Learning from positive data

Stephen Muggleton
Oxford University Computing Laboratory,
Parks Road,
Oxford, OX1 3QD,
United Kingdom.

Abstract

Gold showed in 1967 that not even regular grammars can be exactly identified from positive examples alone. Since it is known that children learn natural grammars almost exclusively from positives examples, Gold's result has been used as a theoretical support for Chomsky's theory of innate human linguistic abilities. In this paper new results are presented which show that within a Bayesian framework not only grammars, but also logic programs are learnable with arbitrarily low expected error from positive examples only. In addition, we show that the upper bound for expected error of a learner which maximises the Bayes' posterior probability when learning from positive examples is within a small additive term of one which does the same from a mixture of positive and negative examples. An Inductive Logic Programming implementation is described which avoids the pitfalls of greedy search by global optimisation of this function during the local construction of individual clauses of the hypothesis. Results of testing this implementation on artificially-generated data-sets are reported. These results are in agreement with the theoretical predictions.

1 Introduction

Gold's [5] seminal paper not only formed the foundations of learnability theory but also provided an important negative result for the learnability of grammars. It was shown that not even the class of all regular languages could be *identified in the limit* from an arbitrary finite sequence of positive examples of the target language. In the same paper Gold pointed out the implications for theories of language acquisition in children. He notes that psycholinguistic studies by McNeill and others had shown that

> ... children are rarely informed when they make grammatical errors and those that are informed take little heed.

Gold's negative results have been taken by [14] as theoretical support for Chomsky's theory [4] of innate human linguistic abilities.

In this paper Gold's requirements for exact identification of a language are replaced by a need to converge with arbitrarily low error. In a previous paper [10] the author derived a function for learning logic programs from positive examples only. In the present paper the Bayes' function for maximising posterior probability is derived. The solution is representation independent, and therefore equally applicable to grammar learning, scientific theory formation or even automatic programming. The expected error of an algorithm which maximises this function over a high prior probability segment of the hypothesis space is analysed and shown to be within a small additive term of that obtained from a mixture of positive and negative examples.

An implementation of this approach within the Inductive Logic Programming (ILP) system Progol4.2 is described. A novel aspect of this implementation is the use of global optimisation during local construction of individual clauses of the hypothesis. The technique avoids the local optimisation pitfalls of cover-set algorithms. Experiments on three separate domains (animal taxonomy, KRK-illegal and grammar learning) are shown to be in accordance with the theoretical predictions.

This paper is organised as follows. In Section 2 a Bayes' framework is described which is compatible with the U-learnability framework [9, 13]. The Bayes' function for the posterior probability of hypotheses given positive examples only is derived in Section 3. The expected error of an algorithm which maximises the Bayes' function over a high prior probability segment of the hypothesis space is given in Section 4. In Section 5 the ILP system Progol4.2, which implements this function is described. An experiment is presented in Section 6, in which Progol4.2 is tested on varying amounts of randomly generated data for three target concepts. The results of these experiments are discussed in Section 7. The paper is concluded in Section 8 by a comparison to related work and a discussion of directions for future research.

2 Bayes' positive example framework

The following is a simplified version of the U-learnability framework presented in [9, 13]. X is taken to be a countable class of instances and $\mathcal{H} \subseteq 2^X$ to be a countable class of concepts. D_X and $D_{\mathcal{H}}$ are probability distributions over X and $\mathcal{H}$ respectively. For $H \in \mathcal{H}$, $D_X(H) = \sum_{x \in H} D_X(x)$ and the conditional distribution of D_X associated with H is as follows.

$$D_{X|H}(x) = D_X(x|H) = \frac{D_X(x \cap H)}{D_X(H)} = \begin{cases} 0 & \text{if } x \notin H \\ \frac{D_X(x)}{D_X(H)} & \text{otherwise} \end{cases}$$

The teacher randomly chooses a target theory T from $D_{\mathcal{H}}$ and randomly and independently chooses a series of examples $E = \langle x_1, .., x_m \rangle$ from T according to $D_{X|T}$. Given E, $D_{\mathcal{H}}$ and D_X a learner L chooses an hypothesis $H \in \mathcal{H}$ for which all elements of E are in H. The teacher then assesses Error(H) as $D_X(H \setminus T) + D_X(T \setminus H)$.

Unlike the setting in U-learnability it is assumed in the present paper that L is given $D_{\mathcal{H}}$ and D_X.

3 Bayes' posterior estimation

Gold's negative result for identification of the regular languages over the symbol set Σ is based on the fact that for any sequence of positive examples E there will always be at least two possible candidate hypotheses, 1) Σ^*, the language containing all possible sentences and 2) the language corresponding to elements of E alone. It is clear that 1) is the most general hypothesis, and has a compact finite automaton description, while 2) is the least general hypothesis and has a complex finite state automaton description. Since these two extreme hypotheses will maximise errors of commission and omission respectively it would seem desirable to find a compromise between the size of the hypothesis description and its generality. Size and generality of an hypothesis can be defined within the Bayes' framework of the previous section as follows.

$$\begin{aligned} \mathrm{sz}(H) &= -\ln D_{\mathcal{H}}(H) \\ \mathrm{g}(H) &= D_X(H) \end{aligned}$$

Bayes' theorem allows us to derive a tradeoff between sz(H) and g(H). In its familiar form, Bayes' theorem is as follows.

$$p(H|E) = \frac{p(H)p(E|H)}{p(E)}$$

With respect to the Bayes' framework of the previous section $p(H|E)$ is interpreted from the learner's perspective as the probability that $H = T$ given the example sequence is E. Similarly, $p(H)$ is defined as the probability that $H = T$, which is

$$p(H) = D_{\mathcal{H}}(H).$$

Meanwhile $p(E|H)$ is the probability that the example sequence is E given that $H = T$. Since examples are chosen randomly and independently from $D_{X|H}$ then for any consistent hypothesis this is as follows.

$$\begin{aligned} p(E|H) &= \prod_{i=1}^{m} D_{X|H}(x_i) \\ &= \prod_{i=1}^{m} \frac{D_X(x_i)}{D_X(H)} \end{aligned}$$

The prior $p(E)$ is the probability that the example sequence is E irrespective of T. This is as follows.

$$p(E) = \sum_{T \in \mathcal{H}} D_{\mathcal{H}}(T) \prod_{j=1}^{m} D_{X|T}(x_j)$$

The Bayes' equation can now be rewritten as follows.

$$p(H|E) = \frac{D_{\mathcal{H}}(H) \prod_{i=1}^{m} \frac{D_X(x_i)}{D_X(H)}}{p(E)}$$

Since $\frac{\prod_{i=1}^{m} D_X(x_i)}{p(E)}$ is common to all consistent hypotheses, it will be treated as a normalising constant c_m in the following.

$$\begin{aligned} p(H|E) &= D_{\mathcal{H}}(H) \left(\frac{1}{D_X(H)} \right)^m c_m \\ \ln p(H|E) &= m \ln \left(\frac{1}{g(H)} \right) - sz(H) + d_m \end{aligned}$$

In the above $d_m = \ln c_m$. The tradeoff between size and generality of an hypothesis can be seen in the final equation above. The function $\ln p(H|E)$ decreases with increases in $sz(H)$ and $g(H)$. Additionally, as m grows, the requirements on generality of an hypothesis become stricter. A function f_m with similar properties was defined in [10] and it was shown there that for every hypothesis H except T there is a value of m such that for all $m' > m$ it is the case that $f_{m'}(H) < f_m(T)$. This result indicates a form of convergence, somewhat different from Gold's identification in the limit.

