Appears in
Proceedings of the Second European Conference on Computational Learning Theory (EuroCOLT-95), Barcelona, Spain, March 1995.

Characterizing Rational versus Exponential Learning Curves

Dale Schuurmans

Department of Computer Science
University of Toronto
Toronto, Ontario M5S 1A4
CANADA
dale@cs.toronto.edu

Abstract. We consider the standard problem of learning a concept from random examples. Here a *learning curve* can be 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 concept class C converges *rationally* to zero error (*i.e.*, $\Theta(1/t)$ for training sample size t). However, recently Cohn and Tesauro have demonstrated how *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. These results support the experimental findings of Cohn and Tesauro: for *finite* concept classes, *any* consistent learner achieves exponential convergence, even in the worst case; but for *continuous* concept classes, *no* learner can exhibit sub-rational convergence for every target concept and domain distribution. A precise boundary between rational and exponential convergence is drawn for simple concept *chains*. Here we show that *somewhere dense* chains always force rational convergence in the worst case, but exponential convergence can always be achieved for *nowhere dense* chains.

1 Introduction

1.1 Model

We consider 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 $\langle x, 1_c(x)\rangle$ consisting of a domain object $x \in X$ and the value of c's indicator function $1_c(x) \in \{0,1\}$ at x. Formally, a *learner* L maps sequences of training examples $\langle\langle x_1, \ell_1\rangle, \langle x_2, \ell_2\rangle, ..., \langle x_t, \ell_t\rangle\rangle$, $x_i \in X$, $\ell_i \in \{0,1\}$, $t \in I\!N$, to hypotheses $h \subset X$; *i.e.*, $L : (X \times \{0,1\})^* \to 2^X$. After a fixed training period t, any hypothesis L produces is then tested *ad infinitum* on subsequent test examples. A hypothesis h makes an *error* on any test example $\langle x, \ell\rangle$ for which $1_h(x) \neq \ell$.

As in most theoretical analyses of concept learning, we adopt the *iid* 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 (*i.e.*, *noise free* training examples). This is a natural model of most practical learning situations where the precise sequence of training examples is unpredictable and there is no correlation between successive examples. Here, the *error* of a hypothesis h with respect to target concept c and a domain distribution P is simply given by $\mathrm{P}\{x \in X : x \in h \triangle c\} = d_{\mathrm{P}}(h, c)$.

Intuitively, a *learning curve* measures the quality of a learner's hypotheses as a function of training sample size. For a fixed training sample size t, the *expected error* of L with respect to c and P is just the average error of L's hypotheses after observing t training examples; written $\mathrm{E}_{\mathrm{P}^t}\, err(L, c)$.[1] We define L's **learning curve** with respect to c and P by the expected error it obtains as a function of training sample size t. Intuitively, this is what one measures by repeatedly training a learning system (on a fixed problem) with various sample sizes and plotting the average error obtained as a function of training sample size. The *quality* of a learning curve is measured by the rate at which it converges to zero error.

Of course, for specific c and P, the quality of a learner's curve depends on its prior knowledge. For example, if the exact identity of c 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 adopt the model of prior knowledge first introduced by Valiant [Val84]: we assume the target concept c is known to belong to some class C, but that nothing is known about the domain distribution 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 possible target concepts $c \in C$ and domain distributions P.

An analysis of this form has been carried out by Haussler *et al.* [HLW88] who develop a special learning strategy 1IGPS (for "1-inclusion graph prediction strategy") which, given t training examples, always attains an expected error of at most

$$\mathrm{E}_{\mathrm{P}^t}\, err(\mathsf{1IGPS}, c) \leq \frac{2d}{t+1} \tag{1}$$

for any target concept $c \in C$ and any domain distribution P (where d is the "Vapnik-Chervonenkis dimension" of C) [HLW88, Theorem 5.1].[2] Moreover, they show that *no* learner can do significantly better than this in the worst case: given $t > d$ training examples, *any* learner L must obtain an expected error of at least

$$\mathrm{E}_{\mathrm{P}^t}\, err(L, c) \geq \frac{d-1}{2e(t+1)} \tag{2}$$

[1] This is identical to the probability that L correctly classifies the $t+1$st random example after training on the first t random examples [HLW88, Lemma 6.1].

[2] This result has been improved to $d/(t+1)$ for a slightly modified strategy [HLW90].

for some domain distribution P and target concept $c \in C$ [HLW88, Theorem 5.2]. Overall, this shows that the best achievable worst case expected error behaves as a "rational" function of t (*i.e.*, $\Theta(d/t)$) for any C with finite VCdimension.[3]

1.2 Issue

These results would seem to suggest that we should always expect to obtain *rational* learning curves for any concept class C, at least in the worst case (whenever worst case convergence to zero error is in fact possible). However, it turns out that one does not always observe rational learning curves in practice. This is clearly demonstrated in a recent study by Cohn and Tesauro who show that *exponential* learning curves can be obtained in many experimental settings [CT90, CT92]. Of course, these experimental results do not directly contradict the previous theory, since this exponential convergence behavior was only demonstrated for *specific* target concepts and domain distributions, and may not accurately reflect the *worst case* behavior. It could simply be that the worst case distribution free theory does not capture the typical learning curve behavior observed in certain practical situations.

