

**LEARNING COHERENT CONCEPTS**

*Ashutosh Garg, Dan Roth*

Department of Computer Science and the Beckman Institute,  
University of Illinois at Urbana-Champaign  
{ashutosh,danr}@uiuc.edu

**Abstract**

This paper develops a theory for learning scenarios where multiple learners co-exist but there are mutual *coherency constraints* on their outcomes. This is natural in cognitive learning situations, where “natural” constraints are imposed on the outcomes of classifiers so that a valid sentence, image or any other domain representation is produced. We formalize these learning situations, after a model suggested in (Roth & Zelenko, 2000) and study generalization abilities of learning algorithms under these conditions in several frameworks. We show that the mere existence of coherency constraints, even without the learner’s awareness of them, deems the learning problem easier than predicted by general theories and explains the ability to generalize well from a fairly small number of examples. In particular, it is shown that within this model one can develop an understanding to several realistic learning situations such as highly biased training sets and low dimensional data that is embedded in high dimensional instance spaces.

## 1. Introduction

A fundamental research effort in learning theory has been the study of generalization abilities of learning algorithms and their dependence on sample complexity. The importance of this research direction goes beyond intellectual curiosity. Understanding the inherent difficulty of learning problems allows one to evaluate whether learning is at all possible in certain situations, estimate the degree of confidence in the predictions made by learned classifiers and is crucial in understanding and analyzing learning algorithms. In particular, these theoretical considerations played a crucial role in the development of practical learning approaches (Druker et al., 1993; Golding & Roth, 1999; Cortes & Vapnik, 1995).

One puzzling problem from a theoretical and a practical point of view is the contrast between the hardness of learning problems, as suggested by various bounds on sample complexity and generalization - even for fairly simple concepts - and the apparent ease at which the cognitive systems seem to learn those concepts. Cognitive systems seem to use far less examples and learn more robustly than is predicted by the theoretical models developed so far.

This work develops a learning theory that explains this phenomenon. Following (Roth & Zelenko, 2000) our approach is based on the observation that cognitive learning problems do not usually occur in isolation. Rather, the input is observed by multiple learners that may learn different functions on the same input. We pursue this direction by developing a theory for learning scenarios where multiple learners co-exist but there are mutual compatibility constraints on their outcomes. We believe that this is natural in cognitive learning situations, where “natural” compatibility constraints are imposed on the outcomes of classifiers so that a valid sentence, image or other domain representation is produced. In particular, this model can be viewed as a theoretical framework for learning in multi-modal situations.

Assume, for example, that one is trying to learn a function that determines, given a sentence which contains one of  $\{weather, whether\}$  which of the two should actually occur in the sentence. E.g., given the sentence I did not know

weather to laugh or cry determine if weather should be replaced by whether. The function learned to perform this task may be fairly complicated; it could depend on a huge number of features such as words neighboring the target words in sentences, their syntactic tags, etc. (Golding & Roth, 1999). Notice, however, that the same sentence could be supplied as input to a different function that predicts the part-of-speech (pos) of the word *weather* (and others) in this sentence. However, the *predictions* of these functions are not independent. For example, if the pos function determines, in a given context, that the target word is a noun, then the spelling function cannot determine that the correct spelling is *whether*. Other more intricate constraints exist with other functions that can receive this sentence as input. Consequently, perhaps, even though the data for problems of these sort typically reside in very high dimensional space (e.g.,  $10^3$  to  $10^6$ ), one is able to achieve good classification performance (on test data) by looking at relatively few training examples; *very* few relative to what is expected by theory and is needed in simulations of synthetic data of this dimensionality. Similar phenomena exist when learning to detect faces or properties of faces (e.g., gender) in visual learning problems.

This exemplifies our notion of *coherency constraints*: given that these two functions need to produce coherent outputs, the input sentence may not take *any* possible value in the input space (that it could have taken when the function’s learnability is studied in isolation) but rather may be restricted to a subset of the inputs on which the functions outcomes are coherent. In this paper we model these learning situations and develop a learning theory that attempts to explain these phenomena.