4 Analysis of expected error

Haussler et al. [7] argue the advantages of analysing expected error over VC dimension analysis. Analysis of expected error is the approach taken below.

It is assumed that class membership of instances is decidable for all hypotheses. Also the hypotheses in $\mathcal{H}$ are assumed to be ordered according to decreasing prior probability as $H_1, H_2, \ldots$. For the purposes of analysis the distribution $D_{\mathcal{H}}(H_i) = \frac{a}{i^2}$ is assumed, where a is a normalising constant. This is similar to the prior distribution assumptions used in Progol4.1 [10] and is a smoothed version of a distribution which assigns equal probability to the 2^b hypotheses describable in b bits, where the sum of the probabilities of such hypotheses is 2^{-b}. Within this distribution i has infinite mean and variance. It is also assumed that the hypothesis space contains only targets T for which $D_X(T) \leq \frac{1}{2}$. This assumption, which holds for most target concepts used in Inductive Logic Programming, is not a particularly strong restriction on the hypothesis space since if $\overline{T}$ is the complement of T and $D_X(T) > \frac{1}{2}$ then clearly $D_X(\overline{T}) \leq \frac{1}{2}$.

The following theorem gives an upper bound on the expected error of an algorithm which learns from positive examples only by maximising the Bayes' posterior probability function over the initial am hypotheses within the space.

Theorem 1 Expected error for positive examples only. *Let X be a countable instance space and D_X be a probability distribution over X. Let $\mathcal{H} \subseteq 2^X$ be a countable hypothesis space containing at least all finite subsets of X, and for which all $H \in \mathcal{H}$ have $D_X(H) \leq \frac{1}{2}$. Let $D_{\mathcal{H}}$ be a probability distribution over $\mathcal{H}$. Assume that $\mathcal{H}$ has an ordering $H_1, H_2, \ldots$ such that $D_{\mathcal{H}}(H_i) \geq D_{\mathcal{H}}(H_j)$ for all $j > i$. Let $D_{\mathcal{H}}(H_i) = \frac{a}{i^2}$ where $\frac{1}{a} = \sum_{i=1}^{\infty} \frac{1}{i^2} \approx \frac{1}{0.608}$. Let $\mathcal{H}_n = \{H_i : H_i \in \mathcal{H} \text{ and } i \leq n\}$. T is chosen randomly from $D_{\mathcal{H}}$ and the x_i in $E = \langle x_1, .., x_m \rangle$ are chosen randomly and independently from $D_{X|T}$. Let $f_m(H) = D_{\mathcal{H}}(H)(\frac{1}{D_X(H)})^m$ and let $n = am$. L is the following learning algorithm. If there are no hypotheses $H \in \mathcal{H}_n$ such that $H \supseteq H_E = \{x_1, .., x_m\}$ then $L(E) = H_E$. Otherwise $L(E) = H_n(E) = H$ only if $H \in \mathcal{H}_n$, $H \supseteq H_E$ and for all $H' \in \mathcal{H}_n$ for which $H' \supseteq H_E$ it is the case that $f_m(H) \geq f_m(H')$. The error of an hypothesis H is defined as $Error(H, T) = D_X(T \setminus H) + D_X(H \setminus T)$. The expected error of L after m examples, $EE(m)$, is at most $\frac{2.33+2\ln m}{m}$.*
Proof. *Given in Appendix A.* □

Note that this result is independent of the choice of D_X and that L considers only $O(m)$ hypotheses to achieve an expected error of $O(\frac{\ln m}{m})$. For comparison a similar algorithm which learns from a mixture of positive and negative examples is analysed for the same choice of $D_{\mathcal{H}}$.

Theorem 2 Expected error for positive and negative examples. *Let X be a countable instance space and $\mathcal{H} \subseteq 2^X$ be a countable hypothesis space containing at least all finite subsets of X. Let $D_{\mathcal{H}}$, D_X be probability distributions over $\mathcal{H}$ and X. Assume that $\mathcal{H}$ has an ordering $H_1, H_2, \ldots$ such that $D_{\mathcal{H}}(H_i) \geq D_{\mathcal{H}}(H_j)$ for all $j > i$. Let $D_{\mathcal{H}}(H_i) = \frac{a}{i^2}$ where $\frac{1}{a} = \sum_{i=1}^{\infty} \frac{1}{i^2}$. Let $\mathcal{H}_n = \{H_i : H_i \in \mathcal{H} \text{ and } i \leq n\}$. T is chosen randomly from $D_{\mathcal{H}}$. Let $ex(x, H) = \langle x, v \rangle$ where $v = True$ if $x \in H$ and $v = False$ otherwise. Let $E = \langle ex(x_1, T), .., ex(x_m, T) \rangle$ where each x_i is chosen randomly and independently from D_X. $H_E = \{x : \langle x, True \rangle \text{ in } E\}$. Hypothesis H is said to be consistent with E if and only if $x_i \in H$ for each $\langle x_i, True \rangle$ in E and $x_j \notin H$ for each $\langle x_j, False \rangle$ in E. Let $f(H) = D_{\mathcal{H}}(H)$ and let $n = am$. L is the following learning algorithm. If there are no hypotheses $H \in \mathcal{H}_n$ consistent with E then $L(E) = H_E$. Otherwise $L(E) = H_n(E) = H$ only if $H \in \mathcal{H}_n$, H consistent with E and for all $H' \in \mathcal{H}_n$ consistent with E it is the case that $f(H) \geq f(H')$. The error of an hypothesis H is defined as $Error(H, T) = D_X(T \setminus H) + D_X(H \setminus T)$. The expected error of L after m examples, $EE(m)$, is at most $\frac{1.51+2\ln m}{m}$.*
Proof. *Given in Appendix A.* □

Note that this is within a small additive term of the bound for learning from

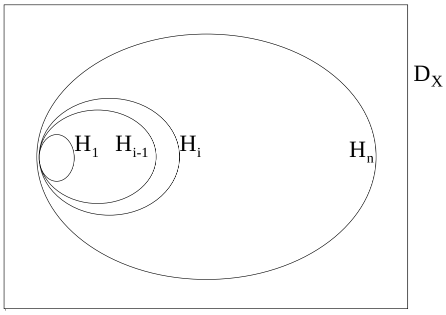

Figure 1: Generality of partial theories

positive examples only. Again the result is independent of the choice of D_X and again L considers only $O(m)$ hypotheses to achieve an expected error of $O(\frac{\ln m}{m})$.