However, Cohn and Tesauro also observe *rational* learning curves in many situations — in fact, obtaining results that closely match the worst case predictions of (1). Specifically, they ran their experiments in pairs, testing identical neural network architectures defined on $\{0,1\}^n$ and $[0,1]^n$, respectively. These paired concept classes have the same (or comparable) VCdimension, and hence are *isomorphic* under the previous theory. However, training two identical networks defined on $\{0,1\}^n$ and $[0,1]^n$ yields dramatically different learning curve behavior in each case, even when given analogous target concepts and domain distributions (*e.g.*, uniform). Invariably, exponential convergence is observed in the finite $\{0,1\}$ case, and rational (near worst case) convergence is observed in the continuous $[0,1]$ case. Therefore, merely stating that the worst case theory is not representative of the "typical" learning curves encountered in practice is not very satisfying. An adequate explanation of empirical learning curve behavior must explain how the worst case results *are* typical in some cases, while not in others.

It turns out that this discrepancy can be resolved via a simple observation about the previous theory. A close inspection of the results of [HLW88] reveals that the analysis is *nonuniform* in training sample size t. In particular, the lower bound result (2) chooses a *different* domain distribution (and possibly a different target concept) for *each* training sample size t. This does not accurately reflect the situation encountered in practice where these are held *fixed* (for example, as experimentally investigated by [CT90, CT92]). This raises the obvious 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 the answer to these questions is *yes*.

---

[3] This is also known as "power law" convergence [HKST94].

1.3 Results

By carrying out 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 distribution free setting. Specifically, we address the worst case asymptotic form of learning curves:

Definition 1 (Worst case learning curve). We say that a concept class C has a $\Theta'(g(t))$ worst case learning curve, written $LC(C) = \Theta'(g(t))$, if

1. there *exists* a learner L that achieves $\mathrm{E}_{\mathrm{P}^t}\, err(L, c) = O(g(t))$ 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 $\mathrm{E}_{\mathrm{P}^t}\, err(L, c) = \Omega'(g(t))$.[4]

(*I.e.*, some learner achieves $O(g(t))$ worst case convergence, but every learner can be forced to have an expected error of at least $ag(t)$ for some domain distribution and target concept in C (for infinitely many training sample sizes t).)

Our first results establish a basic dichotomy between rational and exponential convergence in the distribution free setting. We first show (Proposition 3) that for a *finite* concept class C, any learner that guesses consistent concepts from C obtains an exponential learning curve, even in the worst case. Then we show (Proposition 4) that $LC(C) = e^{\Omega'(-t)}$ for any non-trivial concept class C (containing at least two non–mutually-exclusive concepts); which establishes overall that $LC(C) = e^{\Theta'(-t)}$ for any finite, non-trivial C.

However, exponential convergence cannot always be achieved in the distribution free setting. In fact, there *are* concept classes C for which $LC(C) = \Theta'(1/t)$. Specifically, this can be shown for any class that contains a *continuous chain*:

Definition 2 (Concept chains). A *concept chain* is a concept class which is totally ordered under inclusion (*i.e.*, for $c_1, c_2 \in C$, either $c_1 \subset c_2$, $c_1 \supset c_2$, or $c_1 = c_2$). A *continuous* concept chain is order-isomorphic to the reals (*i.e.*, C can be indexed $\{c_y : y \in [0, 1]\}$ so that $c_{y_1} \subset c_{y_2}$ for $y_1 < y_2$).

We show (Theorem 7) that $LC(C) = \Omega'(1/t)$ for any continuous concept chain C. This is accomplished by exhibiting a single domain distribution P that forces *any* learner to obtain rational convergence for a non-trivial proportion of the concepts in C. Notice this is a stronger result than (2): it is much easier to force bad behavior by choosing a different target concept and domain distribution for each training sample size t, than it is to show that bad behavior results even when these are held fixed.

[4] The notation $f(t) = \Omega'(g(t))$ means there exists a constant a such that $f(t) \geq ag(t)$ 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 the target concept" on arbitrarily large training samples.

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, infinite concept classes which contain a continuous chain always force rational convergence in the worst case. Of course, there is a significant gap between finite and continuous concept classes; for example, these results say nothing about what happens for *countably* infinite concept classes. This gap also leaves open the question of identifying the precise boundary between rational and exponential learning curves, and determining whether alternative (intermediate) forms of convergence are possible (*e.g.*, $\Theta'(1/t^n)$ or some other form).

It turns out that precise answers to these questions can be obtained for the special case of concept *chains*. Here, the precise boundary between rational and exponential learning curves is determined by the presence of a *dense* sub-chain C (where between any two concepts $c_1 \subset c_2$ in C there is a third concept $c_1 \subset c_3 \subset c_2$ in C). Specifically, we can adapt the arguments used for continuous chains to show $LC(C) = \Theta'(1/t)$ for a *somewhere* dense chain C (Theorem 8). Establishing that exponential convergence can always be obtained for *nowhere* dense ("scattered") chains however (Theorem 9), is quite hard. Here, we devise a special learning strategy CHOLC that achieves exponential convergence by attempting to identify the target concept after some finite number of training examples with non-zero probability (avoiding hypothesis sequences that slowly converge to the target concept without ever reaching it). These two results precisely characterize the boundary between rational and exponential worst case convergence for concept chains, and also show that no other form of worst case convergence is possible (for concept chains) in the distribution free setting.

Finally, 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*. Although this means the learning curves they observed must have been fundamentally exponential, we demonstrate how, at the *scale* of training sample sizes they considered, convergence can *appear* rational.

