mathrm{U}, \mathbf{i}', \mathbf{x}^t) = \Omega(t^{-1})$ for infinitely many training sample sizes t. To this end, we show that any learner will be forced to obtain an expected error above the bound infinitely often for a non-trivial portion of the concepts in I, as measured by to the prior distribution Q.

LEMMA 5. *For any* $\lambda > 0$,

$$\mathrm{Q}\left\{\mathbf{i} \in I : \mathrm{E}_{\mathbf{x}^t} \operatorname{err}(L, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) \geqslant \frac{1 - \lambda}{2(t + 2)} \; i.o.\; t\right\} > 0.$$

Applying this lemma gives the result. ∎

(Notice that this is a stronger result than the lower bound of Haussler *et al.* [14]: we have shown that there is a *single* domain distribution P that forces any learner to exhibit rational convergence for a non-trivial portion of the concepts in C. In work subsequent to [14], Haussler *et al.* [15, Theorem 3.2] have independently established a similar result to Eq. (2) above, however their argument is quite different and they do not supply the final step (Lemma 5).

The proof presented here generalizes more readily to Theorem 5 below.)

Combining Proposition 3 and Theorem 4 shows that the worst case learning curve for a continuous concept chain must be exactly rational.

COROLLARY 2. *Any continuous concept chain C has a rational worst case learning curve*: $\mathrm{LC}(C) = \Theta'(t^{-1})$.

Together, Corollaries 1 and 2 establish the existence of the fundamental dichotomy between rational and exponential worst case learning curves in the distribution-free model.

4. EXACT BOUNDARY

The previous section revealed the basic dichotomy between rational and exponential learning curves by demonstrating how continuous concept chains can have rational worst case learning curves, while finite concept classes always have exponential such curves. Of course, this characterization is far from complete as there is a significant gap between finite and continuous concept classes. For example, none of the preceding results apply to countably infinite concept classes. This gap leaves open the question of identifying the precise conditions that dictate between rational and exponential learning curves, and determining whether other (intermediate) forms of convergence are possible under a uniform worst case analysis (e.g., $\Theta'(t^{-n})$ or some other form).

In this section we derive the exact boundary between rational and exponential learning curves for the special case of simple concept chains. Specifically, this boundary is determined by the presence or absence of a *dense* subchain in the original class C.

DEFINITION 5 (Dense versus Scattered Concept Chains). A chain C is dense if between any two concepts $c_1 \subset c_2$ in C there is a third $c_3 \in C$ such that $c_1 \subset c_3 \subset c_2$. (We require that a dense chain contain at least two, and hence, infinitely many concepts.) Thus we say a chain C is *somewhere*-dense if it contains a dense subchain (not necessarily a subinterval), and *nowhere*-dense if it contains no such subchain. Nowhere-dense chains are also referred to as *scattered* [21].

The main contribution here is to show that any somewhere-dense concept chain has a rational worst case learning curve, whereas nowhere-dense chains have exponential such curves. This gives an exact and complete characterization of worst case convergence forms (for concept chains) under the distribution-free model.

4.1. *Dense Concept Chains*

From Corollary 2 above we know that continuous concept chains have rational worst case learning curves. Since any continuous chain is obviously dense, it is natural to consider whether this is the key property that forces rational worst case convergence. Here we show that density is indeed sufficient to force any learner to exhibit rational learning curves for some fixed target concepts and domain distributions.

THEOREM 5 (Dense LB). *For any dense concept chain C: there is a domain distribution* P *that forces any learner L to obtain a rational learning curve* (*i.e.*, $\mathrm{E}_{\mathbf{x}^t}\mathrm{err}(L, \mathrm{P}, c', \mathbf{x}^t) = \Omega^*(t^{-1})$) *for some target concept $c' \in C$.*

Proof. We establish the slightly weakened proposition that $\mathrm{E}_{\mathbf{x}^t}\mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) = \Omega'(t^{-1-\varepsilon})$ for any $\varepsilon > 0$ (hence the notation Ω^*). The basic idea is to generalize the proof of Theorem 4 to handle arbitrary dense chains. However, we must face the fact that C need not be continuous in general (e.g., C might only be countably infinite) so we cannot directly reduce the problem to a uniform chain as before. Instead, we have to define a domain distribution P that *simulate*s the structure of a uniform chain as closely as possible. To this end, we explicitly construct a dense subchain of C on a countable subdomain of X and then define appropriate distributions P and Q.

Construction. First construct a dense chain C_0 on a countable domain X_0 by selecting concepts from C and domain objects from X in a series of Stages $k = 0, 1, 2, \ldots$ as follows: At Stage 0, select any two concepts $c_0 \subset c_1$ from C and choose a domain object x_1 between them (i.e., choose $x_1 \in c_1 - c_0$ such that there remains $c_2, c_3 \in C$ with $c_0 \subset c_2 \subset c_3 \subset c_1$ and $x_1 \in C_3 - c_2$; see Fig. 2). Next, at Stage 1, choose a domain object x_2 and concept c_2 between c_0 and x_1, and then a concept c_3 and domain object x_3 between x_1 and c_1; maintaining the alternation between domain objects and target concepts shown in Fig. 2. Then for all subsequent Stages $k \geqslant 2$, choose a domain object and target concept in each gap left from previous stages, maintaining the alternation between target concepts and domain objects after each stage, again, as shown in Fig. 2. Note that the density of C permits us to continue this process indefinitely, and therefore we obtain a dense subchain C_0 defined on a countable subdomain X_0. This construction provides us with a canonical structure on which to define our probability distributions. (We now drop the subscript 0 for the remainder of this proof, with the understanding that C and X now refer to the constructed C_0 and X_0.)

Now to define a domain distribution P on X: Note that P cannot be uniform since X is only countably infinite. However, we can approximate a uniform distribution by assigning probabilities as follows. First, at Stage 0, assign $\mathrm{P}\{x_0\} = 0$ to the only domain object added at Stage 0. Then for Stages $k = 1, 2, \ldots$, assign a probability of $p_k = (3^{\varepsilon} - 1)/(2 \cdot 3^{k(1+\varepsilon)-1})$ to each of the domain objects x_i added at Stage k. This gives a well-defined probability distribution

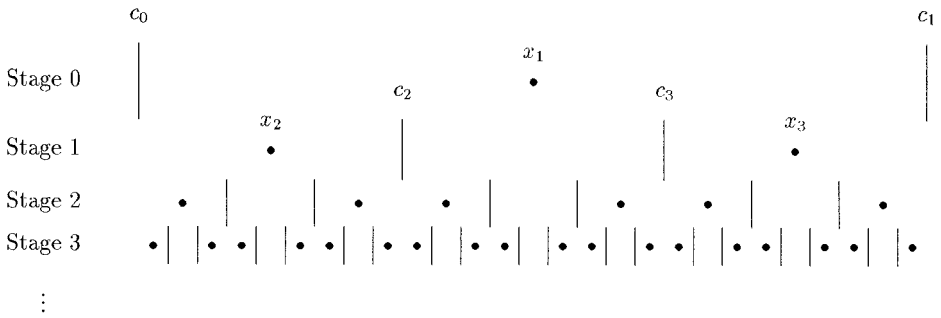

FIG. 2. Constructing a dense subchain on a countable subdomain: Each line indicates a concept that contains the domain objects (indicated by bullets) to the left of the line. Repeating this construction for stages $k = 0, 1, 2, \ldots$ results in a dense chain $C = \{c_0, c_1, \ldots\}$ being defined on a countable domain $X = \{x_1, x_2, \ldots\}$.

for any $\varepsilon > 0$, since there are a total of $N_k = 2 \cdot 3^{k-1}$ objects added at each Stage k, and summing over stages 1, 2, ... yields a total probability of $(3^{\varepsilon} - 1) \sum_{k=1}^{\infty} 3^{-\varepsilon k} = 1$. Notice that we can use the parameter ε to control the "uniformity" of P. That is, choosing smaller values of ε forces the probabilities p_k assigned at each stage to converge more slowly to 0, and hence make the distribution more uniform.

Finally, to define the prior distribution Q on C we proceed in exactly the same way as P on X above: In particular, at Stage 0 assign $\mathrm{Q}\{c_0\} = \mathrm{Q}\{c_1\} = 0$, and then at each subsequent Stage $k \geqslant 1$ assign a probability $q_k = p_k$ to each of the concepts c_j added at Stage k. This yields a well defined probability distribution Q on C exactly as above since during each Stage $k > 0$ an equal number of domain objects and target concepts are added to the construction.

Given this explicit construction of C, X, P, and Q, we can now repeat the lower bound argument from Theorem 4. First we must verify that P is indeed a "hard" domain distribution in the sense that it forces uncertainty intervals to shrink rationally as a function of t.

LEMMA 6. *For any $c \in C$, there is a constant $\alpha_c > 0$ such that*

$$\mathrm{E}_{\mathbf{x}^t} \operatorname{wid}(C, \mathrm{P}, c, \mathbf{x}^t) > \frac{\alpha_c}{t+1}.$$

Given that this slow convergence holds for all concepts in C it would be surprising if a learner could achieve significantly faster convergence for every possible target in C. To prove this, we follow the same averaging argument used in Proposition 2 and Theorem 4 above: Given the prior distribution Q over C, we argue that any learner L must achieve a large expected error *on average* over the target concepts in C. To do this, we consider a simple learning strategy MC that achieves near-optimal average expected error for the distributions P and Q defined above.

Strategy MC. For a training sequence $c\mathbf{x}^t$, guess the consistent concept $c^* \in [s(c\mathbf{x}^t), \ell(c\mathbf{x}^t)]$ that has maximum prior probability according to Q.

It can be shown that no learner L can obtain a smaller average expected error than MC, up to a small multiplicative constant, when target concepts $\mathbf{i}$ are chosen randomly according to Q.

LEMMA 7. *There is a constant $\beta > 0$ such that for any learner L,*

$$\mathrm{E}_c \mathrm{E}_{\mathbf{x}^t} \operatorname{err}(L, \mathrm{P}, c, \mathbf{x}^t) \geqslant \beta \, \mathrm{E}_c \mathrm{E}_{\mathbf{x}^t} \operatorname{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t).$$

Therefore it suffices to establish a rational lower bound on MC's average expected error. Not surprisingly, MC must obtain an average error that is at least some fixed fraction of the width of the uncertainty interval (discounting the fact that MC will guess the target concept exactly with a small non-zero probability).