5 Implementation

The Bayes' function f_m has been implemented to guide the search of the ILP system Progol [10] when learning from positive examples only. The new version, Progol4.2, is available by anonymous ftp from ftp.comlab.ox.ac.uk in directory pub/Packages/ILP/progol4.2. The earlier version, Progol4.1, uses a cover-set algorithm to construct the set of clauses, but for each clause does a pruned admissible search to maximise compression. Progol4.2 has a similar overall search algorithm, but when constructing each clause carries out an admissible search which optimises a global estimate of f_m for the complete theory containing the clause under construction. The basis for this global estimate is as follows. Suppose a clause C_i has been constructed as the ith clause of an overall theory (set of clauses) $H_n = \{C_1, .., C_n\}$. It is found that $H_i = C_1, .., C_i$ implies p more of the m positive examples than H_{i-1}. Figure 1 shows this situation with respect to the sample distribution D_X. According to the Law of Large Numbers when m is large

$$\frac{\mathrm{g}(H_i) - \mathrm{g}(H_{i-1})}{\mathrm{g}(H_n)} \approx \frac{p}{m}$$

and therefore

$$\mathrm{g}(H_n) \approx \frac{m}{p}(\mathrm{g}(H_i) - \mathrm{g}(H_{i-1})).$$

The surprising conclusion is that for large m it is possible to estimate the generality of H_n from p, m, $\mathrm{g}(H_i)$ and $\mathrm{g}(H_{i-1})$.

By assuming that the size of an hypothesis can be measured in bits for any hypothesis and that the number of examples covered per bit of an hypothesis is approximately uniform the following should also hold.

$$\mathrm{sz}(H_n) \approx \frac{m}{p}\ \mathrm{sz}(C_i)$$

In Progol4.2 the value of $\mathrm{sz}(C_i)$ is measured crudely as the number of atoms in the clause.

Since it is possible to estimate both $\mathrm{sz}(H_n)$ and $\mathrm{g}(H_n)$ during the local construction of each individual clause, it is possible to maximise an estimate of $f_m(H_n)$ during the construction of each of the clauses. The polynomial time-complexity bounds on the search carried out by Progol4.1 [10] are unaltered for Progol4.2.

5.1 Estimation of $g(H_i)$

The function $g(H_i)$ in the above is estimated in Progol4.2 using Stochastic Logic Programs (SLPs) [11]. An SLP is a range-restricted logic program P with numeric labels associated with each of the clauses. An SLP can be used to randomly derive elements of the Herbrand Base of P using SLD derivation with a stochastic selection rule. In order to estimate $g(H_i)$ an SLP, representing D_X, is used to randomly generate a sample of size s from the domain of the predicate p/n being learned. If s' of these instances are entailed by p/n then the Laplace corrected estimate of $g(H_i)$ is $\frac{s'+1}{s+2}$.

In order to construct the SLP for the domain of p/n, Progol4.2 uses the *modeh* declaration of p/n (see [10]). For instance, suppose in a chess domain the mode declaration is modeh(1,move(+piece,pos(+file,+rank),pos(+file,+rank))). Then Progol4.2 will construct the following generating clause for the domain.

```
'*move'(A,pos(B,C),pos(D,E)) :- piece(A), file(B), rank(C),
                                file(D), rank(E).
```

The clauses of the SLP consist of the above clause and the definitions of *piece/1, file/1* and *rank/1*. The labels for the SLP are built up by recording the total number of times each clause is visited in the derivations of the positive examples of *move*/3. In this way it is possible to estimate the distribution D_X from the examples themselves. For instance, in the example set we might find that half the examples involve the queen, a quarter involve rooks and the other quarter involve bishops. When randomly generating examples from the conditioned SLP these proportions are maintained.

6 Experiment

6.1 Experimental hypotheses

There remains a question as to how appropriate the assumptions made in Theorems 1 and 2 are in practice. In this section the predictive accuracy of three reasonably typical ILP target theories are compared against the theoretical predictions. If the distributional assumptions made are correct then we would expect

Examples.

class(dog,mammal).	class(dolphin,mammal).
class(trout,fish).	class(eel,fish).
class(lizard,reptile).	class(snake,reptile).
class(eagle,bird).	class(penguin,bird).

Target.

class(A,mammal) :- has_milk(A).
class(A,fish) :- has_gills(A).
class(A,bird) :- has_covering(A,feathers).
class(A,reptile) :- has_covering(A,scales),
not has_gills(A).

Figure 2: Animal taxonomy

that the expected error functions of theorems 1 and 2 should bound the mean error of a sample of real world domains. The experiments described in this section will test the following two hypotheses.

1. **Upper bound.** In every domain $EE(m) \leq \frac{2.33+2\ln\ m}{m}$.
2. **Positive versus positive and negative data.** In every domain error is of a similar order when learning from positives examples only compared to learning from a mixture of positive and negative examples.

6.2 Materials

The experimental hypotheses will be tested using Progol4.2 on the following target theories.

Animal taxonomy. Figure 2 shows the target and form of examples for the animal taxonomy domain.

KRK illegality. Figure 3 shows the target and form of examples for the KRK illegality domain (originally described in [12]).

Natural language grammar. Figure 4 shows the target and form of examples for the natural language grammar domain.

Examples sets and background knowledge for the domains above are available from the ftp site described in Section 5.