1.4 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 [HLW88] suggested that rational convergence was the only possible worst case form; however this is based on a nonuniform analysis. By pursuing a uniform analysis, we are able to distinguish the conditions under which rational and exponential worst case convergence take place, based solely on the structure of the concept class C and ignoring any special properties of the domain distribution P (contrary to common suggestion [SST91, HKST94]).

Most theoretical studies of learning curve behavior adopt a distribution *specific* model that assumes the domain distribution P is known *a priori*. Given these stronger assumptions, many researchers have shown that both rational and exponential learning curves are possible. For example, exponential convergence has been demonstrated in many distribution specific analyses of particular concept spaces [GM93, BL91, PS90, SSSD90, SST91], and rational convergence has been demonstrated for other spaces [OH91].[5] This paper shows how, in a general way, this dichotomy can still be revealed under much weaker assumptions.

We also draw a clean boundary between these two modes of convergence in terms of a simple structural property of concept classes, namely the presence of a dense sub-chain. As most distribution specific analyses address particular case studies, they do not provide a precise, *general* characterization of the concept space properties that permit or prohibit exponential convergence. (Some progress towards such a characterization has been obtained for *finite* concept spaces [HKST94].)

Interestingly, only two possible modes of worst case convergence appear possible in the distribution free setting: rational and exponential (proved for concept chains, but conjecture in general). This is quite unlike the distribution specific case where all intermediate forms of convergence are apparently possible [HKST94]. Clearly the distribution specific analyses give tighter characterizations of the actual learning curves one might observe in practice, but require more problem specific information [HKST94] — in fact more than is generally available in practice. The benefits of the distribution free theory is its wider range of applicability in practical situations. Perhaps the simple, intuitive results reported here will help in the development of a similar characterization for the distribution specific case.

2 Technical discussion

2.1 Finite concept classes

We first observe that it is fairly obvious that any consistent learner obtains exponential worst case convergence for any finite concept class.

Proposition 3 (Finite UB). *For any finite concept class C, any learner L that guesses consistent concepts from C obtains* $\mathrm{E}_{\mathrm{P}^t}\, err(L, c) = e^{O(-t)}$ *for every target concept c in C, regardless of the distribution* P.

Proof. Fix a target concept $c \in C$ and a 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 examples is at most $N(1 - p_0)^t$. Observing a domain object from *each* difference set implies that a consistent learner L will produce a hypothesis with zero error. Therefore, $\mathrm{E}_{\mathrm{P}^t}\, err(L, c) \leq N(1 - p_0)^t = e^{O(-t)}$. □

Furthermore, it is *impossible* to achieve better than exponential worst case convergence for any non-trivial concept class.

[5] Amari *et al.* [AFS92] also provide a general Bayesian analysis of "average case" learning curves, but demonstrate only rational forms.

Proposition 4 (Universal LB). *For* any *non-trivial concept class C, there is a domain distribution* P *that forces* any *learner L to obtain* $\mathrm{E}_{\mathrm{P}^t}\, err(L, c) = e^{\Omega'(-t)}$ *for some c in C.*

Proof. Any non-trivial C contains non–mutually-exclusive concepts $c_1, c_2 \in C$. Fix P so that $\mathrm{P}(c_1 \triangle c_2) = p$ for some $0 < p < 1$. Then, for *any* learner L we get

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

The last inequality holds since any hypothesis h produced by L must satisfy $d_{\mathrm{P}}(h, c_1) + d_{\mathrm{P}}(h, c_2) \geq d_{\mathrm{P}}(c_1, c_2) = p$ by the triangle inequality. Finally, an average expected error of at least $e^{\Omega(-t)}$ for every $t \geq t_0$ implies that L must obtain at least this expected error on one of c_1 or c_2 for infinitely many t. $\square$

It is interesting to see how these results compare to the non-uniform theory of [HLW88]. Although we obtain exponential learning curves for *any* fixed distribution, there is no single "worst case" domain distribution here. That is, we obtain a *different* worst case domain distribution for *each* training sample size t. Therefore, although each individual curve is exponential, any *universal* upper bound over all curves is rational. Figure 1 illustrates this discrepancy between the worst case bounds of [HLW88] which consider a different P for each t, and the experimental results of [CT90] which consider a single P for all t.

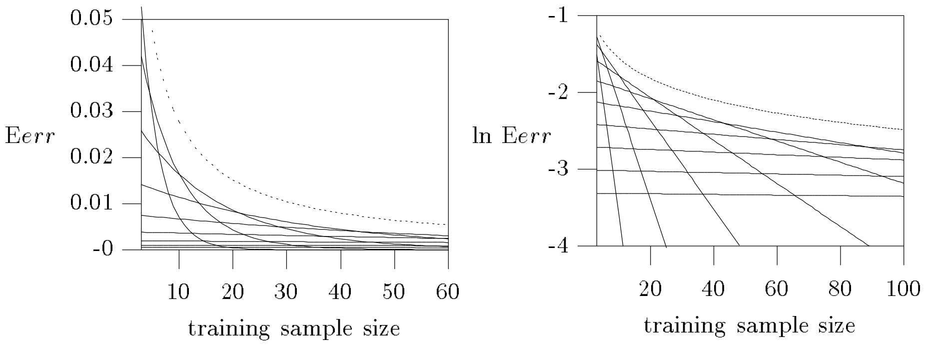

Fig. 1. Comparing uniform versus non-uniform bounds. This demonstrates how a series of exponential learning curves can have a rational upper envelope.