LEMMA 8. *For a constant $\delta > 0$,*

$$\mathrm{E}_c \mathrm{E}_{\mathbf{x}^t} \operatorname{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t) \geqslant \frac{\delta}{(t+2)^{2\varepsilon}} \mathrm{E}_c \mathrm{E}_{\mathbf{x}^t} \operatorname{wid}(C, \mathrm{P}, c, \mathbf{x}^t).$$

Finally, combining Lemmas 7, 8, and 6, we see that any learner L must obtain

$$\mathrm{E}_c \mathrm{E}_{\mathbf{x}^t} \operatorname{err}(L, \mathrm{P}, c, \mathbf{x}^t) \geqslant \frac{\gamma}{(t+2)^{1+2\varepsilon}}, \tag{3}$$

for a constant $\gamma = \bar{\alpha}\beta\delta > 0$ (where $\bar{\alpha} > 0$ is the average value of α_c over $c \in C$). Clearly, since this bound holds on average over all concepts in C it must hold for some $c_t \in C$ for each training sample size t. However as in Theorem 4, we need to establish the stronger claim that there is a single c' in C that forces $\mathrm{E}_{\mathbf{x}^t} \operatorname{err}(L, \mathrm{P}, c', \mathbf{x}^t) = \Omega(t^{-1-\varepsilon})$ for infinitely many training sample sizes t. To this end, we show that any learner must exhibit (near) rational convergence for a non-trivial portion of the concepts in C, as measured by the prior distribution Q.

LEMMA 9. *For any* $\lambda > 0$,

$$\mathrm{Q}\left\{c \in C\colon \mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) \geqslant \frac{(1-\lambda)\,\gamma}{(t+2)^{1+2\varepsilon}}\ i.o.\ t\right\} > 0.$$

Finally, notice that we can freely choose ε to be any positive quantity arbitrarily close to zero, establishing the theorem. ▌

Combining Theorem 5 and Proposition 3 shows that the worst case learning curve for a somewhere-dense concept chain must be rational.

COROLLARY 3. *Any somewhere-dense concept chain* C *has a rational worst case learning curve*: $\mathrm{LC}(C) = \Theta^*(t^{-1})$.

4.2. *Scattered Concept Chains*

We now turn our attention to the complementary class of nowhere-dense (scattered) concept chains. Here we wish to show that exponential learning curves can always be obtained for any fixed target concept and domain distribution. From Proposition 2 above, we know that it is impossible to achieve better than exponential worst case convergence for any non-trivial chain C, so it remains only to show that exponential convergence can always be achieved in this case. This turns out to be hard: we must *demonstrate* a learning strategy that achieves exponential convergence for any scattered concept chain, or least prove that such a strategy exists. Below, we develop a special learning strategy, CHOLC, that achieves this for arbitrary scattered chains.

THEOREM 6 (Scattered UB). *For any scattered concept chain* C: *there is a special learning strategy* CHOLC *that obtains an exponential learning curve* (*i.e.*, $\mathrm{E}_{\mathbf{x}^t}\,\mathrm{err}(\mathsf{CHOLC}, \mathrm{P}, c, \mathbf{x}^t) = e^{O(-t)}$) *for every target concept* $c \in C$, *regardless of the domain distribution* P.

Proof. Fix an arbitrary domain distribution P, and consider the chain (C, P) that results from collapsing equivalent concepts under P together. Clearly, this chain is still scattered since a nowhere-dense chain cannot be made somewhere-dense by removing concepts, so it suffices to consider a scattered chain where no two concepts are equivalent under the d_{P} metric. We saw in the proof of Proposition 1 that exponential convergence results whenever the learner can identify the target concept (up to P-equivalence) with non-zero probability after a finite number of training examples. Here, we attempt to apply this idea to arbitrary scattered concept chains.

First, for an isolated target concept c (i.e., a concept that has a least larger neighbor $\ell \supset c$ and a greatest smaller neighbor $s \subset c$ in (C, P)) it is clear that c is identified with non-zero probability after just two training examples (eliminating both s and ℓ). Thus, any learner L that guesses consistent hypotheses from C will achieve exponential convergence to an isolated target. The only difficulty then must be in dealing with *limit* concepts; i.e., concepts that are the limits of infinite ascending $(\bigcup_1^\infty c_i)$ or descending $(\bigcap_1^\infty c_i)$ sequences of concepts in C. (Note that this can easily happen without the chain being dense; see Fig. 3a.) The problem with limit concepts is that they permit a consistent learner L to guess an infinite sequence of hypotheses that converges to, but never reaches the target concept. For example, in Fig. 3a, if the rightmost concept is the target, then a learner which guesses the smallest consistent concept will never reach this concept, and as we saw in the proof of Theorem 5 (Lemma 6) this can lead to rational convergence for certain distributions. Therefore, the trick to achieving exponential convergence must be to avoid guessing

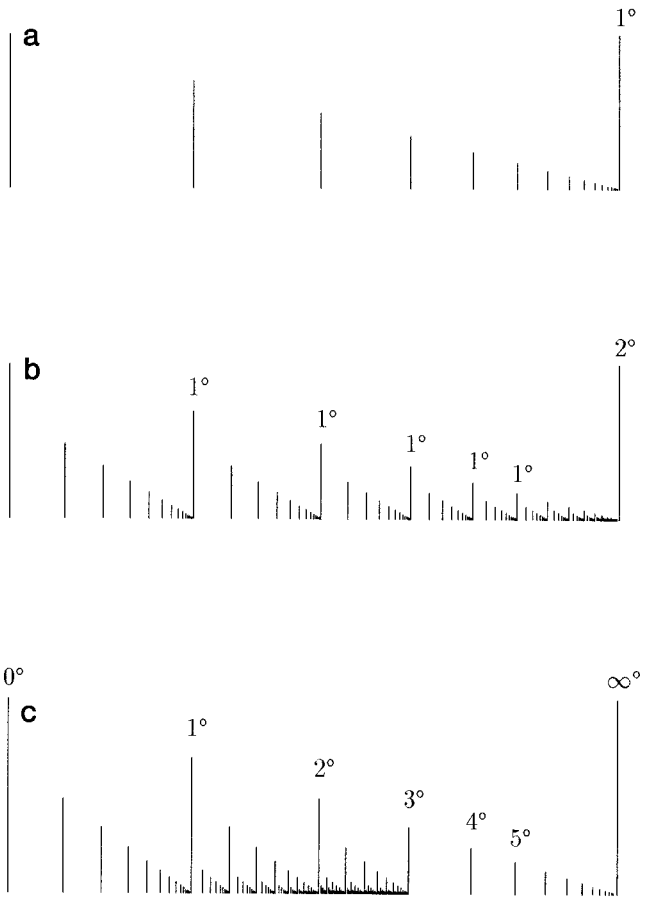

FIG. 3. This figure depicts three scattered concept chains defined on $X = [0, 1]$. Here, each line y indicates the concept c_y that contains all points $x \leqslant y$. Note that each of these chains is nowhere-dense since every concept (except the last) has a *least* larger concept. (a) A scattered concept chain with a first order limit concept at the end (i.e., a chain of order type $\omega + 1$). (b) A scattered concept chain with a series of first order limits and a finally a second order limit concept at the end (i.e., a chain of order type $\omega^2 + 1$). (c) A scattered concept chain with a series of limit concepts of progressively higher order, followed finally by a limit concept with *infinite* order (i.e., a chain of order type $\omega^\omega + 1$).

infinite sequences of hypotheses that slowly converge to limit concepts without ever reaching them. (Notice that the special learning strategy 1IGPS developed by Haussler *et al.* [14, 15] does not automatically do this, and hence might produce a rational learning curve when in fact exponential convergence is possible.)

The obvious way to avoid this difficulty is to guess limit concepts *before* isolated concepts. Then provided that the limit concepts are themselves isolated from one another, we always expect to achieve exponential convergence since any target will be guessed with non-zero probability after just two training examples. Of course, it also possible to have *limits* of limit concepts in a scattered chain (i.e., second order limits; cf. Fig. 3b). In fact, limit concepts of each order 1, 2, 3, ... are certainly possible; and in general one can even have limits of these (i.e., concepts of infinite order!); see Fig. 3c. The following lemma shows that it is possible to have target concepts of any *ordinal* order in a scattered chain, but nevertheless, all concepts of a given order are *isolated* from concepts of the same or higher order. (This fact follows from Hausdorff's Theorem [21], which provides a suitably constructive characterization of the class of scattered linear orderings.)

LEMMA 10 (Corollary to Hausdorff's Theorem). *For a scattered concept chain C, there is some least ordinal γ such that* (i) *every concept in C has order $\beta \leqslant \gamma$, and* (ii) *all limit concepts of a particular order β are* isolated *in concepts of the same or higher order.*

This is the key property of scattered concept chains that permits us to develop a learning strategy that always achieves exponential convergence. In fact, the following strategy almost works in general.

Strategy HOLC. Guess the highest order limit concept consistent with all the training examples.

Intuitively, we expect this strategy to always achieve exponential convergence since any target concept is isolated in the class of concepts with the same or higher order, and this means HOLC will guess a zero-error hypothesis with non-zero probability after two training examples (eliminating the greatest smaller neighbor and least larger neighbor). However, there is one final problem: HOLC requires that a consistent concept of maximal order always *exists* for any sequence of training examples. Unfortunately, this need not be the case in general and therefore HOLC is not always well defined. (For example, consider removing the last concept in Fig. 3c.)

This difficulty can be circumvented by first *compactifying* the chain in a natural way: Let $\mathscr{C}(C)$ denote the closure of C under $\cap$ and $\cup$. We will call such a chain *compact* for the following natural reasons.

LEMMA 11. *Any chain C that is closed under $\cup$, $\cap$ is also complete and bounded, and satisfies a natural version of the Bolzano–Weierstrass property.*

A key property of this closure $\mathscr{C}(C)$ is that it cannot make a scattered chain dense.

LEMMA 12. *If C is a scattered chain, then the chain $\mathscr{C}(C)$ formed by closing C under $\cup$, $\cap$ is still scattered.*

Moreover, because of compactness, $\mathscr{C}(C)$ provides a maximum order limit concept in any uncertainty interval.

LEMMA 13. *For a scattered chain C that is closed under $\cap$, $\cup$, there is always a concept $c \in C$ of* maximal *order consistent with any finite sequence of training examples.*

This leads to the final proposal for a learning strategy that is guaranteed to achieve exponential convergence for any scattered concept chain:

Strategy CHOLC. First close the chain C under $\cup$, $\cap$ to obtain $\mathscr{C}(C)$, then guess the highest order limit concept in $\mathscr{C}(C)$ consistent with the training examples.

By Lemmas 12 and 13, CHOLC is well defined for any scattered concept chain C, and by Lemma 10, CHOLC is guaranteed to achieve exponential convergence for any target concept c in C. (Note that fixing a domain distribution preserves the order structure of the chain, or collapses subintervals of the chain together. Collapsing subintervals cannot produce new limit concepts or increase the order of existing target concepts; beyond identifying them with already existing such concepts. Therefore, the result holds for any domain distribution P.) ▮

Obviously Procedure CHOLC has little practical impact, but the issues it addresses shed light on the fundamental nature of worst case learning curves: Combining Theorem 6 with Proposition 2 shows that any non-trivial, nowhere-dense concept chain must have exactly an exponential worst case learning curve.