Examples.

```
illegal(3,5,6,7,6,2).   illegal(3,6,7,6,7,4).
illegal(5,1,2,1,2,1).   illegal(4,3,1,1,4,2).
```

Target.

```
illegal(A,B,A,B,_,_).
illegal(_,_,A,B,A,B).
illegal(A,B,_,_,C,D) :- adj(A,C), adj(B,D).
illegal(A,_,B,_,B,_) :- not A=B.
illegal(_,A,_,B,_,B) :- not A=B.
illegal(_,A,B,C,B,D) :- A<C, A<D.
illegal(_,A,B,C,B,D) :- A>C, A>D.
illegal(A,_,B,C,D,C) :- A<B, A<D.
illegal(A,_,B,C,D,C) :- A>B, A>D.
```

Figure 3: KRK illegality

Examples.

```
s([every, nice, dog, barks], []).
s([the,man,hits,the,ball,at,the,house],[]).
s([the,dog,walks,to,the,man],[]).
```

Target.

```
s(A,B) :- np(A,C), iverb(C,B).
s(A,B) :- np(A,C), vp(C,D), np(D,B).
s(A,B) :- np(A,C), tverb(C,D), np(D,E),
          prep(E,F), np(F,B).
```

Figure 4: Natural language grammar

6.3 Method

For the first two domains instances were generated randomly using the appropriate SLP (see Section 5.1) with uniform values of labels on all clauses. In the grammar domain it was found that only around 4 in 10,000 randomly generated sentences were positive examples of the target grammar T. Thus the distribution D_X was skewed so that $D_X(T) = D_X(\overline{T}) = 0.5$. In all domains instances were classified according to the target theory in order to construct training and test sets. In the case of learning from positive examples only, training sets had all negative examples removed.

For each domain Progol4.2 was tested on 1) learning from positive examples only and 2) learning from a mixture of positive and negative examples. In both cases m was varied according to the series $m = 5, 10, 20, 40, 80, 160, 320, 640, 1280$. For each size of sample the predictive accuracy of the hypothesis returned by Progol4.2 was estimated on a test set of size 10,000. For each m the estimate of predictive accuracy was averaged over 10 resamplings of the same sized training set. The series was discontinued for a particular domain if the estimate error was 0 for several successive values of m.

7 Results

7.1 Predictive accuracy versus bound

The results of testing the first experimental hypothesis (expected error upper bound) are graphed in Figures 5, 6 and 7. Labels on these graphs have the following meanings.

P. The predictive accuracy of learning from positive examples only is shown as the mean and standard deviation (error bars) of the 10 retrials for each value of m.

L(P). The theoretical lower bound on positive examples only accuracy from Theorem 1 (Accuracy$= 100(1 - EE(m))$).

M. Majority class for domain $(100 D_X(T))$.

Since each data point in each of the three domains lies above the bound, the first experimental hypothesis of Section 6 is confirmed [1].

[1] The non-monotonic behaviour of **P** in Figure 5 was found to be caused by large fluctuations in errors of commission. This is due to the gradual allowance of larger theories by the Bayes' function as m grows, together with the fact that generality does not vary monotonically with increasing size of a clausal theory.

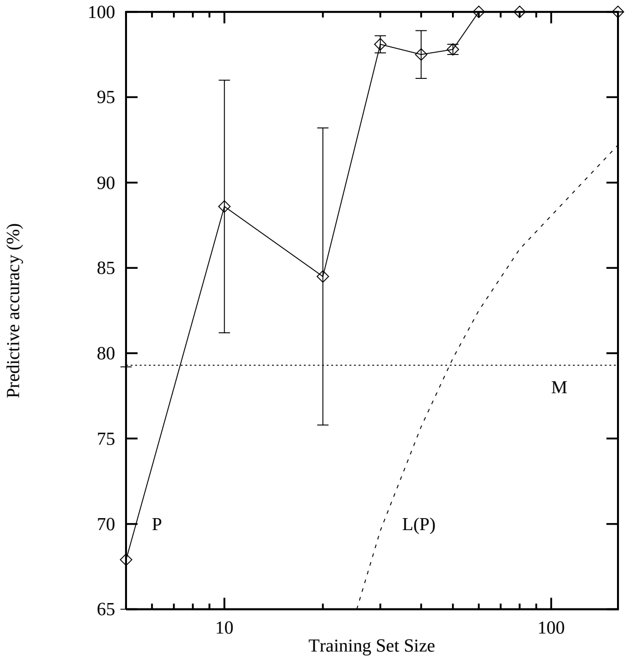

Figure 5: Predictive accuracy versus bound for animal taxonomy

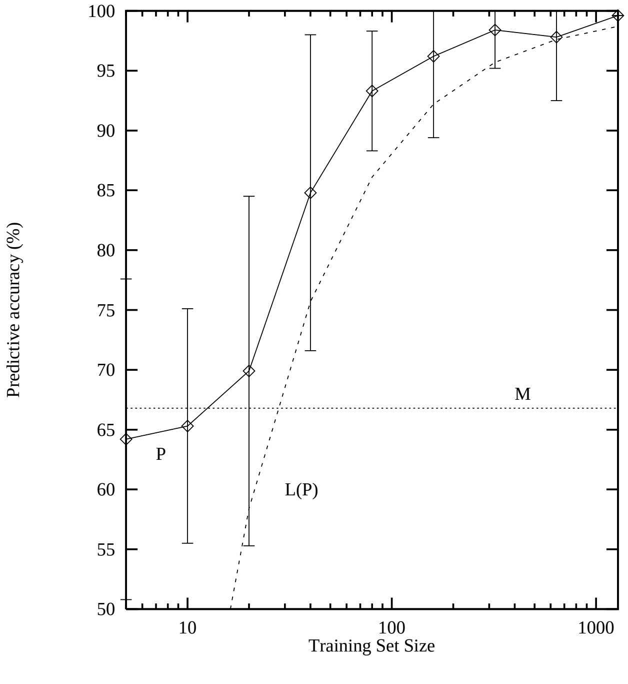

Figure 6: Predictive accuracy versus bound for KRK illegal

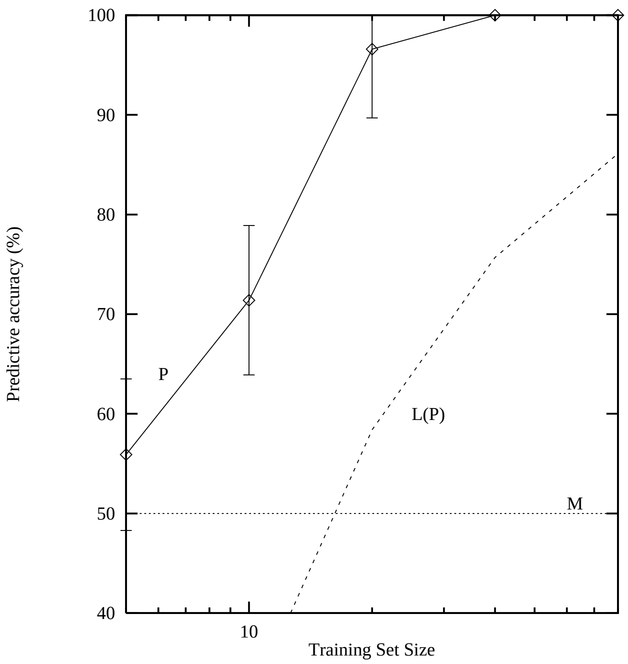

Figure 7: Predictive accuracy versus bound for natural language grammar

7.2 Positive versus positive and negative

The results of testing the second experimental hypothesis (similar expected error for positive versus positive and negative) are graphed in Figures 8, 9 and 10. Labels on these graphs have the following meanings.