2.2 Continuous concept chains

However the situation is quite different for concept classes containing a continuous chain. Here it turns out rational convergence is the *best* that can be attained in the worst case. Moreover, there is a *single* domain distribution P that forces this worst case behavior for *any* learner L. Before going on to prove the lower bound, we first observe that, for simple chains, rational convergence can always be achieved by any consistent learner.

Proposition 5 (Chain UB). *For* any *chain C, any learner L that guesses consistent concepts from C obtains* $\mathrm{E}_{\mathrm{P}^t}\, err(L,c) = O(1/t)$ *for every target concept c in C, regardless of the domain distribution* P.

Proof. (Sketch) Since the VCdimension of any non-trivial chain is obviously 1, the special learning strategy 1IGPS obtains $\mathrm{E}_{\mathrm{P}^t}\, err(\text{1IGPS}, c) = O(1/t)$ by (1). Furthermore, *any* learner L that guesses consistent concepts from C obtains $\mathrm{E}_{\mathrm{P}^t}\, err(L,c) = O(\ln t/t)$ [HLW88]. Here, we strengthen these results slightly by showing that *any* consistent learner L actually obtains $\mathrm{E}_{\mathrm{P}^t}\, err(L,c) = O(1/t)$. This is done by noticing that the worst case situation is represented by the *uniform* chain (proof omitted):

Definition 6 (Uniform chain). A *uniform chain* is a concept space (C,P) where $C = \{c_y : y \in [0,1]\}$ is a continuous chain (Definition 2), and P is defined so that $d_{\mathrm{P}}(c_{y_1}, c_{y_2}) = |y_1 - y_2|$ for $y_1, y_2 \in [0,1]$. For a target concept $c_y \in C$ and a sequence of domain objects $\mathbf{x}^t$, we denote the *consistent segment* of C by $C[c_y\mathbf{x}^t]$, and the *diameter* of this set (under d_{P}) by $|C[c_y\mathbf{x}^t]|$.

Lemma 14 in Appendix A shows that $\mathrm{E}_{\mathrm{P}^t}|C[c\mathbf{x}^t]| < 2/(t+1)$ for any target concept c from a uniform chain (C,P), which obviously forces any consistent learner L to obtain $\mathrm{E}_{\mathrm{P}^t}\, err(L,c) \le \mathrm{E}_{\mathrm{P}^t}|C[c\mathbf{x}^t]| < 2/(t+1)$. □ (Proposition 5)

It remains to show that rational convergence is the *best* that can be achieved in this case.

Theorem 7 (Continuous LB). *For any continuous chain C, there is a domain distribution* P *that forces* any *learner L to obtain* $\mathrm{E}_{\mathrm{P}^t}\, err(L,c) = \Omega'(1/t)$ *for a non-trivial portion of the concepts c in C.*

Proof. (Sketch) Recall the definition of a continuous chain (Definition 2): $C = \{c_y : y \in [0,1]\}$ such that $c_{y_1} \subset c_{y_2}$ for $y_1 < y_2$. Fix a domain distribution P so that (C,P) forms a *uniform* chain (Definition 6).[6] In broad outline, the proof of this theorem follows the same strategy as Proposition 4: First, we argue that any learner must obtain a minimum expected error *on average* over the concepts in C for any fixed training sample size t, and use this to force a minimum expected error on infinitely many t for some $c \in C$. Specifically, the theorem is proved in 3 steps (Lemmas 14 to 16 in Appendix A): First we show (Lemma 14) that the expected diameter of the consistent segment of C shrinks no faster than $\Omega(1/t)$ for any $c \in C$. Next, we define a *"uniform" prior* on the concept class C, and use the first result to show that *any* learner L must obtain an expected error of at least $\Omega(1/t)$ *on average* over $c \in C$ (Lemma 15). Finally, we use these results to show that any learner L must obtain $\mathrm{E}_{\mathrm{P}^t}\, err(L,c) = \Omega'(1/t)$ for a non-trivial portion of the $c \in C$ (Lemma 16). □

[6] Such a measure can always be constructed by the same procedure used to construct the Lebesgue measure on $[0,1]$; see *e.g.*, [Ash72, Chapter 1].

Notice this is *stronger* than (2) as it shows how rational convergence can be forced by a *single* domain distribution and target concept (rather than choosing a different distribution and target concept for each training sample size).[7]

Overall, these results reveal the basic dichotomy between rational and exponential convergence in the distribution free setting. However, there are a wide range of concept classes not covered by this theory (*e.g.*, *countable* concept classes). We now draw a precise boundary between rational and exponential learning curves for the special case of *concept chains*. Specifically, we characterize the precise boundary between exponential and rational convergence in terms of the *density* of the chain.

2.3 Dense concept chains

First of all, for a (somewhere) dense chain it is easy to see that any consistent learner obtains at worst rational convergence, simply from Proposition 5. However, as with continuous chains, this is the *best* that can be achieved:

Theorem 8 (Dense LB). *For any dense chain C, there is a domain distribution* P *that forces* any *learner L to obtain* $\mathrm{E}_{\mathrm{P}^t} err(L,c) = \Omega'(1/t)$ *for a non-trivial portion of the concepts c in C.*

Proof. (Sketch) We saw in Theorem 7 that a *continuous* chain with a *uniform* distribution forces any learner L to obtain $\mathrm{E}_{\mathrm{P}^t}\, err(L,c) = \Omega'(1/t)$ for a non-trivial portion of the concepts in the chain. Here we generalize this result to arbitrary *dense* concept chains C. The main trick is to define a suitable domain distribution P that preserves the essential properties of a uniform chain. (Notice that we cannot simply construct a uniform chain directly since, in general, C could just be countably infinite.) Given a dense chain C, construct P as follows:

Construction: Stage 0: Choose arbitrary concepts $c_0, c_1 \in C$ such that $c_0 \subset c_1$. Stage 1: Choose a point $x_1 \in X$ between c_0, c_1 and assign it probability $1/2$. Stage 2: Choose a concept $c_2 \in C$ between c_0, x_1 and $c_3 \in C$ between x_1, c_1. Repeat this process *ad infinitum* for stages $i = 0, 1, 2, \ldots$
Odd Stage $k = 2i+1$: Choose a *point* $x_j \in X$ between each of the preceding elements (adjacent points or concepts) and assign it a probability $p_i = 4/8^{i+1}$.
Even Stage $k = 2i$: Choose a *concept* $c_j \in C$ between each of the preceding elements (adjacent points or concepts). Notice that the density of C ensures that each x_j and c_j can always be found; see Figure 2. Upon completion, each selected concept has neighbors at distances $\sum_{i=n}^{\infty} 4/8^{i+1} = 1/(14 \cdot 8^{n-1})$ for every $n \geq n_0$ for some $n_0 > 0$.

Given this distribution, we proceed as in Theorem 7: Lemma 14 can be adapted to show that the expected diameter of the consistent segment of C can shrink no faster than $\Omega(1/t)$ for any $c \in C$. Next, we define a natural "prior" on the

[7] Haussler *et al.* prove a similar result to Lemmas 14 and 15 in their later technical report [HLW90]; however, they do not supply Lemma 16. The argument presented here generalizes more readily to Theorem 8 below.

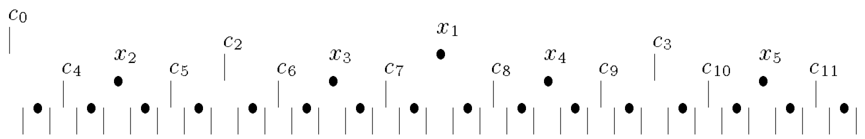

Fig. 2. Constructing a "dense" distribution for a dense concept chain.

concepts in C and use the first result to show that *any* learner L must obtain an expected error of at least $\Omega(1/t)$ *on average* over $c \in C$ (as in Lemma 15). Finally, we proceed exactly as in Lemma 16. (Details omitted.) $\Box$ (Theorem 8)

2.4 Scattered concept chains

Finally, we turn our attention to the complementary class of *scattered* (nowhere dense) chains. In contrast to the previous result establishing rational convergence for dense chains, the hope is show that *exponential* convergence can always be obtained for a scattered chain. First of all, Proposition 4 shows 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 for a scattered chain. (This turns out to be hard: we must actually *demonstrate* a guessing strategy that always obtains exponential convergence for any scattered chain, or at least prove that such a strategy exists.)

Theorem 9 (Scattered UB). *For any scattered chain C, there is a learning strategy CHOLC that obtains $\mathrm{E}_{\mathrm{P}^t}\, err(L, c) = e^{O(-t)}$ for every target concept c in C, regardless of the distribution* P.

Proof. (Intuitive sketch) In the finite case we saw that exponential convergence results whenever the target concept is identified (up to P-equivalence) with non-zero probability after a finite number of training examples. Here we attempt to apply the same idea to scattered concept chains.

It turns out that the only catch is dealing with *limit* concepts in this case; *i.e.*, concepts which are the limits of infinite ascending ($\bigcup_1^\infty c_i$) or descending ($\bigcap_1^\infty c_i$) sequences of $c_i \in C$. (Note that this can easily happen without the chain being *dense*; see Figure 3.) The problem is that guessing an infinite sequence of hypotheses that converges to, but never reaches the target concept c, can eas-

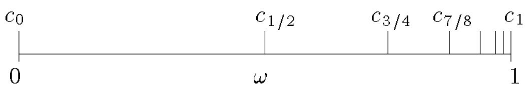

Fig. 3. A scattered concept chain defined on $[0, 1]$ with a limit concept c_1. (Endpoints shown: for concept c_y defined by endpoint y we have $c_y(x) = 1$ for all $x \le y$.) Notice that concept c_1 is the limit of concepts $c_{1-2^{-i}}$, but the chain remains nowhere dense (*i.e.*, there is no *set* of concepts for which between any two there is a third).

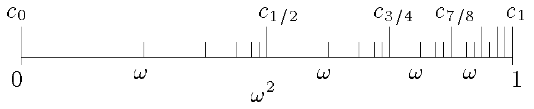

Fig. 4. A scattered concept chain defined on $[0, 1]$ with a second order limit concept. Notice that c_1 is a limit of concepts $c_{1-2^{-i}}$, *each* of which is itself a limit concept. Also notice that the chain is still nowhere dense.

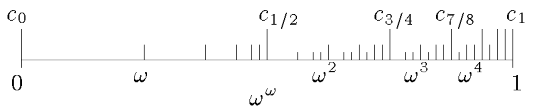

Fig. 5. A scattered chain of concepts defined on $[0, 1]$ with a limit concept of *infinite* order. Notice that c_1 is a limit of concepts $c_{1-2^{-i}}$, each of which is a limit concept of order i. Therefore, c_1 is a limit concept which can have no finite order. Notice that the chain remains nowhere dense (since for each concept there is a *least* larger concept).

ily yield rational convergence (*cf.* Theorem 8).[8] The obvious way to circumvent this difficulty is to always guess limit concepts before isolated concepts. Then, provided that the limit concepts are isolated from one another, we will always achieve exponential convergence, since any target concept is identified (up to P-equivalence) after just 2 training examples with non-zero probability. (Establishing that this must be the case in a scattered chain is the main part of the proof (Lemma 10 below).)