COROLLARY 4. *Any non-trivial, scattered concept chain C has an exponential worst case learning curve*: $\mathrm{LC}(C) = e^{\Theta'(-t)}$.

Together, Corollaries 3 and 4 establish a complete and exact boundary between rational and exponential learning curves (for concept chains) in the worst case, distribution-free model.

COROLLARY 5 (Exact Boundary). *Any concept chain C must have either a* rational *or* exponential *worst case learning curve*: *rational if and only if the chain is somewhere-dense*; *exponential if and only if the chain is nowhere-dense.*

5. SCALING EFFECTS

Although the previous results draw a precise boundary between rational and exponential worst case learning curves, in one sense they miss the point demonstrated by the experimental results of Cohn and Tesauro [6, 7]: Since their computer simulations were conducted with finite precision, every concept class Cohn and Tesauro considered must have been fundamentally *finite*, and hence every learning curve they obtained must have been asymptotically exponential. The fact that they obtain rational learning curves seems to contradict the theoretical results presented here. However, the real source of the dichotomy in these experiments is a scaling effect: Every learning curve observed by Cohn and Tesauro really is asymptotically exponential, it is just that at the scale of training sample sizes they consider (relative to the size of the inter-concept distances) convergence *appears* rational. This is easily demonstrated by a simple example. Consider a finite concept chain C_n consisting of $n+1$ concepts $c_0 \subset c_1 \subset \cdots \subset c_n$, and fix a domain distribution P_n that imposes a distance of $1/n$ between adjacent concepts.

PROPOSITION 7 [22, Proposition 4.24]. *For any* $c \in (C_n, \mathrm{P}_n)$,

$$\left(1-\frac{1}{n}\right)^t \frac{1}{4(t+1)} < \mathrm{E}_{\mathbf{x}^t}\,\mathrm{wid}(C_n, \mathrm{P}_n, c, \mathbf{x}^t) < \left(1-\frac{2}{n}\right)^t \frac{2}{t+1}.$$

Notice that this convergence is exponential in t and well approximated by $e^{-t/n}a(t+1)^{-1}$ for large n. Now if we focus on training sample sizes t that are small relative to n, the $e^{-t/n}$ factor will behave like a constant near 1 and the $(t+1)^{-1}$ factor will dominate. In this case we would observe apparently rational convergence, even though convergence remains asymptotically exponential. To reveal exponential convergence, we must consider training sample sizes on the order of $t = n, 2n, 3n, \ldots$, which in the "continuous" case considered by [6, 7] is on the order of "computer BIGNUM." This partly explains the dichotomy observed by Cohn and Tesauro, as they considered the *same* training sample sizes for concept classes with vastly different interconcept distances: observing exponential convergence when the gaps were large, and rational convergence when the gaps were small. (A similar observation has been made by Barnard [3].)

This suggests that there is no definitive scale where training sample sizes are inherently "interesting" for a given gap size in practical settings. That is, contrary to the common suggestion in statistical mechanical analyses [13], one cannot simply look at a concept space and ascertain what sample sizes are inherently interesting to consider: sometimes the practically available sample sizes are small relative to the inverse gap size, and sometimes they are large. *This* is what yields the practical dichotomy between rational and exponential convergence.

6. DISCUSSION AND RESEARCH DIRECTIONS

This paper establishes the fundamental dichotomy between rational and exponential learning curves under a worst case, distribution-free analysis; contrary to the common suggestion that strong distributional assumptions and average case analyses are needed to reveal this distinction [13, 24]. Our results reveal that the precise characteristic that dictates between rational and exponential convergence is the existence of a dense subchain in the concept class.

However, these results are still limited in a practical sense. The main limitation is the restriction to simple concept chains, and the theory needs to be scaled-up to handle arbitrary concept classes. Note that scaling-up to product chains is trivial, but characterizing the boundary between rational and exponential convergence for general concept classes is a difficult open problem. One source of difficulty is the fact that a concept class might not contain a dense chain directly and yet still form a dense chain over a restricted subset of the domain. (For example, consider the concept class $C = \{[0, y] \cup \{1+y\} : y \in [0, 1]\}$ on $X = [0, 2]$. This class directly contains no chain longer than one, and yet defines a continuous chain on the subdomain $[0, 1]$.) Proving that such a class forces rational worst case convergence is easy. However, proving that exponential convergence can always be achieved in the contrary case is hard: one would have to generalize CHOLC to somehow cope with arbitrary nowhere-dense concept classes (with finite VC dimension).

Beyond characterizing the conditions when rational versus exponential convergence can be achieved, it would also be useful to have tight bounds on the specific convergence *rates*. From the discussion in Section 3.1 we saw that, for exponential learning curves at least, it is impossible to obtain a tight characterization of convergence rates without taking into account the specific domain distribution. This means that any reasonable characterization of learning curve convergence rates must adopt a distribution-specific analysis rather than the distribution-free model adopted in this paper. Unfortunately, it turns out that a distribution-specific analysis of worst case convergence rates is much more difficult than a distribution-free analysis. The most comprehensive theory to date is due to Haussler *et al.* [13], which characterizes the rates with which the diameters of consistent neighborhoods converge to target concepts. However this approach cannot be fully general, since there are concept spaces which have consistent neighborhoods that do not converge and yet special learning strategies can still achieve asymptotic convergence to zero error in these cases [22]. A step towards a general distribution-specific theory of worst case learning curves is taken in [22] where a

generic learning procedure BC is developed that achieves asymptotic convergence to zero error for any concept space (C, P) that is finitely α-coverable at all scales $\alpha > 0$. Moreover, it can be shown that *no* learning procedure can converge to all targets if this is not the case. (An α-cover of a concept space (C, P) is a set of concepts A such that for every $c \in C$ there is a $c' \in A$ such that $d_{\mathrm{P}}(c, c') < \alpha$.)

THEOREM 8 [22, Theorem 4.29]. *For any concept space (C, P) where the size of the minimal α-covers of (C, P) behaves as $N_\alpha(C, \mathrm{P}) = \Theta(1/\alpha)^d$ for some d as $\alpha \to 0$:*

1. *The learning strategy* BC *achieves* $\mathrm{E}_{\mathbf{x}^t} \mathrm{err}(\mathsf{BC}, \mathrm{P}, c, \mathbf{x}^t) = O((d/t)\ln(t/d))$ *for any target $c \in C$.*

2. *Any learner L must obtain* $\mathrm{E}_{\mathbf{x}^t} \mathrm{err}(L, \mathrm{P}, c', \mathbf{x}^t) = e^{\Omega'(-t/d)}$ *for some target $c' \in C$.*

So BC is a universal learning strategy in the sense that it achieves worst case convergence to zero error whenever this is possible under the distribution-specific model. Unfortunately, these bounds on the *rate* of worst case convergence admit both rational and exponential forms.

Surprisingly, just determining the boundary between rational and exponential convergence is *harder* in the distribution-specific case than the distribution-free model. In fact, it is not even obvious what property distinguishes rational from exponential convergence for simple chains in this case. It has often been suggested [3, 23, 24] that density in the metric d_{P} should distinguish rational from exponential worst case learning curves. However, this suggestion can easily be shown to be false: Consider a concept chain (C, P) where $C = \{\mathbf{q} = [0, q] : q \in \mathbb{Q}[0, 1]\}$ consists of rational initial segments $[0, q]$ of the unit interval $X = [0, 1]$, and P is any distribution on $[0, 1]$ such that $\mathrm{P}(q) > 0$ for all (and only) rational points $q \in \mathbb{Q}[0, 1]$. Clearly, C is dense under d_{P} (since for any $\mathbf{q} \in C$ and $\varepsilon > 0$ there is an $\mathbf{r} \in C$ such that $d_{\mathrm{P}}(\mathbf{q}, \mathbf{r}) < \varepsilon$) and yet a simple learning procedure always achieves exponential convergence for this space: simply guess the smallest rational initial segment $\mathbf{r}$ consistent with the training sequence. Since each $\mathbf{q} \in C$ has a gap between it and all smaller concepts (because $q \in \mathbf{q}$ but $q \notin \mathbf{r}$ for all $\mathbf{r} \subset \mathbf{q}$) we will guess $\mathbf{q}$ with non-zero probability for each training example, and therefore achieve exponential convergence. (Notice that this result does not contradict Theorem 5 as we have only shown that exponential convergence can be achieved for some, but not every domain distribution.)

This example shows that density in the induced metric d_{P} is not sufficient to force rational convergence in general. Instead we need some property of "density from all sides." Unfortunately, generalizing this notion to arbitrary concept spaces appears difficult, and the prospects for a general theory seem remote. If we cannot even distinguish between rational and exponential convergence it seems unlikely that we can ever develop a tight characterization of empirical convergence rates for general spaces. Therefore, this difficulty has serious implications for any theory (distribution-specific or otherwise) that purports to provide a general, tight characterization of the learning curves observed in practice.

APPENDIX A: PROOFS OF LEMMAS

Proof of Lemma 1. (There is a slight ambiguity here as the sequences $\mathbf{x}^t$ actually range over different domains X^t in each case, but this has no bearing on the result.) First, consider the space (C, P) and choose an arbitrary target $c \in (C, \mathrm{P})$. For this c, define the random variables $S^c(x) = \mathrm{P}(c - s(cx))$ and $H^c(x) = \mathrm{P}(\ell(cx) - c)$, which measure the distance between c and the smallest and largest concepts in the uncertainty interval $[s(cx), \ell(cx)]$ respectively. Then the variables $\underline{S^c}_t(\mathbf{x}^t) = \min\{S^c(x_1), ..., S^c(x_t)\}$ and $\underline{H^c}_t(\mathbf{x}^t) = \min\{H^c(x_1), ..., H^c(x_t)\}$ measure the distance between c and the smallest and largest concepts in $[s(c\mathbf{x}^t), \ell(c\mathbf{x}^t)]$. Thus,

$$\begin{aligned}\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t) &= \mathrm{P}(\ell(c\mathbf{x}^t) - s(c\mathbf{x}^t)) \\ &= \mathrm{P}(\ell(c\mathbf{x}^t) - c) + \mathrm{P}(c - s(c\mathbf{x}^t)) \\ &= \underline{S^c}_t(\mathbf{x}^t) + \underline{H^c}_t(\mathbf{x}^t).\end{aligned}$$

Now, turning our attention to the space (I, U), consider a corresponding target concept $\mathbf{i} = [0, \mathrm{P}(x)] \in I$. For this concept we can define $\underline{S^{\mathbf{i}}}_t$ and $\underline{H^{\mathbf{i}}}_t$ as above. Notice that, by construction, this concept $\mathbf{i}$ has the property that the width of its uncertainty interval can shrink no faster than c's:

P1. *For any $y \leqslant \mathrm{P}(c)$ we have $\mathrm{P}\{S^c < y\} \geqslant \mathrm{U}\{S^{\mathbf{i}} < y\}$, and for any $y \leqslant 1 - \mathrm{P}(c)$ we have $\mathrm{P}\{H^c < y\} \geqslant \mathrm{U}\{H^{\mathbf{i}} < y\}$.*

Given P1 the result is obvious, since for positive random variables X and Y, $F_X \geqslant F_Y$ implies $\mathrm{E}X \leqslant \mathrm{E}Y$. To prove P1, note that $\mathrm{U}\{S^{\mathbf{i}} < y\} = y$ for $y \leqslant \mathrm{P}(c)$ by construction, whereas $\mathrm{P}\{S^c < y\} \geqslant y$ over this range; similarly $\mathrm{U}\{H^{\mathbf{i}} < y\} = y$ for $y \leqslant 1 - \mathrm{P}(c)$, and yet $\mathrm{P}\{H^c < y\} \geqslant y$ over this range. (To see this for H^c, let $\ell_y = \bigcap\{h \in C : \mathrm{P}(h - c) \geqslant y\}$ and notice that $\mathrm{P}\{H^c < y\} = \mathrm{P}(\ell_y - c) \geqslant y$. A similar argument also works for S^c.) ∎

Proof of Lemma 2. Consider an arbitrary target concept $\mathbf{i} = [0, i] \in I$. Let $S(x)$ and $H(x)$ be random variables for $\mathbf{i}$ as defined in the proof of Lemma 1 above (dropping the superscript $\mathbf{i}$). Then by definition we have $\mathrm{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) = \underline{S}_t(\mathbf{x}^t) + \underline{H}_t(\mathbf{x}^t)$, and hence

$$\begin{aligned}&\mathrm{E}_{\mathbf{x}^t} \mathrm{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) \\ &\quad = \mathrm{E}_{\mathbf{x}^t}[\underline{S}_t(\mathbf{x}^t) + \underline{H}_t(\mathbf{x}^t)] \\ &\quad = \sum_{k=0}^{t} \binom{t}{k} \mathrm{U}(\mathbf{i})^k \, \mathrm{U}(\mathbf{i}^c)^{t-k} \, (\mathrm{E}_{\mathbf{x}^k}[\underline{S}_k \,|\, x_1, ..., x_k \in \mathbf{i}] \\ &\qquad + \mathrm{E}_{\mathbf{x}^{t-k}}[\underline{H}_{t-k} \,|\, x_{k+1}, ..., x_t \notin \mathbf{i}]).\end{aligned}$$

Obviously in this case $\mathrm{U}(\mathbf{i}) = i$ and $\mathrm{U}(\mathbf{i}^c) = 1 - i$. Now notice that for $S(x)$ and $H(x)$ defined as above, we have $S \mid (x \in \mathbf{i}) \sim \text{uniform}[0, i)$ and $H \mid (x \in \mathbf{i}^c) \sim \text{uniform}(0, 1 - i)$. For any random variable $R(x)$ with a uniform$(0, r)$ distribution, it is not hard to show that $\mathrm{E}_{\mathbf{x}^t}[\underline{R}_t(\mathbf{x}^t)] = r/(t+1)$ [17, pp. 99]. Therefore, we get

$$\mathrm{E}_{\mathbf{x}^t} \operatorname{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) = \sum_{k=0}^{t} \binom{t}{k} i^k (1-i)^{t-k} \left(\frac{i}{k+1} + \frac{1-i}{t-k+1} \right).$$

Finally, each of the two terms in the summation can be reduced via the Binomial Theorem [5, Chapter 4] to yield the stated result. ∎

Proof of Lemma 3. We will show that the hypotheses produced by MP are Bayes-optimal for the uniform prior Q on I and the uniform domain distribution U on $[0, 1]$. The result then follows by a well known fact about Bayes-optimal prediction; cf. [8, Chapter 2].

Intuitively, this result is clear. Given a training sequence $\mathbf{z}^t = \langle\langle x_1, y_1\rangle, ..., \langle x_t, y_t\rangle\rangle$ yielding the uncertainty interval $[\mathbf{s}(\mathbf{z}^t), \mathbf{l}(\mathbf{z}^t))$, the posterior probability that a domain object $x \in [s(\mathbf{z}^t), \ell(\mathbf{z}^t))$ gets classified as $y = 1$ is just the proportion of initial segment concepts $\mathbf{i} \in [\mathbf{s}(\mathbf{z}^t), \mathbf{l}(\mathbf{z}^t))$ that contain x. Therefore, the Bayes-optimal classification for x is $y = 1$ just when $x \leqslant m = (s(\mathbf{z}^t) + \ell(\mathbf{z}^t))/2$.

To prove this formally, consider an arbitrary training sequence $\mathbf{z}^t = \langle\langle x_1, y_1\rangle, ..., \langle x_t, y_t\rangle\rangle$ consistent with some target segment $\mathbf{i} \in I$. We can compute the posterior probability that a particular domain object x gets classified as $y = 1$ as follows.

$$\begin{aligned} \mathrm{P}(y \mid x, \mathbf{z}^t) &= \int_0^1 p(y, \mathbf{i} \mid x, \mathbf{z}^t)\, di \\ &= \int_0^1 \mathrm{P}(y \mid \mathbf{i}, x, \mathbf{z}^t)\, p(\mathbf{i} \mid x, \mathbf{z}^t)\, di \\ &= \int_0^1 \mathrm{P}(y \mid \mathbf{i}, x)\, p(\mathbf{i} \mid \mathbf{z}^t)\, di, \end{aligned}$$

since y is independent of $\mathbf{z}^t$ given x and $\mathbf{i}$, and $\mathbf{i}$ is independent of x. Now, notice that

$$p(\mathbf{i} \mid \mathbf{z}^t) = \begin{cases} 0 & \text{if} \quad i \notin [s(\mathbf{z}^t), \ell(\mathbf{z}^t)) \\ \dfrac{1}{\ell(\mathbf{z}^t) - s(\mathbf{z}^t)} & \text{if} \quad i \in [s(\mathbf{z}^t), \ell(\mathbf{z}^t)), \end{cases}$$

where $s(\mathbf{z}^t)$ is the largest positive example and $\ell(\mathbf{z}^t)$ is the smallest negative example in $\mathbf{z}^t$, and notice that $\mathrm{P}(y = 1 \mid \mathbf{i}, x) = 1$ if $x > i$ and $\mathrm{P}(y = 1 \mid \mathbf{i}, x) = 0$ if $x \leqslant i$. Therefore, the posterior probability that an object x is classified as $y = 1$ given $\mathbf{z}^t$ is given by

$$\begin{aligned} \mathrm{P}(y = 1 \mid x, \mathbf{z}^t) &= \int_x^{\ell(\mathbf{z}^t)} \frac{1}{\ell(\mathbf{z}^t) - s(\mathbf{z}^t)}\, di \\ &= \frac{\ell(\mathbf{z}^t) - x}{\ell(\mathbf{z}^t) - s(\mathbf{z}^t)}. \end{aligned} \tag{4}$$

for $x \in [s(\mathbf{z}^t), \ell(\mathbf{z}^t))$. (Note that this posterior probability is 1 if $x < s(\mathbf{z}^t)$, and 0 if $x \geqslant \ell(\mathbf{z}^t)$.) Then, following the standard Bayes decision procedure: given $\mathbf{z}^t$, we classify x as 1 exactly when $\mathrm{P}(y = 1 \mid x, \mathbf{z}^t) \geqslant 1/2$. But by (4) this occurs when $x \leqslant (s(\mathbf{z}^t) + \ell(\mathbf{z}^t))/2$, which is exactly what MP's hypothesis does. ∎

Proof of Lemma 4. First, notice that $\mathrm{E}_{\mathbf{i}} \mathrm{E}_{\mathbf{x}^t} \operatorname{err}(\mathsf{MP}, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) = \mathrm{E}_{\mathbf{x}^t} \mathrm{E}_{\mathbf{i}} \operatorname{err}(\mathsf{MP}, \mathrm{U}, \mathbf{i}, \mathbf{x}^t)$ by Fubini's Theorem. Now consider an arbitrary (ordered) object sequence $\mathbf{x}^t = \{x_1 < x_2 < \cdots < x_t\}$ that partitions the chain I into $t+1$ subintervals $I_{[0, x_1)}, I_{[x_1, x_2)}, ..., I_{[x_t, 1]}$, where $I_{[x_n, x_{n+1})} = \{\mathbf{i} \in I : x_n \leqslant i < x_{n+1}\}$. I.e., each subinterval contains initial segment concepts that identically label the objects in $\mathbf{x}^t$. Consider an arbitrary subinterval $I_{[x_n, x_{n+1})} = [\mathbf{s}, \mathbf{l})$. For this subinterval, MP always guesses the same hypothesis, $\mathbf{m}$, defined by the endpoint $m = (s + \ell)/2$. Thus,

$$\begin{aligned} \int_s^\ell \operatorname{err}(\mathsf{MP}, \mathrm{U}, \mathbf{i}, \mathbf{x}^t)\, di &= \int_s^\ell d_{\mathrm{U}}(\mathbf{m}, \mathbf{i})\, di \\ &= \int_s^\ell |m - i|\, di \\ &= 2 \int_s^m m - i\, di \\ &= \tfrac{1}{4} \int_s^\ell \ell - s\, di \\ &= \tfrac{1}{4} \int_s^\ell \operatorname{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t)\, di. \end{aligned}$$

The result then follows since

$$\begin{aligned} \mathrm{E}_{\mathbf{i}} \operatorname{err}(\mathsf{MP}, \mathrm{U}, \mathbf{i}, \mathbf{x}^t) &= \sum_{n=0}^{t} \int_{x_n}^{x_{n+1}} \operatorname{err}(\mathsf{MP}, \mathrm{U}, \mathbf{i}, \mathbf{x}^t)\, di \\ &= \tfrac{1}{4} \int_0^1 \operatorname{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t)\, di \\ &= \tfrac{1}{4} \mathrm{E}_{\mathbf{i}} \operatorname{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t). \end{aligned}$$ ∎

Proof of Lemma 5. We want to use the result that a large error is forced on average over all $\mathbf{i} \in I$ to show that a large error must be forced for a significant proportion of the

concepts $\mathbf{i} \in I$ for each t, and then show that this means a large error must be forced for some particular $\mathbf{i}' \in I$ for infinitely many t.