**Notations:** We consider the standard scenario of concept learning from examples. A learner is trying to identify a binary classifier  $c : X \rightarrow \{0, 1\}$  when presented with examples  $(x, y)$ , where instances  $x \in X (= \mathfrak{R}^n)$  are drawn according to a fixed (but unknown) distribution on  $X$  and labeled  $y = c(x)$ .  $m$  denotes the number of training examples.  $\mathcal{H}$  denotes the hypothesis space (the class of

functions from which a hypothesis is selected),  $|\mathcal{H}|$  is its cardinality and  $h \in \mathcal{H}$  refers to the learned hypothesis.

While our goal is to learn a single target concept,  $c : \mathcal{X}^n \rightarrow \{0, 1\}$ , we are interested in studying situations in which the learning scenario involves several concepts  $c_1, c_2, \dots, c_k$ . We further assume that the concepts are subject to a *constraint*  $g$  ( $g : X \times \{0, 1\}^k \rightarrow \{0, 1\}$ ) which is fixed (but could be probabilistic) and unknown to the learner. The constraint reflects the fact that all these functions represent different aspects of some natural data, as in the example above. We formalize this learning scenario and show that the mere existence of the other functions along with the constraints Nature imposes on the relations between these functions – all unknown to the learner – contribute to the effective simplification of the task of learning  $c_1$ .

The effect of constraints on the learning is analyzed by studying three models, with increased generality. We start by a pac analysis of the finite hypothesis class case, under coherency constraints. We then relax some of the assumptions and move to study constraints in the more general equivalence class framework. This allows us to develop a view of coherency as an equivalence relation on the hypothesis class. This view is beneficial in understanding conditions under which learning becomes easier and supports better generalization, even when the same hypothesis class is used. Finally, we develop a general VC-dimension view of coherency constraints. We show that these can be analyzed as a way to restrict the effective number of dichotomies and thus VC-dimensions techniques can be used to derive generalization bounds.

We also provide some examples that serve to motivate the framework and exemplify its power as well as some experimental evidence to its validity. In particular, we show that within this framework one can study and develop an understanding to several realistic learning situations such as highly biased training sets and low dimensional data that is embedded in high dimensional instance spaces.

## 2. Coherency Constraints

The usual way to constrain the learning task is to explicitly restrict the concept class. Instead, here we are concerned with the case in which the restriction is imposed implicitly via interaction among concepts. More precisely, we are interested in learning the concept  $c_1$  in a situation that involves several concepts  $c_1, c_2, \dots, c_k$ ;  $c_i : X \rightarrow \{0, 1\}$ , and a global constraint  $g : X \times \{0, 1\}^k \rightarrow \{0, 1\}$  on the outcomes of these concepts.

The concept of coherency constraints has been formalized in (Roth & Zelenko, 2000) where it was shown that it can be used to explain the “easiness” and robustness of learning in some restricted situations. Several semantics for *coherency* were discussed there. The notion of *Class Coherency* was developed to indicate coherency at the level of the outcome of the classifiers; this turns out to be too restrictive in that it restricted the hypothesis class to include only functions which are coherent with each other over all samples. This notion was then relaxed to define *Distributional Coherency*. In this case the hypothesis space is not restricted; rather, the effect of distributional coherency is in disallowing some of the instances – those on which the constraints are not satisfied – to occur in the input. Results are given for the case of mistake bound learning of half spaces under specific constraints. It was also shown that learning concepts under this model results in hypothesis that are more robust to attribute noise. Similar ideas on robustness have also been discussed in (Arriaga & Vempala, 1999).