Notice first that it is actually possible to have concepts which are *limits* of limit concepts (*i.e.*, second order limit concepts) in a scattered chain; see Figure 4. In fact, limit concepts of each order $1, 2, 3, \ldots$ are certainly possible, and in general one can even have *limits* of such limit concepts (*i.e.*, concepts of infinite order!); see Figure 5. The following lemma shows that it is possible to have limit concepts of any *ordinal* order in a scattered chain, but regardless, all concepts of any particular order are *isolated* in the concepts of the same or higher order. (A proof of this fact follows as a corollary to Hausdorff's Theorem, which provides a suitably constructive characterization of the class of scattered linear orderings [Ros82, Chapter 5].)

Lemma 10. **(Corollary to Hausdorff's Theorem)** *For a* scattered *concept chain C, there is some* least *ordinal α for which:*
(i) *every limit concept in C has ordinal order $\beta < \alpha$,*
(ii) *C has limit concepts of every ordinal order $\beta < \alpha$,*
(iii) *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 chains which permits us to develop a learning strategy that successfully obtains exponential convergence. The following strategy *almost* works in general: **Strategy** HOLC: guess the highest order limit concept consistent with all the training examples. Intuitively, this strategy al-

[8] Note that guessing strategy 1IGPS does not automatically do this, and can actually produce rational learning curves when in fact exponential curves are possible.

ways yields exponential convergence since any concept is isolated from the class of concepts of the same or higher order, meaning that HOLC guesses a zero-error hypothesis with non-zero probability after just 2 training examples (eliminating both the nearest "left" and "right" same-order neighbors).[9]

However there is one final problem: HOLC requires that a consistent concept of maximal order always *exist* for any sequence of training examples. Unfortunately, this need not always be the case in general (*e.g.*, consider removing the last concept c_1 in Figure 5), and therefore HOLC is not always well defined. This difficulty can be circumvented by first *compactifying* the chain in a natural way:[10]

Lemma 11. *If C is a scattered chain, the the chain $\mathcal{C}(C)$ formed by closing C under limits (countable unions and intersections) is still scattered.*[11]

Lemma 12. *For a* compact *scattered chain C, there is a* unique *concept c' in C of* maximal *order that is consistent with any sequence of training examples.*

This leads to the final proposal for a guessing strategy to obtains exponential convergence for scattered chains: **Strategy** CHOLC: First compactify the chain C to obtain $\mathcal{C}(C)$, then guess the highest order limit concept in $\mathcal{C}(C)$ consistent with all of the training examples. CHOLC always obtains exponential convergence for any scattered concept chain. □ (Theorem 9)

Obviously the procedure CHOLC has little practical impact, but the issues it addresses shed light on the fundamental nature of worst case learning curves.

2.5 Scaling effects

Although the previous results draw a precise boundary between exponential and rational worst case learning curves (for concept chains), in a fundamental sense they miss the point demonstrated by the experimental results of [CT90, CT92]: since their experiments were computer simulations conducted with finite precision, *all* of the concept classes Cohn and Tesauro considered were fundamentally *finite*. The fact that they observe rational convergence in some cases seems to directly contradict the theoretical results presented here. However, the real source of the dichotomy between rational and exponential convergence in these

[9] Note that the effect of fixing a domain distribution is always to preserve the ordering structure of the chain, or to collapse segments of the ordering by identifying adjacent concepts. Collapsing segments of the chain cannot produce new limit concepts or increase the order of old limit concepts (beyond identifying them with already existing such concepts). So, for example, a nowhere dense chain cannot be made somewhere dense by *removing* concepts. Thus, the result holds for all domain distributions.

[10] There are good reasons for calling this a compactification: the resulting chain satisfies natural versions of the Bolzano-Weierstrass and Heine-Borel properties, as well as being complete-and-bounded, *et cetera*.

[11] Note that measurability is not a problem here so long as C was measurable to begin with.

computer simulations has to do with a *scaling* effect: all of the learning curves obtained by Cohn and Tesauro *are* fundamentally exponential, its just that at the *scale* of training sample sizes they considered (relative to the distances between neighboring concepts) convergence *appears* rational. This is easily demonstrated by a simple example.

Example: Consider a finite chain of concepts C_n consisting of $n+1$ concepts $c_0 \subset c_1 \subset ... \subset c_n$, and assume a "uniform" domain distribution P_n that imposes a distance of $1/n$ between adjacent concepts.

Proposition 13. *For* (C_n, P_n) *and for any target* c *in* C_n

$$\left(1-\tfrac{1}{n}\right)^t \tfrac{1}{2(t+1)} \quad < \quad \mathrm{E}_{\mathrm{P}^t}|C[c\mathbf{x}^t]| \quad < \quad \left(1-\tfrac{1}{n}\right)^t \tfrac{2}{t+1}.$$

These bounds are clearly exponential in t, and for large n are well approximated by $e^{-t/n}1/(2(t+1))$ and $e^{-t/n}2/(t+1)$, respectively. Now if we focus on training sample sizes t that are small relative to n, then the factor $e^{-t/n}$ will behave like a constant near 1 and the $1/(t+1)$ factor will dominate. So here we could wind up observing what would appear to be rational convergence, even when convergence is asymptotically exponential. To reveal exponential convergence we would have to consider training sample sizes on the order of $t = n, 2n, 3n, ...$, which in the "continuous" case is on the order of "computer BIGNUM." Therefore, we would have to observe astronomical training sample sizes in order for the exponential behavior to be revealed. This explains the dichotomy obtained in [CT90], as they tested the *same* training sample sizes for concept classes with vastly different inter-concept distances; observing exponential convergence when the gaps were large, and apparently rational convergence when the gaps were small.