Let $\mathrm{err}(\mathbf{i}, t) \stackrel{\text{def}}{=} \mathrm{E}_{\mathbf{x}^t} \mathrm{err}(L, \mathrm{U}, \mathbf{i}, \mathbf{x}^t)$. Focusing on the distribution Q over I, we are first interested in the event

$$B_t \stackrel{\text{def}}{=} \left\{ \mathbf{i} \in I : \mathrm{err}(\mathbf{i}, t) \geqslant \frac{1-\lambda}{2(t+2)} \right\},$$

which contains the concepts $\mathbf{i} \in I$ that force a large error for training sample size t. To prove that this event has significant positive measure under Q for every t, fix an arbitrary $t > 0$ and think of $R = \mathrm{err}(\mathbf{i}, t)$ as a random variable over $\mathbf{i}$. Then by (2) we know that $\mathrm{E}R \geqslant 1/(2(t+2))$, and by Lemma 2 we know that if L is consistent for I then $R \leqslant 2/(t+1)$. Combining these two facts gives $0 \leqslant R \leqslant 8\mathrm{E}R$. Now, letting $q \stackrel{\text{def}}{=} \mathrm{Q}(B_t) = \mathrm{Q}\{\mathbf{i} : R(\mathbf{i}) \geqslant (1-\lambda)\,\mathrm{E}R\}$, we have $\mathrm{E}R \leqslant (1-q)(1-\lambda)\,\mathrm{E}R + 8q\,\mathrm{E}R$. It is easy to see that this holds if and only if $q \geqslant \lambda/(\lambda+7)$. (Note that if L is not consistent for I then we can always construct a consistent learner L' such that $\mathrm{err}(L', \mathrm{U}, \mathbf{i}, \mathbf{x}^t) \leqslant \min\{\mathrm{err}(L, \mathrm{U}, \mathbf{i}, \mathbf{x}^t), \mathrm{wid}(I, \mathrm{U}, \mathbf{i}, \mathbf{x}^t)\}$ for all $\mathbf{i} \in I$, $\mathbf{x}^t \in X^t$; so it suffices to consider a consistent learner.)

Now we consider the event

$$B \stackrel{\text{def}}{=} \bigcap_{n=1}^{\infty} \bigcup_{t=n}^{\infty} B_t,$$

which contains the concepts $\mathbf{i} \in I$ that force L to exhibit $\mathrm{err}(\mathbf{i}, t) \geqslant (1-\lambda)/(2(t+2))$ for infinitely many training sample sizes t. We wish to prove that $\mathrm{Q}(B) > 0$ and hence there exists some $\mathbf{i}' \in B$ that forces $\mathrm{err}(\mathbf{i}', t) \geqslant (1-\lambda)/(2(t+2))$ for infinitely many t. To do this, let $B^T = \bigcap_{n=1}^{T} \bigcup_{t=n}^{\infty} B_t$. Notice that $B^T \downarrow B$, and hence $\mathrm{Q}(B^T) \downarrow \mathrm{Q}(B)$ by [2, Theorem 1.2.7]. But now see that $B_T \subseteq B^T$ for all T, and therefore $\mathrm{Q}(B^T) \geqslant \mathrm{Q}(B_T) \geqslant \lambda/(\lambda+7)$ for all T. This implies $\mathrm{Q}(B) \geqslant \lambda/(\lambda+7)$, and we are done. ▮

Proof of Lemma 6. The key reason we obtain rational convergence here is that, by construction, the region around any target concept is sufficiently dense to ensure (C, P) behaves like a uniform chain. To show this, consider an arbitrary concept $c \in C$ and let $[c-D, c+D]$ denote the subinterval of X containing all objects within a d_{P}-distance D of c. Also, let $r = 3^{1+\varepsilon}$. Then we have

P2.

$$\mathrm{E}_{\mathbf{x}^n}[\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^n) \mid x_1, \ldots, x_n \in [c-D, c+D]] > \frac{D}{r(n+1)}.$$

Given P2, it is easy to prove the lemma by a simple application of the Binomial Theorem

$$\begin{aligned}
&\mathrm{E}_{\mathbf{x}^t} \mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t) \\
&\quad > \sum_{n=0}^{t} \binom{t}{n} (2D)^n (1-2D)^{t-n}\, \mathrm{E}_{\mathbf{x}^n}[\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^n) \mid \\
&\qquad x_1, \ldots, x_n \in [c-D, c+D]] \\
&\quad > \sum_{n=0}^{t} \binom{t}{n} (2D)^n (1-2D)^{t-n} \frac{D}{r(n+1)} \\
&\quad \geqslant \frac{D}{r(t+1)}.
\end{aligned}$$

To prove P2: Consider an arbitrary concept c added at some stage K of the construction. Note that by the definition of P, c must have neighboring concepts (on both sides) at each distance $D_n = (3^{\varepsilon}-1)(3/2) \sum_{k=n+1}^{\infty} r^{-k} = \lambda \cdot r^{-n}$ for every $n \geqslant K$ (where λ is just a fixed positive constant). Let $D = D_K = \lambda \cdot r^{-K}$. This means that inside a local neighborhood $[c-D, c+D]$ of c, the chain (C, P) must behave as if it were a compressed version of the uniform chain (I, U). To see this, consider the right hand neighborhood $[c, c+D]$ and notice that the distribution of distances from c to its right-hand neighbors is bounded by r times a uniform(0, 1) distribution (Fig. 4). In particular, for $d \leqslant D$ we have $\mathrm{P}\{H^c(x) < d\} \leqslant r\mathrm{P}\{R < d\}$ where $R \sim$ uniform(0, 1). That is, given the event $x \in [c, c+D]$, the distribution for $H^c(x) \mid x \in [c, c+D]$ is upper bounded by a uniform(0, D/r) distribution, as shown in Fig. 4.

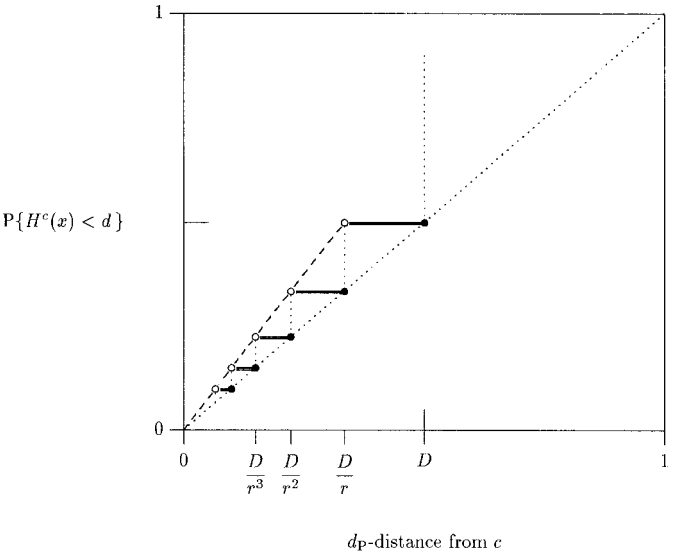

FIG. 4. The solid lines indicate the distribution of d_{P}-distances from a target concept c to its right-side neighbors in a dense chain C. Here $D = \lambda \cdot r^{-K}$, where $K = \mathrm{stage}(c)$, $r = 3^{1+\varepsilon}$, and λ is a fixed positive constant. The dashed line shows how the distribution of d_{P}-distances, given that the distance is less than D, is bounded by a *uniform* $(0, D/r)$ distribution.

Now, recalling that for positive random variables X and Y, $F_X \leqslant F_Y$ implies $\mathrm{E}X \geqslant \mathrm{E}Y$, we can simply apply the Binomial Theorem to obtain

$$\begin{aligned}
&\mathrm{E}_{\mathbf{x}^n}[\,\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^n) \mid x_1, ..., x_n \in [\,c-D, c+D\,]\,] \\
&\quad = \sum_{i=0}^{n} \binom{n}{i}\left(\frac{1}{2}\right)^n (\mathrm{E}_{\mathbf{x}^i}[\,\underline{S}^c_i \mid x_1, ..., x_i \in [\,c-D, c\,]\,] \\
&\qquad + \mathrm{E}_{\mathbf{x}^{n-i}}[\,\underline{H}^c_{n-i} \mid x_{i+1}, ..., x_n \in [\,c, c+D\,]\,]) \\
&\quad \geqslant \sum_{i=0}^{n} \binom{n}{i}\left(\frac{1}{2}\right)^n \left[\frac{D}{r(i+1)} + \frac{D}{r(n-i+1)}\right] \\
&\quad > \frac{D}{r(n+1)}. \quad \blacksquare
\end{aligned}$$

Proof of Lemma 7. The main reason MC obtains near-optimal average expected error is that the max-weight concept in any uncertainty interval, J, possesses a minimum fraction of J's total weight under Q. Let $q^* = \max_{c \in J} \mathrm{Q}(c)$ denote the maximum prior probability of any concept in J.

P3. *For any uncertainty interval* J: $q^* \geqslant \rho \mathrm{Q}(J)$ *for a fixed constant* $\rho = (1 - 3^{-\varepsilon})/2$.

Given P3, we prove the lemma as follows: First note that we can rearrange the order of summation to obtain $\mathrm{E}_c \mathrm{E}_{\mathbf{x}^t} \mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t) = \mathrm{E}_{\mathbf{x}^t} \mathrm{E}_c \mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t)$, so it suffices to consider an arbitrary fixed $\mathbf{x}^t$. Note that a sequence $\mathbf{x}^t = \{x_1 < x_2 < \cdots < x_t\}$ partitions the chain C into $t+1$ subintervals $C_{(\leftarrow, x_1)}$, $C_{(x_1, x_2)}$, ..., $C_{(x_t, \rightarrow)}$, where $C_{(x_n, x_{n+1})} = \{c \in C: x_n \in c \text{ and } x_{n+1} \notin c\}$. That is, the concepts in each subinterval identically label $\mathbf{x}^t$.

Now, consider an arbitrary subinterval $J = C_{(x_n, x_{n+1})}$ and compare the performance of MC to an arbitrary learner L in this subinterval. Since $\mathbf{x}^t$ is fixed, we can think of the target concepts $c \in J$ as being drawn randomly according to the distribution Q. Note that, since $c_1 \mathbf{x}^t = c_2 \mathbf{x}^t$ for any $c_1, c_2 \in J$, any learner must produce a fixed hypothesis for all target concepts $c \in J$. Thus, given targets $c \in J$, L produces a fixed hypothesis $L(c\mathbf{x}^t) = h$, and MC guesses the concept $c^* \in J$ with maximum prior probability according to Q. (The proof of P3 below notes that there can be (at most) two concepts with maximum prior probability in an uncertainty interval J, so we just assume MC deterministically picks one of them.) We now show that any hypothesis h must obtain an average error over random concepts drawn from J that is at least a fixed fraction of c^*'s average error. Let $q^* = \mathrm{Q}(c^*) = \max_{c \in J} \mathrm{Q}(c)$, $Q = \sum_{c \in J} \mathrm{Q}(c)$, $Q^* = Q - q^*$, and $d^* = d_{\mathrm{P}}(h, c^*)$. Then, by the triangle inequality, we get

$$\begin{aligned}
D(h) &\stackrel{\mathrm{def}}{=} \sum_{c \in J} d_{\mathrm{P}}(h, c)\, \mathrm{Q}(c) \\
&\geqslant d_{\mathrm{P}}(h, c^*)\, q^* + \sum_{c \in J - \{c^*\}} \theta[\, d_{\mathrm{P}}(c, c^*) \\
&\qquad - d_{\mathrm{P}}(h, c^*)\,]\, \mathrm{Q}(c) \\
&\geqslant d^* q^* + \theta[\, D(c^*) - d^* Q^*\,], \qquad (5)
\end{aligned}$$

where θ is a threshold function such that $\theta(x) = x$ if $x > 0$ and $\theta(x) = 0$ otherwise.