The model studied in this paper builds on the distributional coherency model but extends it in several directions. First, we generalize it to general constraints; we extend it to a probabilistic setting and allow constraints to apply only with some probability; and, under these conditions we develop techniques to analyze general classifiers

Although we are interested in learning a single function  $c_1$ , in the following definition we will consider it along with the (possibly) constraining functions  $c_2, \dots, c_k$ , and denote  $\bar{c} = (c_1, c_2, \dots, c_k)$ . Thus the constraints in the following definition are imposed on the direct product  $\mathcal{C}^k$ . The semantics is that for

each  $\bar{c} \in \mathcal{C}^k$ , we restrict the domain of  $\bar{c}$  to  $X'$  where, with high probability, the constraint is satisfied; that is,  $\forall x \in X', g(\bar{c}(x)) = 1$ . Moreover, we allow  $g$  to depend on  $x$ , so that the constraints can take a very general form.

**Definition 1 (Distributional Coherency)** *Let  $\mathcal{C}$  be a class of functions  $c : X \rightarrow \{0, 1\}$ ,  $g : X \times \{0, 1\}^k \rightarrow \{0, 1\}$  a Boolean constraint, and  $\alpha \in [0, 1]$  a constant. We define the class of  $g$ -coherent functions  $\mathcal{C}_g^*$  to be the collection of all functions  $\bar{c}^* : X \rightarrow \{0, 1\}^k \cup \{\star\}$  in  $\mathcal{C}^k$  defined by*

$$\bar{c}^*(x) = \begin{cases} \bar{c}(x) & \text{if } P\{x \in X | g_x(\bar{c}(x)) = 1\} \geq \alpha \\ \star & \text{otherwise} \end{cases}$$

We interpret the value of “ $\star$ ” as a forbidden value for the function  $\bar{c}$ . In this way we restrict the domain of  $\bar{c}$  to the subset  $X'$  of  $X$  satisfying (w.h.p.) the constraint  $g$ .

In the pac learning model the above constraint can be interpreted as restricting the class of distributions when learning a function  $c_1 \in \mathcal{C}$ . Only distributions giving zero (or small, depending on  $\alpha$ ) weight to the region  $X \setminus X'$  are allowed.

We note that this is different from the model of distribution specific learning (e.g., (Benedek & Itai, 1991)). There, the learner is explicitly aware of the underlying distribution and can utilize it directly. Our model is based on assuming that this is unrealistic; instead, we assume a distribution free model in which the distribution could be constrained in intricate ways. The learner is unaware of this. However, as we show, under this model the learning problem nevertheless becomes easier and we can justify the generalization abilities of the learned hypothesis even in the presence of relatively small number of training examples.

### 3. PAC Analysis of Coherency Constraints

Consider the pac model (Valiant, 1984) analysis for the admissible case when the hypothesis space is finite. That is, it is assumed that during the training phase one is provided with  $m$  training examples drawn independently according to  $P$  and labeled according to some target concept  $c_1$ . The learning al-

gorithm chooses a hypothesis  $h \in \mathcal{H}$  that is consistent with target function on the training data. In this setting (Blumer et al., 1987) we know that for the true error of  $h$  (that is,  $Pr_P(h(x) \neq c_1(x))$ ) to be bounded by  $\epsilon$  with probability at least  $1 - \delta$ , the number of training examples required needs to be greater than

$$m \geq \frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{\epsilon}. \quad (1)$$

This analysis can be extended to the non admissible case (when the target function is not in  $\mathcal{H}$ ) and the assumption on  $|\mathcal{H}|$  being finite can be relaxed to a finite VC dimension. Eqn. 1 gives the relation between the true error and the sample complexity. On fixing the confidence ( $\delta$ ) and the hypothesis class ( $\mathcal{H}$ ), the sample complexity is inversely proportional to the true error.