P. The predictive accuracy of learning from positive examples only, shown as the mean of the 10 retrials for each value of m.

P+N. The predictive accuracy of learning from a mixture of positive and negative examples, shown as the mean of the 10 retrials for each value of m.

L(P+N). The theoretical lower bound on positive examples only accuracy from Theorem 2 (Accuracy$= 100(1 - EE(m))$).

M. Majority class for domain ($100D_X(T)$).

In the taxonomy and grammar domains (Figures 8 and 10) learning from positive examples only requires consistently fewer examples for any given ϵ than learning from a mixture of positive and negative examples. In the KRK-illegality domain the converse is true. In every domain accuracy for all values of m is comparable when learning from positive examples compared to learning from a mixture of positive and negative examples. This confirms the experimental hypothesis.

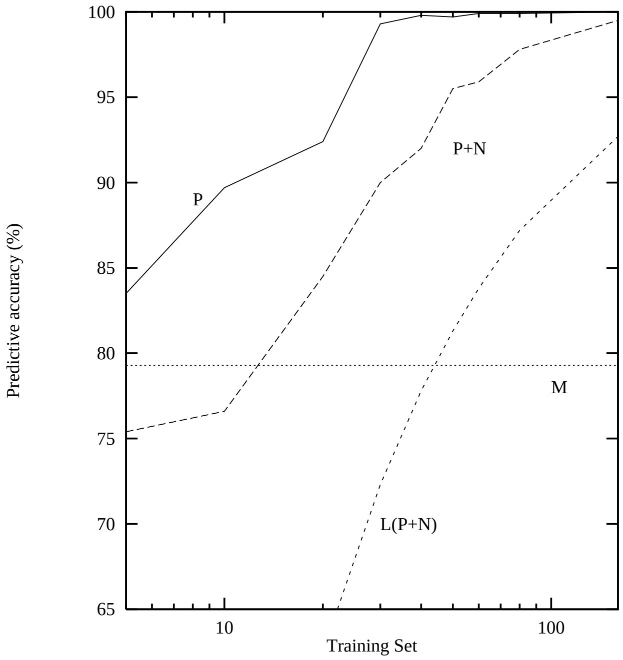

Figure 8: Positives versus positives and negatives for animal taxonomy

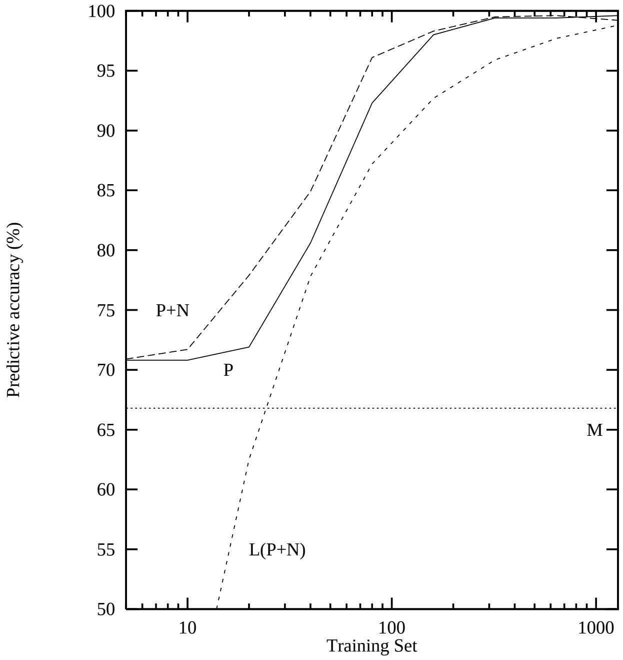

Figure 9: Positives versus positives and negatives for KRK illegal

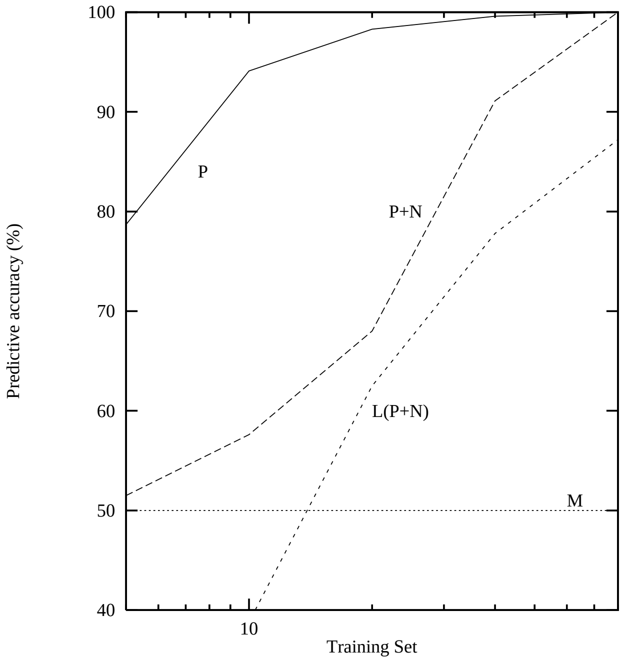

Figure 10: Positives versus positives and negatives for natural language grammar

8 Conclusion

In 1967 Gold demonstrated negative results for learnability in the limit of various classes of formal languages. This has provided a strong impetus for the investigation of constrained hypothesis languages, within which learning from positive examples is possible. For instance, Plotkin [15] demonstrated the existence of unique least general generalisations of positive examples represented as first-order clauses. Biermann and Feldman [2] and later Angluin [1] demonstrated that certain parameterised subsets of the regular languages could be identified in the limit from positive examples only. Within the framework of PAC-learning Valiant demonstrated [19] that k-CNF propositional logic formulae are learnable from positive examples. More recently Shinohara [18] demonstrated that certain size-bounded classes of elementary formal systems are identifiable in the limit from positive examples.