Now we just minimize this lower bound as a function of d^*. Here we have two cases: If $q^* > Q^*$ then (5) is minimized by choosing $d^* = 0$, which gives $D(h) \geqslant D(c^*)$. If, on the other hand, $q^* \leqslant Q^*$ then (5) is minimized by choosing $d^* = D(c^*)/Q^*$, which gives $D(h) \geqslant D(c^*)\, q^*/Q^*$. In this case we can just apply P3 to obtain $D(h) \geqslant D(c^*)\, \rho/(1-\rho)$. Thus, in either case we obtain $D(h) \geqslant \beta D(c^*)$ for a fixed constant $\beta \geqslant \rho/(1-\rho) > 0$.

Finally, we note that this shows any learner L must obtain

$$\begin{aligned}
\mathrm{E}_c \mathrm{err}(L, \mathrm{P}, c, \mathbf{x}^t) &= \sum_{n=0}^{t} \sum_{c \in C_{(x_n, x_{n+1})}} d_{\mathrm{P}}(L(c\mathbf{x}^t), c)\, \mathrm{Q}\{c\} \\
&\geqslant \beta \sum_{n=0}^{t} \sum_{c \in C_{(x_n, x_{n+1})}} d_{\mathrm{P}}(\mathsf{MC}(c\mathbf{x}^t), c)\, \mathrm{Q}\{c\} \\
&= \beta\, \mathrm{E}_c \mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t).
\end{aligned}$$

*Proof of P*3. Note that any subinterval J of C contains at least one, and at most two concepts of maximum weight under Q. (This is true since, by construction, any interval that contains three concepts from a stage K, must contain at least one concept from an earlier stage $N < K$; cf. Fig. 2.) Therefore, if c^* is a max-weight concept of J, then J can have: at most one other concept of maximum weight $q^* = \mathrm{Q}(c^*)$; at most 6 concepts of weight q^*/r, $r = 3^{1+\varepsilon}$; at most $2 \cdot 3^k$ concepts of weight q^*/r^k for all stages $k \geqslant 0$; etc. (That is, this is symmetric to considering $J = C - \{c_0, c_1\}$ in Fig. 2 and choosing $c^* = c_2$.) This gives a total weight of $\mathrm{Q}(J) \leqslant q^* \sum_{k=0}^{\infty} 2 \cdot 3^k r^{-k} = 2q^*/(1 - 3^{-\varepsilon})$. $\blacksquare$

Proof of Lemma 8. As in Lemma 7 we can rearrange the order of summation to obtain the identity $\mathrm{E}_c \mathrm{E}_{\mathbf{x}^t} \mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t) = \mathrm{E}_{\mathbf{x}^t} \mathrm{E}_c \mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t)$. So consider an arbitrary fixed $\mathbf{x}^t = \{x_1 < x_2 < \cdots < x_t\}$; think of the target concept c as being randomly drawn according to Q, and consider MC's performance for $\mathbf{x}^t$. First note that there is a nonzero probability that MC guesses the target concept exactly. So we need to argue that (a) MC does not guess the target with too high a probability; and (b) given that MC does not guess the target, it must achieve an average error that is at least a fixed fraction of the uncertainty interval width. Let $\neg\mathsf{MC}(\mathbf{x}^t) \stackrel{\mathrm{def}}{=} \{c: \mathsf{MC}(c\mathbf{x}^t) \neq c\}$ denote the set of concepts in

C that MC does *not* guess given any possible labelling of the object sequence $\mathbf{x}^t$. Then for any fixed $\mathbf{x}^t$ we have

P4. $\mathrm{Q}(\neg\mathsf{MC}(\mathbf{x}^t)) \geqslant 1/3^{\varepsilon}(t+2)^{\varepsilon}$.

P5.
$$\mathrm{E}_c[\mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t) \mid \neg\mathsf{MC}(\mathbf{x}^t)]$$
$$\geqslant \frac{1}{8r^3}\,\mathrm{E}_c[\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t) \mid \neg\mathsf{MC}(\mathbf{x}^t)]$$
$$\times\, \mathrm{Q}(\neg\mathsf{MC}(\mathbf{x}^t)) \qquad \text{for} \quad r = 3^{1+\varepsilon}.$$

Given these two facts, it is easy to prove the lemma as follows. Applying P4 and P5 yields

$$\mathrm{E}_c\,\mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t)$$
$$= \mathrm{E}_c[\mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t) \mid \neg\mathsf{MC}(\mathbf{x}^t)]\; Q(\neg\mathsf{MC}(\mathbf{x}^t))$$
$$\geqslant \mathrm{E}_c[\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t) \mid \neg\mathsf{MC}(\mathbf{x}^t)]\; \delta(t+2)^{-2\varepsilon},$$

for a constant $\delta = (8r^3 3^{2\varepsilon})^{-1} > 0$. Now, averaging over $\mathbf{x}^t$ and re-arranging the sum yields

$$\mathrm{E}_{\mathbf{x}^t}\mathrm{E}_c\,\mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t)$$
$$\geqslant \delta(t+2)^{-2\varepsilon}\,\mathrm{E}_{\mathbf{x}^t}\mathrm{E}_c[\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t) \mid \neg\mathsf{MC}(\mathbf{x}^t)]$$
$$= \delta(t+2)^{-2\varepsilon}\,\mathrm{E}_c\mathrm{E}_{\mathbf{x}^t}[\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t) \mid \neg\mathsf{MC}(c)],$$

where $\neg\mathsf{MC}(c) \stackrel{\text{def}}{=} \{\mathbf{x}^t : \mathsf{MC}(c\mathbf{x}^t) \neq c\}$ is the set of object sequences $\mathbf{x}^t \in X^t$ where MC does not guess c. This proves the lemma, since for any c we have

$$\mathrm{E}_{\mathbf{x}^t}[\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t) \mid \neg\mathsf{MC}(c)] \geqslant \mathrm{E}_{\mathbf{x}^t}\,\mathrm{wid}(C, \mathrm{P}, c, \mathbf{x}^t).$$

(Intuitively, this follows because the uncertainty intervals where c is the max-weight concept tend to be small. That is, consider a fixed left boundary of an uncertainty interval around c and notice that every right boundary that gives $c \in \neg\mathsf{MC}(c)$ is strictly further away from c than any boundary where $\mathsf{MC}(c\mathbf{x}^t) = c$.)

Proof of P4. We get this bound because $\mathbf{x}^t$ partitions C into at most $t+1$ subintervals and MC can guess at most one concept per subinterval. Thus, the probability that MC guesses a random target concept $c \in C$ is bounded by the sum of the largest $t+1$ probabilities in C. That is, $\mathrm{Q}(\neg\mathsf{MC}(\mathbf{x}^t)) \geqslant \sum_{i=t+2}^{\infty} q_i$ where $\{q_i\}_{i=2}^{\infty}$ is the sequence of probabilities assigned in the construction of Q. So, letting $Q_T = \sum_{i=T}^{\infty} q_i$, we seek a lower bound on Q_T for $T = t+2$.

To determine this lower bound, note that by the construction of C, the total number of concepts added in Stages 1 through K inclusive is $\sum_{k=1}^{K} 2 \cdot 3^{k-1} = 3^K - 1$, so the index of the last concept added at Stage K is 3^K. This means that $Q_T \geqslant \sum_{k=\lceil 1+\log_3 T\rceil}^{\infty} Q_k$, where $Q_k \stackrel{\text{def}}{=} (3^{\varepsilon}-1)\,3^{-\varepsilon k}$ is the total probability assigned at Stage k of the construction. Thus

$$\mathrm{Q}_T \geqslant (3^{\varepsilon}-1) \sum_{k=\lceil 1+\log_3 T\rceil}^{\infty} 3^{-\varepsilon k}$$
$$= (3^{\varepsilon}-1)\,\frac{(3^{-\varepsilon})^{\lceil 1+\log_3 T\rceil}}{1-3^{-\varepsilon}}$$
$$= 3^{\varepsilon}(3^{-\varepsilon})^{\lceil 1+\log_3 T\rceil}$$
$$\geqslant 3^{\varepsilon}(3^{-\varepsilon})^{2+\log_3 T} = 3^{-\varepsilon}T^{-\varepsilon}.$$

Proof of P5. This inequality holds because the region around the max-weight concept in any uncertainty interval is sufficiently dense in both P and Q to simulate the effects of a uniform prior on a uniform chain (as in Lemma 4). Here we are interested in the conditional distribution of Q given $\neg\mathsf{MC}(\mathbf{x}^t)$, which is defined by

$$\mathrm{Q}^*(c) = \begin{cases} \mathrm{Q}(c)/Q^* & \text{if} \quad c \in \neg\mathsf{MC}(\mathbf{x}^t), \\ 0 & \text{otherwise,} \end{cases}$$

where Q^* is the normalizing constant given by $Q^* = \mathrm{Q}(\neg\mathsf{MC}(\mathbf{x}^t))$. Let E_c^* denote expectation over c with respect to this conditional distribution. We seek a lower bound on

$$\mathrm{E}_c^*\,\mathrm{err}(\mathsf{MC}, P, c, \mathbf{x}^t)$$
$$= \sum_{n=0}^{t} \int_{c \in (x_n, x_{n+1})} \mathrm{err}(\mathsf{MC}, P, c, \mathbf{x}^t)\, d\mathrm{Q}^*(c), \qquad (6)$$

where (x_n, x_{n+1}) denotes the subinterval of concepts $c \in C$ between x_n and x_{n+1}.