We prove next an analogous result that exhibits the effect of the coherency constraints. W.l.o.g we present it for the case  $k = 2$ . As above, our goal is to learn a hypothesis  $h$  that approximates  $c_1$ ; the hypothesis is chosen such that it is consistent on  $m$  training examples with the target concept  $c_1$  and thus, based on the above, (with confidence  $\delta$ , which we fix for the rest of the discussion) it has a true error of  $\epsilon_1$ . We now analyze the effect the presence of the coherency constraint has on  $h$ 's performance. Before we do that, and in order to simplify the discussion that follows, we note that it is always possible to think about the coherency constraint as an *equality* constraint. The reason is that we can always replace  $c_2$  by  $c_2'$  deterministically, via the graph of  $g$ . Namely,  $c_2'$  is defined so that when  $g_x(c_1(x), c_2(x)) = 1$ ,  $c_2'(x) = c_2(x)$  if  $c_1(x) = c_2(x)$  and  $c_2'(x) = \neg c_2(x)$  otherwise. When  $g_x(c_1(x), c_2(x)) = 0$  we define  $c_2'$  exactly in the opposite way, yielding  $\forall x, g_x(c_1, c_2) \equiv [c_1 \equiv c_2']$ .

Assume the existence of a concept  $c_2$ , such that the learned hypothesis  $h$  has a true error of  $\epsilon_2$  w.r.t.  $c_2$ . Also, we assume that  $c_2$  coheres with the target function  $c_1$  via  $g$ , that is:

$$P\{x | g_x(c_1, c_2)\} = \alpha. \quad (2)$$

Consequently, we care about the performance of  $h$

only under these coherency constraints. In the following discussion we assume that the outcomes of the concepts  $c_1, c_2$  are independent<sup>1</sup> given the outcome of the hypothesis  $h$ . In addition, we make the following technical assumption. We assume that  $Pr(c_2 = 0|h = 1) = Pr(c_2 = 1|h = 0)$ , that is, that the labels of  $c_2$  are symmetric with respect to  $h^2$ . As we show in the next theorem, under these conditions, the hypothesis  $h$  learned to approximate  $c_1$ , actually achieves true error – relative to instances that are subject to coherency constraints – that is smaller than  $\epsilon_1$ . Equivalently, in order to achieve true error of  $\epsilon_1$  one needs to train on less than the  $m$  examples of Eqn. 1.

**Theorem 1** *Let  $c_1$  be a target concept and  $h$  a learned hypothesis that has true error  $\epsilon_1$  relative to it, based on Eqn. 1. Assume that  $h$  has true error  $\epsilon_2$  with respect to  $c_2$ . Then, the true error of  $h$  with respect to  $c_1$  on the data satisfying the constraint  $g$  (Eqn. 2.), and under the conditions given, is given by*

$$\epsilon = \frac{\epsilon_1 \epsilon_2 \alpha}{(1 - \epsilon_1)(1 - \epsilon_2) + \epsilon_1 \epsilon_2} + \frac{\epsilon_1(1 - \epsilon_2)(1 - \alpha)}{\epsilon_1(1 - \epsilon_2) + (1 - \epsilon_1)\epsilon_2}. \quad (3)$$

The proof is given in Appendix 1. Note that

**Lemma 1**  $\forall \epsilon_1$ , if  $(\epsilon_2 - 0.5)(\alpha - (1 - \epsilon_1)(1 - \epsilon_2) + \epsilon_1 \epsilon_2) < 0$  then the bound on  $\epsilon$  in Eqn. 3 satisfies  $\epsilon < \epsilon_1$ .

The lemma simply means that for values of  $\alpha$  which actually constraints the instances,  $\epsilon < \epsilon_1$ . The proof is by direct algebraic manipulation. Lemma. 1 and Thm. 1 together show that for coherency constrained data, the true error of  $h$  w.r.t.  $c_1$  is lower than the true error in general. Equivalently, one can achieve the same generalization using a smaller number of training examples. This reduction in sample complexity depends on (1) the degree ( $\alpha$ ) of coherency and (2) the performance ( $\epsilon_2$ ) of (the  $g$ -map of)  $h$  on  $c_2$ . An important point to note

<sup>1</sup>This is a reasonable assumption in many situations, e.g, the example given in the introduction. Moreover, many studies that use multiple classifiers (e.g., across modalities) make this assumption.