A Proofs of some lemmas

Lemma 14. *For any target concept* c_y *in the uniform chain* (C, P)

$$\mathrm{E}_{\mathrm{P}^t}|C[c_y\mathbf{x}^t]| \quad = \quad \tfrac{2-(1-y)^{t+1}-y^{t+1}}{t+1} \quad = \quad \Theta(1/t).$$

Proof. (Outline) Fix a target concept $c_y \in C$. Let random variables U and V measure the distance between c_y and the smallest and the largest consistent concepts in C respectively. Clearly $U|\{x \in c_y\} \sim \mathsf{uniform}(0, y)$ and $V|\{x \notin c_y\} \sim \mathsf{uniform}(0, 1-y)$. Given t training examples, these minimum distances become $\underline{U}_t = \min\{U_1, ..., U_t\}$ and $\underline{V}_t = \min\{V_1, ..., V_t\}$, giving $C[c_y\mathbf{x}^t] = \underline{U}_t(\mathbf{x}^t) + \underline{V}_t(\mathbf{x}^t)$. In general, we have $\mathrm{E}\underline{R}_t = r/(t+1)$ for any random variable $R \sim \mathsf{uniform}(0, r)$ (see *e.g.*, [LM81, pp.99]). Thus we get

$$\mathrm{E}_{\mathrm{P}^t}|C[c_y\mathbf{x}^t]| = \sum_{i=0}^{t} \tbinom{t}{i} y^i (1-y)^{t-i} \big[\, \mathrm{E}\underline{U}_i|\{x_1, .., x_i \in c_y\} + \underline{V}_{t-i}|\{x_{i+1}, .., x_t \notin c_y\}\,\big]$$

$$= \sum_{i=0}^{t} \tbinom{t}{i} y^i (1-y)^{t-i} \left[\tfrac{y}{i+1} + \tfrac{1-y}{t-i+1}\right].$$

This can be simplified via the Binomial Theorem (see *e.g.*, [Bru77, Chapter 4]) to obtain the stated result. □

Lemma 15. *For the uniform prior* P_C *on* C*, any learner* L *obtains*

$$\mathrm{E}_{\mathrm{P}_C}\,\mathrm{E}_{\mathrm{P}_X^t}\,err(L,c) \geq \tfrac{1}{3(t+2)}.$$

Proof. (Outline) Fix an arbitrary learner L. An application of Fubini's Theorem and the definition of P_C yields

$$\mathrm{E}_{\mathrm{P}_C}\,\mathrm{E}_{\mathrm{P}_X^t}\,err(L,c) \;=\; \mathrm{E}_{\mathrm{P}_X^t}\,\mathrm{E}_{\mathrm{P}_C}\,err(L,c) \;=\; \mathrm{E}_{\mathrm{P}_X^t}\int_{[0,1]} err(L,c_y)\,dy.$$

Now notice that each fixed $\mathbf{x}^t \in X^t$ partitions the chain C into (at most) $t+1$ subintervals $C_{[0,y_1]}, C_{(y_1,y_2]}, \ldots, C_{(y_t,1]}$ where the concepts in each subinterval $c_y \in C_{(y_i,y_{i+1}]}$ identically label all training objects $x_j \in \mathbf{x}^t$. Therefore,

$$\mathrm{E}_{\mathrm{P}_X^t}\int_{[0,1]} err(L,c_y)\,dy \;=\; \mathrm{E}_{\mathrm{P}_X^t}\sum_{i=0}^{t}\int_{(y_i,y_{i+1}]} err(L,c_y)\,dy.$$

Now for any subinterval $C_{(y_i,y_{i+1}]}$ it can be shown (via the triangle inequality) that the average error (under d_{P_X}) of any hypothesis h can be lower bounded by a fixed fraction of the subinterval width, which leads to

$$\int_{(y_i,y_{i+1}]} err(L,c_y)\,dy \;\geq\; \int_{(y_i,y_{i+1}]} \tfrac{|C_{(y_i,y_{i+1}]}|}{6}\,dy \;=\; \int_{(y_i,y_{i+1}]} \tfrac{|C[c_y\mathbf{x}^t]|}{6}\,dy.$$

So we get

$$\mathrm{E}_{\mathrm{P}_X^t}\sum_{i=0}^{t}\int_{(y_i,y_{i+1}]} err(L,c_y)dy \;\geq\; \mathrm{E}_{\mathrm{P}_X^t}\sum_{i=0}^{t}\int_{(y_i,y_{i+1}]} \tfrac{|C[c_y\mathbf{x}^t]|}{6}dy \;=\; \mathrm{E}_{\mathrm{P}_X^t}\int_{[0,1]} \tfrac{|C[c_y\mathbf{x}^t]|}{6}dy$$

$$=\; \int_{[0,1]} \mathrm{E}_{\mathrm{P}_X^t}\tfrac{|C[c_y\mathbf{x}^t]|}{6}\,dy \;=\; \int_{[0,1]} \tfrac{2-(1-y)^{t+1}-y^{t+1}}{6(t+1)}\,dy \;=\; \tfrac{1}{3(t+2)}$$

by applying Fubini's Theorem and Lemma 14.[12] □

Lemma 16. *For the uniform prior* P_C *on* C *and any constant* $\alpha > 0$*, any learner* L *obtains* $\mathrm{P}_C\left\{c \in C \;:\; \mathrm{E}_{\mathrm{P}_X^t}\,err(L,c) \geq \tfrac{1-\alpha}{3(t+2)}\ i.o.\ t\right\} > 0$.