To establish this lower bound, consider an arbitrary subinterval $J = (x_n, x_{n+1})$ and let $c^* = \arg\max_{c \in J} \mathrm{Q}(c)$. Note that we can split the summation over J into two halves

$$\int_{c \in J} \mathrm{err}(\mathsf{MC}, \mathrm{P}, c, \mathbf{x}^t)\, d\mathrm{Q}^*(c)$$
$$= \int_{c \in (x_n, c^*)} d_{\mathrm{P}}(c^*, c)\, d\mathrm{Q}^*(c)$$
$$+ \int_{c \in (c^*, x_{n+1})} d_{\mathrm{P}}(c^*, c)\, d\mathrm{Q}^*(c). \qquad (7)$$

So consider one of the half intervals $M = (c^*, x_{n+1})$. It is not hard to show that the average d_{P}-distance from c^* to $c \in M$ is at least a fixed fraction of M's width under P: To see this, note that for each $c \in (c^*, x_{n+1})$ added at Stage K of the construction we can assign a distinct x from Stage $K+1$ between c^* and c—namely, the x at Stage $K+1$ closest to c in (c^*, c); see Fig. 2. Then, for any subinterval (c^*, c) we

get $P(c^*, c) \geq Q(c^*, c)/r = Q^*(c^*, c)\, Q^*/r$, where $r = 3^{1+\varepsilon}$. (Here we are using the notation (c^*, c) to refer ambiguously to both the set of concepts and the set of domain objects between c^* and c. The intended meaning should be clear from context.) This means that for any d_P-distance d such that $d \leq P(M)$ we get $Q^*\{c \in M: d_P(c^*, c) \leq d\} < rd/Q^*$. So, thinking of $d_P(c^*, c)$ as a random variable over c, we can see that the distribution function for $d_P(c^*, c) \mid c \in M$ is bounded by a uniform$(0, P(M)\, Q^*/r)$ distribution, as shown in Fig. 5. Thus we get $E_c^*[d_P(c^*, c) \mid M] \geq P(M) Q^*/(2r)$, and hence

$$\int_{c \in M} d_P(c^*, c)\, dQ^*(c) = E_c^*[d_P(c^*, c) \mid M]\; Q^*(M) \geq \frac{Q^*}{2r} P(M)\, Q^*(M). \tag{8}$$

Now, reconsidering the complete interval J, note that one of the half intervals (x_n, c^*) or (c^*, x_{n+1}) must be at least half the d_P-width of J. Without loss of generality, assume $P(M) \geq P(J)/2$. Then we can argue that $Q^*(M) \geq P(M)/r \geq P(J)/(2r) \geq Q^*(J)/(2r^2)$ as above. Combining this with (7) and (8) gives

$$\int_{c \in J} \mathrm{err}(\mathsf{MC}, P, c, \mathbf{x}^t)\, dQ^*(c) = \int_{c \in J} d_P(c^*, c)\, Q^*(c) \geq \frac{Q^*}{8r^3} P(J)\, Q^*(J) = \frac{Q^*}{8r^3} \int_{c \in J} \mathrm{wid}(J, P)\, dQ^*(c).$$

Substituting this back into (6) yields the stated bound. ∎

Proof of Lemma 9. Follows from essentially the same argument as Lemma 5. ∎

Proof of Lemma 10. First we need to formalize the notion of the *order* of a limit concept.

DEFINITION 6 (Limits and Order). For a chain C, let $i(C)$ denote the set of *isolated* concepts in C; i.e., the concepts with a least-larger and greatest-smaller neighbor in C. Also let $C^0 = i(C)$. Then we define $C^{1+} = C - C^0$ to be the *limit* concepts of C. The limit concepts with order exactly 1 are given by $C^1 = i(C^{1+})$. Continuing in this way for arbitrary ordinals α, we define the concepts with order *at least* α by $C^{\alpha+} = C - \bigcup_{\beta<\alpha} C^\beta$, and the concepts with order *exactly* α by $C^\alpha = i(C^{\alpha+})$. Notice that for any ordinal α we have $C = C^{\alpha+} \cup \bigcup_{\beta<\alpha} C^\beta$.

The key issue is to show that by collecting successively higher order limit concepts in this way we eventually exhaust a scattered chain.

P6. *For any scattered chain C, there exists some least ordinal γ such that $C = \bigcup_{\beta \leq \gamma} C^\beta$.*

Note that by the definition of C^β, P6 implies both properties (i) and (ii) of Lemma 10. To prove P6 we must resort to an inductive characterization of scattered linear orderings first developed by Hausdorff (an excellent treatment of this subject is given in Rosenstein's monograph [21]). This characterization is based on constructing the following "condensation" map: We say that two concepts are a finite distance apart if there are only finitely many concepts between them in the ordering. Then the *finite condensation* map $f: C \to 2^C$ is defined by $f(c) = \{d \in C: c \text{ and } d \text{ are a finite distance apart}\}$. The effect of this map is to collapse the chain into a collection of subintervals (that is, subintervals of the chain C, not the domain X). The key point is to notice that these subintervals *themselves* form a linear-ordering, so we can naturally define iterates of this map as follows: For a *successor* ordinal $\beta + 1$, define $f^{\beta+1}(c) = \bigcup \{f^\beta(d): f^\beta(d) \text{ and } f^\beta(c) \text{ are a finite distance apart}\}$; and for any *limit* ordinal λ, define $f^\lambda(c) = \bigcup_{\beta<\lambda} \{f^\beta(c)\}$. Then we have the following relations.

P7. *$f^\beta(C) = f^\alpha(C)$ for all $\beta \geq \alpha$ if and only if $f^\alpha(C)$ is dense or a singleton.*

(This proposition is more or less immediate from the definitions; see e.g., [21, p. 81].) Now define γ to be the *least* ordinal for which $f^\beta(C) = f^\gamma(C)$ for all $\beta \geq \gamma$. (We know that such a γ must exist, since for any chain C with cardinality κ there is an ordinal $\alpha < \kappa+$ such that $f^\beta(C) = f^\alpha(C)$ for all $\beta \geq \alpha$ [21, Theorem 5.9].) From P7 it is easy to see that

P8. *$f^\gamma(C)$ is a singleton if and only if C is scattered* [21, *Exercise* 5.11.2].

This gives a necessary and sufficient characterization of scattered concept chains in terms of $f^\gamma(C)$. So now all we need to do is related this characterization of a scattered chain C to its decomposition into limit points given in Definition 6 above. Below we prove

P9. *For any ordinal β, there can be at most 2 concepts from $C^{\beta+}$ in any internal of $f^\beta(C)$.*

This gives the result, since: From P8 we know that if C is scattered then $f^\gamma(C)$ is a single interval. Combined with P9, this means $C^{\gamma+}$ contains at most 2 concepts, and hence $C^{(\gamma+1)+} = \varnothing$. Since by definition $C = C^{\alpha+} \cup \bigcup_{\beta<\alpha} C^\beta$ for any α, we have shown that $C = \bigcup_{\beta \leq \gamma} C^\beta$; establishing P6 and hence the lemma.

*Proof of P*9. Proof is by induction on ordinals. *Base*: Simply define f^0 to be the singleton intervals of C. *Successor ordinal*: For any successor ordinal $\beta + 1$ we know there are at most 2 concepts from $C^{\beta+}$ in any subinterval of f^β by the induction hypothesis. Now assume there are 3 concepts

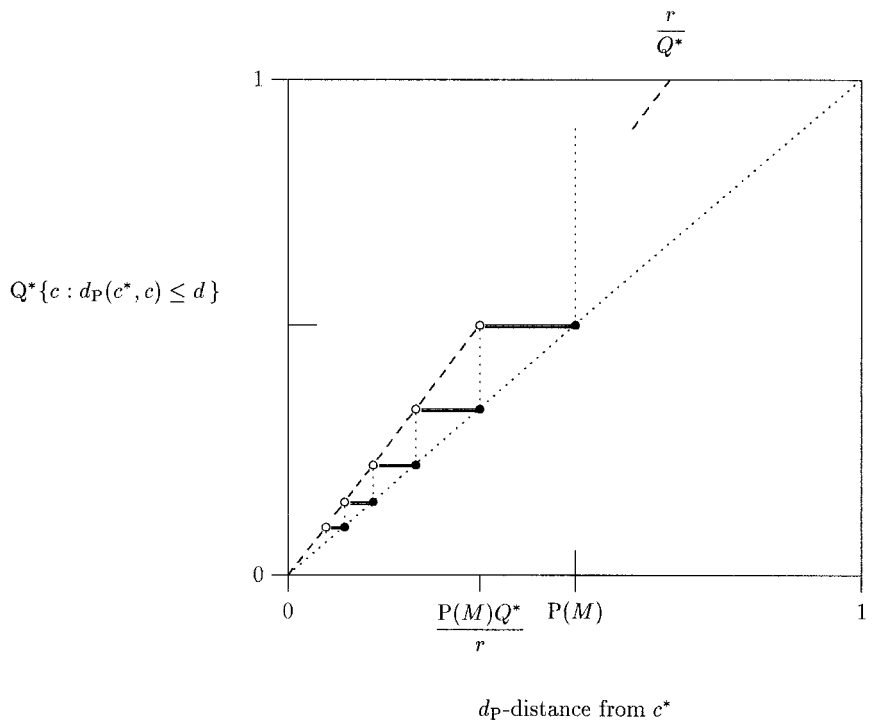

FIG. 5. The solid lines indicate the Q*-distribution of d_P-distances from a max-weight target concept c^* to its right-side neighbors in a dense chain C. The dashed line shows how the Q*-distribution of d_P-distances, given that the distance is less than $P(M)$, is bounded by a *uniform* $(0, P(M)\,Q^*/r)$ distribution.

$a \subset b \subset c$ from $C^{(\beta+1)+}$ in a single interval of $f^{\beta+1}$. Since they are in a single interval in $f^{\beta+1}$, these concepts must have belonged to intervals in f^{β} that were only a finite distance apart. But then, by the induction hypothesis, $a \subset b \subset c$ must only be a finite distance apart in $C^{\beta+}$. But this means b must be isolated in $C^{\beta+}$, and hence cannot belong to $C^{(\beta+1)+}$; a contradiction. *Limit ordinal*: For any limit ordinal λ we know there are at most 2 concepts from $C^{\beta+}$ in any subinterval of f^{β}, for all $\beta < \lambda$, by the induction hypothesis. Now assume there are 3 concepts $a \subset b \subset c$ from $C^{\lambda+}$ in a single interval of f^{λ}. Since they are in a single interval in f^{λ}, there must be some $\beta_1 < \lambda$ such that a and b belong to an interval of f^{β_1}, and some $\beta_2 < \lambda$ such that b and c belong to an interval of f^{β_2}. But then all 3 concepts must belong to a single interval of f^{β} for $\beta = \max\{\beta_1, \beta_2\} < \lambda$; a contradiction. ∎

Proof of Lemma 11. We define the usual topological concepts for linear orderings (see e.g., [21, Chapter 2]).

DEFINITION 7 (Compactness Properties). For a chain C, a *Dedekind cut* of C is a partition of C into two nonempty subsets $\langle U, V\rangle$ where $u \subset v$ for all $u \in U$ and $v \in V$. A *gap* is a Dedekind cut $\langle U, V\rangle$ where U has no maximal concept and V has no minimal concept. We say that a chain C is: (i) *closed* if it is closed under $\cap$ and $\cup$; (ii) *bounded* if it has a minimal and maximal concept; (iii) *Dedekind-complete* if it has no gaps; (iv) *complete* if every upper (lower) bounded subchain of C has a least upper (greatest lower) bound in C; and (v) *Bolzano–Weierstrass* if every infinite subchain of C has a limit in C.