Unlike the approaches above, the techniques used in this paper for learning from positive examples are representation independent. That is to say, the representation of hypotheses does not play a part either in the development of the Bayes' function (Section 3) or the analysis of expected error (Section 4). It might legitimately be claimed that two strong assumptions are made in Section 2: 1) that the learner knows $D_{\mathcal{H}}$ and 2) that the learner can estimate D_X by conditioning a Stochastic Logic Program. The second assumption seems less pernicious since it only requires a logic program which defines the Herbrand base. Further theoretical research is required to analyse the effect of discrepancy between the learner's prior $p(H)$ and the teacher's distribution $D_{\mathcal{H}}$.

Theorems 1 and 2 define bounds on expected error for randomly chosen targets and example sets. To machine learning practitioners it might seem surprising that these bounds are a function of only m, the number of examples. However, the bounds are achieved by considering expectations over randomly chosen targets from a given distribution, $D_{\mathcal{H}} = \frac{a}{i^2}$. Experiments could be devised in which targets were randomly chosen from such a distribution, though this would tell us nothing about expected error in real domains. The experiments in Section 6 were devised to check whether the assumptions within Theorems 1 and 2 were appropriate for typical domains to which the ILP system Progol has been applied. Owing to limitations of time it was only feasible to test three such domains, for which the bounds held in all three cases. Further testing of domains is necessary to determine whether the bounds hold on average for existing Progol datasets.

Various researchers including [3, 6] have advocated and demonstrated the use of Bayesian analysis in machine learning. The success of the Bayesian solution to learning from positive examples reinforces this trend.

Several techniques [16, 17, 8] for learning from positive examples only have been investigated within Inductive Logic Programming. However, all these approaches differ from this paper in assuming some form of completeness within the example set.

In the light of the results in this paper it would seem worth reconsidering the degree of support that Gold's learnability results provide for Chomskian linguistics. Clearly, Chomsky's theory of innate linguistic ability is consistent with the results in this paper. However, the results in this paper show that weaker assumptions concerning the innate properties of natural language can be made than those that Gold's results might seem to suggest.

Acknowledgements

The author would like to thank David Haussler for influential discussions on the topic of learning from positive examples. The author's investigations with David Haussler of PAC upper-bound results for learning from positive examples will be reported elsewhere. Many thanks are due to my wife, Thirza Castello-Cortes, who has shown me great support during the writing of this paper. The author is grateful to Nick Chater of the Experimental Psychology Department in Oxford for pointing out the relevant literature on language learning in children. Thanks also for useful discussions on the topics in this paper with Donald Michie, John McCarthy, Tony Hoare, David Page, Ashwin Srinivasan and James Cussens. This work was supported partly by the Esprit Long Term Research Action ILP II (project 20237), EPSRC grant GR/J46623 on Experimental Application and Development of ILP, EPSRC grant GR/K57985 on Experiments with Distribution-based Machine Learning and an EPSRC Advanced Research Fellowship held by the author. The author is also supported by a Research Fellowship at Wolfson College Oxford.

References

[1] D. Angluin. Inference of reversible languages. *Journal of the ACM*, 29:741–765, 1982.

[2] A.W. Biermann and J.A. Feldman. On the synthesis of finite-state machines from samples of their behaviour. *IEEE Transactions on Computers*, C(21):592–597, 1972.

[3] W. Buntine. *A Theory of Learning Classification Rules.* PhD thesis, School of Computing Science, University of Technology, Sydney, 1990.

[4] N. Chomsky. *Knowledge of language: its nature, origin and use.* Praeger, New York, 1986. First published 1965.

[5] E.M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[6] D. Haussler, M Kearns, and R. Shapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. In *COLT-91: Proceedings of the 4th Annual Workshop on Computational Learning Theory*, pages 61–74, San Mateo, CA, 1991. Morgan Kauffmann.

[7] D. Haussler, M Kearns, and R. Shapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning Journal*, 14(1):83–113, 1994.

[8] R.J. Mooney and M.E. Califf. Induction of first-order decision lists: Results on learning the past tense of english verbs. *Journal of Artificial Intelligence Research*, 3:1–24, 1995.

[9] S. Muggleton. Bayesian inductive logic programming. In M. Warmuth, editor, *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory*, pages 3–11, New York, 1994. ACM Press.

[10] S. Muggleton. Inverse entailment and Progol. *New Generation Computing*, 13:245–286, 1995.

[11] S. Muggleton. Stochastic logic programs. In L. De Raedt, editor, *Advances in Inductive Logic Programming.* IOS Press/Ohmsha, 1996.

[12] S. Muggleton, M.E. Bain, J. Hayes-Michie, and D. Michie. An experimental comparison of human and machine learning formalisms. In *Proceedings of the Sixth International Workshop on Machine Learning*, Los Altos, CA, 1989. Kaufmann.

[13] S. Muggleton and C.D. Page. A learnability model for universal representations. Technical Report PRG-TR-3-94, Oxford University Computing Laboratory, Oxford, 1994.

[14] S. Pinker. *Language learnability and language development*. Harvard University Press, Cambridge, Mass., 1984.

[15] G.D. Plotkin. A note on inductive generalisation. In B. Meltzer and D. Michie, editors, *Machine Intelligence 5*, pages 153–163. Edinburgh University Press, Edinburgh, 1969.

[16] J.R. Quinlan and R.M. Cameron. Induction of logic programs: FOIL and related systems. *New Generation Computing*, 13:287–312, 1995.

[17] L. De Raedt and M. Bruynooghe. A theory of clausal discovery. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1993.

[18] T. Shinohara. Inductive inference of monotonic formal systems from positive data. In *Proceedings of the first international workshop on algorithmic learning theory*, Tokyo, 1990. Ohmsha.

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

A Proof of Theorems 1 and 2

Theorem 1 Expected error for positive examples only. *Let X be a countable instance space and D_X be a probability distribution over X. Let $\mathcal{H} \subseteq 2^X$ be a countable hypothesis space containing at least all finite subsets of X, and for which all $H \in \mathcal{H}$ have $D_X(H) \leq \frac{1}{2}$. Let $D_{\mathcal{H}}$ be a probability distribution over $\mathcal{H}$. Assume that $\mathcal{H}$ has an ordering $H_1, H_2, \ldots$ such that $D_{\mathcal{H}}(H_i) \geq D_{\mathcal{H}}(H_j)$ for all $j > i$. Let $D_{\mathcal{H}}(H_i) = \frac{a}{i^2}$ where $\frac{1}{a} = \sum_{i=1}^{\infty} \frac{1}{i^2} \approx \frac{1}{0.608}$. Let $\mathcal{H}_n = \{H_i : H_i \in \mathcal{H} \text{ and } i \leq n\}$. T is chosen randomly from $D_{\mathcal{H}}$ and the x_i in $E = \langle x_1, .., x_m \rangle$ are chosen randomly and independently from $D_{X|T}$. Let $f_m(H) = D_{\mathcal{H}}(H)(\frac{1}{D_X(H)})^m$ and let $n = am$. L is the following learning algorithm. If there are no hypotheses $H \in \mathcal{H}_n$ such that $H \supseteq H_E = \{x_1, .., x_m\}$ then $L(E) = H_E$. Otherwise $L(E) = H_n(E) = H$ only if $H \in \mathcal{H}_n$, $H \supseteq H_E$ and for all $H' \in \mathcal{H}_n$ for which $H' \supseteq H_E$ it is the case that $f_m(H) \geq f_m(H')$. The error of an hypothesis H is defined as*

Error$(H,T) = D_X(T \setminus H) + D_X(H \setminus T)$. *The expected error of* L *after* m *examples,* $EE(m)$*, is at most* $\frac{2.33+2\ln\, m}{m}$.