Proof. (Outline) Fix an arbitrary learner L and a constant $\alpha > 0$. Let $B_t^\alpha = \left\{c \in C : \mathrm{E}_{\mathrm{P}_X^t}\,err(L,c) \geq \tfrac{1-\alpha}{3(t+2)}\right\}$, and notice that $\overline{\lim}\, B_t^\alpha = \bigcap_{n=1}^{\infty}\bigcup_{t=n}^{\infty} B_t^\alpha = \left\{c \in C : \mathrm{E}_{\mathrm{P}_X^t}\,err(L,c) \geq \tfrac{1-\alpha}{3(t+2)} i.o.\, t\right\}$.

First of all, Lemma 15 can be used to show $\mathrm{P}_C B_t^\alpha \geq \alpha/(\alpha+11)$ for all t: To do this we note that for any learner L there is a learner L' that achieves $\mathrm{E}_{\mathrm{P}_X^t}\,err(L',c) \leq \min\{\,\mathrm{E}_{\mathrm{P}_X^t}\,err(L,c)\,,\, \mathrm{E}_{\mathrm{P}_X^t}\,|C[c\mathbf{x}^t]|\,\}$. Then it is easy to show for any random variable R with $0 \leq R \leq 2/(t+1)$ and $\mathrm{E}R \geq 1/(3(t+2))$ that $\mathrm{P}\,[\,R > (1-\alpha)/(3(t+2))\,] \;>\; \alpha/(\alpha+11)$ for any $\alpha > 0$.

Finally, we use this fact to show that $\mathrm{P}_C\overline{\lim}\, B_t^\alpha \geq \alpha/(\alpha+11)$: Let $D_T^\alpha = \bigcap_{n=1}^{T}\bigcup_{t=n}^{\infty} B_t^\alpha$. Clearly, $D_T^\alpha \downarrow \overline{\lim}\, B_t^\alpha$, and so $\mathrm{P}_C D_T^\alpha \downarrow \mathrm{P}_C\overline{\lim}\, B_t^\alpha$. Now notice $B_T^\alpha \subseteq D_T^\alpha$ and therefore $\mathrm{P}_C D_T^\alpha \geq \mathrm{P}_C B_T^\alpha \geq \alpha/(\alpha+11)$ for all T, which gives the desired result. □

[12] Haussler *et al.* prove a similar result in their later technical report [HLW90], but they use a different argument than the one presented here.

References

[AFS92] S. Amari, N. Fujita, and S. Shinomoto. Four types of learning curves. *Neural Computation*, 4:605–618, 1992.

[Ash72] R. B. Ash. *Real Analysis and Probability.* Academic Press, San Diego, 1972.

[BL91] E. B. Baum and Y.-D. Lyuu. The transition to perfect generalization in perceptrons. *Neural Computation*, 3:386–401, 1991.

[Bru77] R. A. Brualdi. *Introductory Combinatorics.* North-Holland, New York, 1977.

[CT90] D. Cohn and G. Tesauro. Can neural networks do better than the Vapnik-Chervonenkis bounds? In D. Touretzky, editor, *Advances in Neural Information Processing Systems 3*, pages 911–917. Morgan Kaufmann, San Mateo, CA, 1990.

[CT92] D. Cohn and G. Tesauro. How tight are the Vapnik-Chervonenkis bounds? *Neural Computation*, 4:249–269, 1992.

[GM93] 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 (COLT-93)*, pages 151–157, 1993.

[HKST94] D. Haussler, M. Kearns, H. S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. In *Proceedings of the Seventh Annual Workshop on Computational Learning Theory (COLT-94)*, pages 76–87, 1994.

[HLW88] 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 (COLT-88)*, pages 280–296, 1988.

[HLW90] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting {0,1}-functions on randomly drawn points. Technical Report UCSC-CRL-90-54, Computer Research Laboratory, University of California at Santa Cruz, 1990.

[LM81] R. J. Larsen and M. L. Marx. *An Introduction to Mathematical Statistics and its Applications.* Prentice-Hall, Englewood Cliffs, NJ, 1981.

[OH91] M. Opper and D. Haussler. Generalization performance of Bayes optimal classification algorithm for learning a Perceptron. *Physical Review Letters*, 66(20):2677–2680, 1991.

[PS90] M. J. Pazzani and W. Sarrett. Average case analysis of conjunctive learning algorithms. In *Proceedings of Seventh International Conference on Machine Learning (ML-90)*, pages 339–347, 1990.

[Ros82] J. G. Rosenstein. *Linear Orderings.* Academic Press, New York, 1982.

[SSSD90] D. B. Schwartz, V. K. Samalam, S. A. Solla, and J. S. Denker. Exhaustive learning. *Neural Computation*, 2:374–385, 1990.

[SST91] H. S. Seung, H. Sompolinsky, and N. Tishby. Learning curves in large neural networks. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory (COLT-91)*, pages 112–127, 1991.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.