The definitions and results of this lemma will be used to prove the next two lemmas below. First observe that being closed under $\cup$, $\cap$ implies being Dedekind-complete and bounded, complete and bounded, and Bolzano-Weierstrass. Therefore we call such a chain *compact*.

((i) ⇒ (ii)) Obvious: the maximal concept is just $\bigcup\{c \in C\}$ and the minimal concept is $\bigcap\{c \in C\}$, which both must be in C since it is closed under $\cap$, $\cup$.

((i) ⇒ (iii)) Consider any partition $\langle U, V\rangle$ of C. Since C is closed under $\cap$, $\cup$, both $\bigcup\{u \in U\}$ and $\bigcap\{v \in V\}$ are in C, meaning that $\langle U, V\rangle$ cannot be a gap.

((iii) ⇒ (iv)) For any subchain A of C consider the partition defined by

$$U = \{u \in C: \exists a \in A \text{ such that } u \subseteq a\}.$$

Since $\langle U, V\rangle$ cannot be a gap, U must have a maximal concept or V a minimal concept. In either case we can supply a least upper bound on A.

((ii) + (iv) ⇒ (v)) Without loss of generality, consider a countably infinite subchain A of C. Since C is complete and bounded, A must have a greatest lower bound $a_1 \in C$. If

$a_1 \notin A$ we are done (since then a_1 would be a limit concept of A), so assume $a_1 \in A$. Continuing in this way, find $a_2 = glb(A - \{a_1\})$, $a_3 = glb(A - \{a_1, a_2\})$, etc. Note that if any a_i is not in A then it must be a limit concept of A, so we are left with the case where every $a_1 \subset a_2 \subset \cdots$ belongs to A. But then, since C is complete and bounded, $b = lub(\{a_i\}_1^\infty)$ must also be in C, and since $b \notin \{a_i\}_1^\infty$ by construction, this must be a limit concept of A. ▮

Proof of Lemma 12. Note that every concept $c \in \mathscr{C}(C)$ either belongs to C, or is given by $c = \bigcup\{u \in U\}$ or $c = \bigcap\{v \in V\}$ for some gap $\langle U, V\rangle$. Thus, $\mathscr{C}(C) = C \cup \mathscr{U}(C) \cup \mathscr{V}(C)$, where $\mathscr{U}(C)$ denotes the concepts added for gaps where U has no maximal concept, and $\mathscr{V}(C)$ denotes the concepts added for gaps where V has no minimal concept. Assume that C is scattered but $\mathscr{C}(C)$ is somewhere-dense. We will show that this leads to a contradiction.

First, since $\mathscr{C}(C)$ is somewhere-dense the following proposition shows that one of $\mathscr{U}(C)$ or $\mathscr{V}(C)$ must also be somewhere-dense.

P10. *Removing a scattered subchain S from a dense chain D leaves a somewhere-dense chain $D - S$.*

(This proposition is easy to prove: Since S is scattered there must be two concepts $s_1 \subset s_2$ in S with no $s_3 \in S$ between them. But then the subinterval (s_1, s_2) of D, which is dense, is properly contained in $D - S$.) So, without loss of generality, assume $\mathscr{U}(C)$ contains a dense subchain U. Notice that C is *between* U in the sense that for every pair $u_1 \subset u_2$ in U there must be a $c \in C$ such that $u_1 \subset c \subset u_2$ (for if not, then there would be two distinct gaps $\langle U_1, V_1\rangle$, $\langle U_2, V_2\rangle$ of C with $U_2 - U_1 = \varnothing$, which cannot be). But then the following proposition shows that C must be somewhere-dense as well; a contradiction.

P11. *If a chain B is between a dense chain D, then B must also be somewhere-dense.*

To prove this proposition, first note that we can find concepts $b_1, b_2 \in B$, $d_1, d_2 \in D$ such that $b_1 \subset d_1 \subset d_2 \subset b_2$ (just pick four concepts $d_3 \subset d_1 \subset d_2 \subset d_4$ from D and choose b_1, b_2 between the first and last pairs respectively). Now, for any such quadruple $b_1 \subset d_1 \subset d_2 \subset b_2$ we can always find $b_3 \in B$ and $d_3, d_4 \in D$ such that $b_1 \subset d_1 \subset d_3 \subset b_3 \subset d_4 \subset d_2 \subset b_2$ (just choose d_3 and d_4 between d_1 and d_2, and then b_3 between d_3 and d_4). Thus, we can continue this process indefinitely to construct a dense subchain of B. ▮

Proof of Lemma 13. Recall that any sequence of training examples $c\mathbf{x}^t$ determines an uncertainty interval $[s(c\mathbf{x}^t), \ell(c\mathbf{x}^t)]$. Also recall from the definition of an uncertainty interval (Definition 3) that s and ℓ are defined by unions and intersections of concepts from C, and hence must also belong to C. This means that $[s, \ell]$ is a *compact* subinterval of C (i.e., $[s, \ell]$ is also closed under $\cap$, $\cup$). Since $[s, \ell]$ is also scattered, by Lemma 10 we know that there exists a *least* ordinal γ such that $[s, \ell] = \bigcup_{\beta \leqslant \gamma} [s, \ell]^\beta$. (That is, γ is the least ordinal such that all limit concepts in $[s, \ell]$ have order at most γ.)

It suffices to show that $[s, \ell]^\gamma$ is non-empty, as this will supply the needed maximal order limit concepts. Assume $[s, \ell]^\gamma$ is empty. Then clearly γ must be infinite and $[s, \ell]^\beta$ must be non-empty for all $\beta < \gamma$ (otherwise γ would not be the least such ordinal). But then consider the subchain $\{c^\beta \in C^\beta : \beta < \gamma\}$ of $[s, \ell]$ formed by choosing a single limit concept of each order $\beta < \gamma$. By the compactness of $[s, \ell]$, this infinite subchain must have a *limit* concept c in $[s, \ell]$ (cf. the Bolzano–Weierstrass property of Lemma 1]). This concept c cannot be in $[s, \ell]^\beta$ for any $\beta < \gamma$, and hence must belong to $[s, \ell]^{\gamma+}$; a contradiction. ▮

REFERENCES

1. S. Amari, N. Fujita, and S. Shinomoto, Four types of learning curves, *Neural Comput.* **4** (1992), 605–618.
2. R. B. Ash, "Real Analysis and Probability," Academic Press, San Diego, 1972.
3. E. Barnard, A model for nonpolynomial decrease in error rate with increasing sample size, *IEEE Trans. Neural Networks* **5** (1994), 994–997.
4. E. B. Baum and Y.-D. Lyuu, The transition to perfect generalization in perceptrons, *Neural Comput.* **3** (1991), 386–401.
5. R. A. Brualdi, "Introductory Combinatorics," North-Holland, New York, 1977.
6. D. Cohn and G. Tesauro, Can neural networks do better than the Vapnik–Chervonenkis bounds?, *in* "Advances in Neural Information Processing Systems," Vol. 3, (D. Touretzky, Ed.), pp. 911–917, Morgan Kaufmann, San Mateo, CA, 1990.
7. D. Cohn and G. Tesauro, How tight are the Vapnik–Chervonenkis bounds?, *Neural Comput.* **4** (1992), 249–269.
8. R. O. Duda and P. Hart, "Pattern Classification and Scene Analysis," Wiley, New York, 1973.
9. R. M. Dudley, S. Kulkarni, T. Richardson, and O. Zeitouni, A metric entropy bound is not sufficient for learnability, *IEEE Inform. Theory* **40** (1994), 883–885.
10. M. Golea and M. Marchand, Average case analysis of the clipped Hebb rule for nonoverlapping perceptron networks, *in* "Proceedings of the Sixth Annual Workshop on Computational Learning Theory, 1993," pp. 151–157.
11. G. Györgyi, First-order transition to perfect generalization in a neural network with binary synapses, *Phys. Rev. A* **41** (1990), 7097–7100.
12. D. Haussler, M. Kearns, and R. Schapire, Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension, *in* "Proceedings of the Fourth Annual Workshop on Computational Learning Theory, 1991," pp. 67–74.
13. D. Haussler, M. Kearns, H. S. Seung, and N. Tishby, Rigorous learning curve bounds from statistical mechanics, *in* "Proceedings of the Seventh Workshop on Computational Learning Theory, 1994," pp. 76–87.
14. D. Haussler, N. Littlestone, and M. K. Warmuth, Predicting $\{0, 1\}$-functions on randomly drawn points, *in* "Proceedings of the First Workshop on Computational Learning Theory, 1988," pp. 280–296.
15. D. Haussler, N. Littlestone, and M. K. Warmuth, Predicting $\{0, 1\}$-functions on randomly drawn points, *Inform. and Comput.* **115** (1994), 248–292.

16. P. Langley, W. Iba, and K. Thompson, An analysis of Bayesian classifiers, *in* "Proceedings of the Tenth National Conference on Artificial Intelligence, 1992," pp. 223–228.
17. R. J. Larsen and M. L. Marx, "An Introduction to Mathematical Statistics and Its Applications," Prentice–Hall, Englewood Cliffs, NJ, 1981.
18. Y.-D. Lyuu and I. Rivin, Tight bounds on transition to perfect generalization in perceptrons, *Neural Comput*. **4** (1992), 854–862.
19. M. Opper and D. Haussler, Generalization performance of Bayes optimal classification algorithm for learning a perceptron, *Phys. Rev. Lett*. **66** (1991), 2677–2680.
20. M. J. Pazzani and W. Sarrett, Average case analysis of conjunctive learning algorithms, *in* "Proceedings of Seventh International Conference on Machine Learning, 1990," pp. 339–347.
21. J. G. Rosenstein, "Linear Orderings," Academic Press, New York, 1982.
22. D. Schuurmans, "Effective Classification Learning," Ph.D. thesis, University of Toronto, Department of Computer Science, January 1996 [also available as Tech. Rep. KRR-TR-96-1, University of Toronto, Department of Computer Science].
23. D. B. Schwartz, V. K. Samalam, S. A. Solla, and J. S. Denker, Exhaustive learning, *Neural Comput*. **2** (1990), 374–385.
24. H. S. Seung, H. Sompolinsky, and N. Tishby, Learning curves in large neural networks, *in* "Proceedings of the Forth Annual Workshop on Computational Learning Theory, 1991," pp. 112–127.
25. L. G. Valiant, A theory of the learnable, *Comm. ACM* **27** (1984), 1134–1142.
26. V. N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory Probab. Appl*. **16** (1971), 264–280.
27. V. N. Vapnik, E. Levin, and Y. Le Cun, Measuring the VC-dimension of a learning machine, *Neural Comput*. **6** (1994), 851–876.