Proof. *Consider the case in which* $T \in \mathcal{H}_n$ *(case 1). Then by definition* $L(E) = H_n(E) = H$*. Since* $H = H_n(E)$ *has the maximum value of* f_m *in* $\mathcal{H}_n$ *it follows that*

$$D_{\mathcal{H}}(H)\left(\frac{1}{D_X(H)}\right)^m \geq D_{\mathcal{H}}(T)\left(\frac{1}{D_X(T)}\right)^m. \tag{1}$$

Also $D_X(T \cap H) = D_X(T) - D_X(T \setminus H) = D_X(H) - D_X(H \setminus T)$ *and therefore*

$$D_X(H) - D_X(T) = D_X(H \setminus T) - D_X(T \setminus H). \tag{2}$$

Furthermore consider the case in which $D_X(H) > D_X(T)$ *(case 1a). Rearranging (1) and taking logs we get*

$$m(ln\ D_X(H) - ln\ D_X(T)) \leq ln\ D_{\mathcal{H}}(H) - ln\ D_{\mathcal{H}}(T). \tag{3}$$

Let $r = \frac{\ln\, D_X(H) - \ln\, D_X(T)}{D_X(H) - D_X(T)}$*. We now show that for* $0 \leq D_X(T) < D_X(H) \leq \frac{1}{2}$ *it is the case that* $r > 2$*. First note that* r *decreases monotonically in* $D_X(T)$ *since*

$$\frac{\partial r}{\partial D_X(T)} = \frac{1}{D_X(H) - D_X(T)^2}(\ln\, \frac{D_X(H)}{D_X(T)} - \frac{D_X(H)}{D_X(T)} + 1) < 0$$

for $D_X(H) > D_X(T)$*. But as* $D_X(T)$ *approaches* $\frac{1}{2}$ *it also approaches* $D_X(H)$ *and* $\lim_{D_X(T) \to D_X(H)} r = \frac{\partial}{\partial D_X(T)} ln\ D_X(T) = \frac{1}{D_X(T)} = 2$*. Thus* $r > 2$ *for* $0 \leq D_X(T) < D_X(H) \leq \frac{1}{2}$ *from which it follows that* $ln\ D_X(H) - ln\ D_X(T) > 2(D_X(H) - D_X(T))$*. Combining this with (2) and (3) gives*

$$2m(D_X(H \setminus T) - D_X(T \setminus H)) \leq ln\ D_{\mathcal{H}}(H) - ln\ D_{\mathcal{H}}(T)$$

and therefore

$$D_X(H \setminus T) \leq \frac{-ln\ D_{\mathcal{H}}(T)}{2m} + D_X(T \setminus H). \tag{4}$$

Now consider the case in which $D_X(H) \leq D_X(T)$ *(case 1b). From (2) it follows that* $D_X(H \setminus T) \leq D_X(T \setminus H)$*. Thus (4) holds in both case 1a and case 1b, and therefore from the definition of Error in the theorem for all of case 1 we get*

$$Error(H_n(E), T) \leq \frac{-\ln\, D_{\mathcal{H}}(T)}{2m} + 2D_X(T \setminus H). \tag{5}$$

Lastly consider the case in which $T \notin \mathcal{H}_n$ *(case 2). In this case we have the trivial bound*

$$Error(H, T) \leq 1. \tag{6}$$

We are now in a position to bound $EE(m)$. First we define $T^1 = T$ and $T^m = T \times T^{m-1}$. Now

$$EE(m) = \sum_{T \in \mathcal{H}} D_{\mathcal{H}}(T) \sum_{E \in T^m} D_X(E|T) Error(L(E), T).$$

Splitting the sum into case 1 and case 2 and making use of (6) gives the following.

$$EE(m) \leq \sum_{T \in \mathcal{H}_n} D_{\mathcal{H}}(T) \sum_{E \in T^m} D_X(E|T) Error(H_n(E), T) + \sum_{T \in \mathcal{H} \setminus \mathcal{H}_n} D_{\mathcal{H}}(T).1$$

But since $D_{\mathcal{H}}(H_i) = \frac{a}{i^2}$,

$$\sum_{T \in \mathcal{H} \setminus \mathcal{H}_n} D_{\mathcal{H}}(T) = \sum_{i=n+1}^{\infty} \frac{a}{i^2} \leq \int_{i=n}^{\infty} \frac{a}{i^2} = \frac{a}{n}.$$

This together with (5) gives

$$\begin{aligned} EE(m) &\leq \frac{a}{n} + \sum_{T \in \mathcal{H}_n} D_{\mathcal{H}}(T) \sum_{E \in T^m} D_X(E|T) \left(\frac{-\ln D_{\mathcal{H}}(T)}{2m} + 2D_X(T \setminus H) \right) \\ &= \frac{a}{n} - \frac{1}{2m} \sum_{T \in \mathcal{H}_n} D_{\mathcal{H}}(T) ln\ D_{\mathcal{H}}(T) \\ &\quad +2 \sum_{T \in \mathcal{H}_n} D_{\mathcal{H}}(T) \sum_{E \in T^m} D_X(E|T) D_X(T \setminus H). \end{aligned}$$

But

$$-\frac{1}{2} \sum_{T \in \mathcal{H}_n} D_{\mathcal{H}}(T) ln\ D_{\mathcal{H}}(T) \leq -\frac{1}{2} \sum_{i=1}^{\infty} \frac{a}{i^2} ln\ (\frac{a}{i^2}) < 0.82.$$

Letting $\tau_{mn}(\epsilon) = \{E : E \in T^m$ and $D_X(T \setminus H_n(E)) \leq \epsilon D_X(T)\}$ and remembering that $D_X(T \setminus H) \leq D_X(T) \leq \frac{1}{2}$ gives

$$\begin{aligned} &\sum_{T \in H_n} D_{\mathcal{H}}(T) \sum_{E \in T^m} D_X(E|T) D_X(T \setminus H) \\ = &\sum_{T \in H_n} D_{\mathcal{H}}(T) \sum_{E \in \tau_{mn}(\epsilon)} D_X(E|T) D_X(T \setminus H) \\ &+ \sum_{T \in H_n} D_{\mathcal{H}}(T) \sum_{E \in T^m \setminus \tau_{mn}(\epsilon)} D_X(E|T) D_X(T \setminus H) \\ \leq &\ \frac{\epsilon}{2} + \sum_{T \in H_n} D_{\mathcal{H}}(T) \sum_{E \in T^m \setminus \tau_{mn}(\epsilon)} D_X(E|T) \frac{1}{2} \\ = &\ \frac{\epsilon}{2} + \frac{1}{2} \sum_{T \in H_n} D_{\mathcal{H}}(T) Pr(for\ random\ E, D_X(T \setminus H_n(E)) > \epsilon D_X(T)) \\ \leq &\ \frac{\epsilon}{2} + \frac{Pr(\exists H \in \mathcal{H}_n . D_X(T \setminus H) > \epsilon D_X(T)\ and\ x_1, .., x_m \in (T \cap H))}{2} \\ \leq &\ \frac{\epsilon + n(1-\epsilon)^m}{2} \\ \leq &\ \frac{\epsilon + ne^{-\epsilon m}}{2}. \end{aligned}$$

Thus

$$EE(m) \leq \frac{a}{n} + \frac{0.82}{m} + \epsilon + ne^{-\epsilon m}.$$

Optimal values of n and ϵ are found by successively setting to zero the partial derivatives of n, ϵ and solving. This gives $\epsilon = \frac{ln\ nm}{m}$ and $n = am$. Substituting gives

$$\begin{aligned} EE(m) &\leq \frac{1 + 0.82 + 2ln\ m + ln\ a + 1}{m} \\ &< \frac{2.33 + 2ln\ m}{m}. \end{aligned}$$

□

Theorem 2 Expected error for positive and negative examples. *Let X be a countable instance space and $\mathcal{H} \subseteq 2^X$ be a countable hypothesis space containing at least all finite subsets of X. Let $D_{\mathcal{H}}$, D_X be probability distributions over $\mathcal{H}$ and X. Assume that $\mathcal{H}$ has an ordering $H_1, H_2, \ldots$ such that $D_{\mathcal{H}}(H_i) \geq D_{\mathcal{H}}(H_j)$ for all $j > i$. Let $D_{\mathcal{H}}(H_i) = \frac{a}{i^2}$ where $\frac{1}{a} = \sum_{i=1}^{\infty} \frac{1}{i^2}$. Let $\mathcal{H}_n = \{H_i : H_i \in \mathcal{H}$ and $i \leq n\}$. T is chosen randomly from $D_{\mathcal{H}}$. Let $ex(x, H) = \langle x, v\rangle$ where $v = True$ if $x \in H$ and $v = False$ otherwise. Let $E = \langle ex(x_1, T), .., ex(x_m, T)\rangle$ where each x_i is chosen randomly and independently from D_X. $H_E = \{x : \langle x, True\rangle$ in $E\}$. Hypothesis H is said to be consistent with E if and only if $x_i \in H$ for each $\langle x_i, True\rangle$ in E and $x_j \notin H$ for each $\langle x_j, False\rangle$ in E. Let $f(H) = D_{\mathcal{H}}(H)$ and let $n = am$. L is the following learning algorithm. If there are no hypotheses $H \in \mathcal{H}_n$ consistent with E then $L(E) = H_E$. Otherwise $L(E) = H_n(E) = H$ only if $H \in \mathcal{H}_n$, H consistent with E and for all $H' \in \mathcal{H}_n$ consistent with E it is the case that $f(H) \geq f(H')$. The error of an hypothesis H is defined as $Error(H, T) = D_X(T \setminus H) + D_X(H \setminus T)$. For $n = am$ the expected error of L after m examples, $EE(m)$, is at most $\frac{1.51 + 2\ln\ m}{m}$.*

Proof. *Let $T^m = \{\langle ex(x_1, T), .., ex(x_m, T)\rangle : x_i \in T\}$. The expected error can be bounded in a similar way to that used in the proof of Theorem 1.*

$$\begin{aligned} EE(m) &= \sum_{T \in \mathcal{H}} D_{\mathcal{H}}(T) \sum_{E \in T^m} D_X(E|T) Error(L(E), T) \\ &\leq \sum_{T \in \mathcal{H}_n} D_{\mathcal{H}}(T) \sum_{E \in T^m} D_X(E|T) Error(H_n(E), T) + \sum_{T \in \mathcal{H} \setminus \mathcal{H}_n} D_{\mathcal{H}}(T).1 \\ &\leq \frac{a}{n} + \sum_{T \in \mathcal{H}_n} D_{\mathcal{H}}(T) \sum_{E \in T^m} D_X(E|T) Error(H_n(E), T) \end{aligned}$$

Letting $\tau_{mn}(\epsilon) = \{E' : E' \in T^m$ and $Error(H_n(E'), T) \leq \epsilon\}$ gives

$$\begin{aligned} &\sum_{T \in H_n} D_{\mathcal{H}}(T) \sum_{E \in T^m} D_X(E|T) Error(H_n(E), T) \\ &= \sum_{T \in H_n} D_{\mathcal{H}}(T) \sum_{E \in \tau_{mn}(\epsilon)} D_X(E|T) Error(H_n(E), T) \end{aligned}$$

$$
\begin{aligned}
& + \sum_{T \in H_n} D_{\mathcal{H}}(T) \sum_{E \in T^m \setminus \tau_{mn}(\epsilon)} D_X(E|T) Error(H_n(E), T) \\
\leq \;& \epsilon + Pr(\exists H \in \mathcal{H}_n . Error(H, T) > \epsilon \ and \ x_1, .., x_m \in (T \cap H)) \\
\leq \;& \epsilon + n(1-\epsilon)^m \\
\leq \;& \epsilon + ne^{-\epsilon m}
\end{aligned}
$$

Thus

$$EE(m) \leq \frac{a}{n} + \epsilon + ne^{-\epsilon m}$$

Again optimal values of n *and* ϵ *are found by successively setting to zero the partial derivatives of* n, ϵ *and solving. This gives* $\epsilon = \frac{\ln\ nm}{m}$ *and* $n = am$*. Substituting gives*

$$
\begin{aligned}
EE(m) &\leq \frac{1 + 2\ln\ m + \ln\ a + 1}{m} \\
&< \frac{1.51 + 2\ln\ m}{m}.
\end{aligned}
$$

□