<sup>2</sup>This can be relaxed by splitting the region measured by  $\epsilon_2$  above to two regions, one given  $h = 1$  and the other given  $h = 0$  and defining  $Pr(h \neq c_2) = \max\{Pr(c_2 = 0|h = 1), Pr(c_2 = 1|h = 0)\}$ .

is that we do not assume that the learning algorithm knows  $c_2$ . It is the mere existence of this concept, which makes the learning of  $c_1$  easier. For the special case of a deterministic constraint, when  $\alpha = 1$ , the number of examples required for  $h$  to have a true error of  $\epsilon_1$  with respect to  $c_1$  on the constrained data, is given by

$$m \geq \frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{\frac{\epsilon_1(1 - \epsilon_2)}{\epsilon_1(1 - \epsilon_2) + (1 - \epsilon_1)\epsilon_2}} \quad (4)$$

It is straightforward to see that for  $0 \leq \epsilon_2 < 0.5$ , this is a better bound than the one given in Eqn. 1.

This case is similar to the one presented in (Blum & Mitchell, 1998). They have introduced the concept of *co-training* and have shown that the presence of unlabeled data can help under some consistency assumptions. They consider the example of labeling web pages, and it is argued that in this case two independent concepts exist which provide consistent labels. Their example can be mapped to our framework which then provides the guarantees missing in (Blum & Mitchell, 1998). What is further highlighted by our framework is the realization that the mere existence of the constraint makes the original learning problem easier (even without using it).

The importance of the sample complexity result above is the following interpretation of it. The presence of constraints reduces the number of training examples needed in order to achieve a certain generalization performance, relative to a constraint-free scenario. Stated differently, it directly addresses one of our concerns in the introduction: in these situations, one can believe the results of the learned predictor even though it was learned using a small number of examples. This can also be thought of as the PAC analysis of the case discussed in (Roth & Zelenko, 2000).

#### 4. Equivalence Class Analysis

In this section we relax one of the assumptions used in Sec. 3. Rather than assuming a finite hypothesis space we consider the more general case in which the hypothesis class is countably infinite and one assumes a probability distribution  $Q$  over it. For the standard learning model this case has been analyzed

in (McAllester, 1999) and others. Consider the following example.

**Example 1** Assume one is trying to learn a target function  $c_1$  in the presence of  $c_2$  and that  $c_2$  coheres with  $c_1$  on the observed data. Let  $\mathcal{H}$  be the class of monotone Boolean conjunctions over  $\{0, 1\}^n$  and assume that  $c_1, c_2$  are also in  $\mathcal{H}$ . We can thus think of  $c_1, c_2$  as elements in  $\{0, 1\}^n$  with the interpretation that  $c_i(j) = 1$  iff the conjunction  $x_i$  contains the variable  $x_j$  ( $i=1,2; j=1,\dots,n$ ). Assume now that  $c_1$  differs from  $c_2$  on  $k$  bits. Given the coherency constraint then, for all observed instances  $x \in \{0, 1\}^n$ , the corresponding  $k$  bits must be zero. As a result, all functions in  $\mathcal{H}$  which differ only on these bits are equivalent for this learning problem. The size of the equivalence class will depend on the number of bits  $k$  on which  $c_1, c_2$  differ.

The assumptions made in the above example can be relaxed in several ways. In particular, we can assume coherency with high probability and can still get a similar result in terms of an equivalence class with high probability. Some of the examples in (Roth & Zelenko, 2000) can also be analyzed via the equivalence class view.

Assume that there is some probability distribution  $Q$  over the hypothesis space. An equivalence class over this hypothesis space would then mean that one can consider a smaller effective hypothesis space. Figure 1 shows the probability distribution over the hypothesis class. Figure 2 maps the equivalence class view over this hypothesis space. That is, all hypotheses belonging to an equivalence class are indistinguishable due to the presence of a constraint or, equivalently, due to a property of the data. The effective probability distribution assigns to the equivalence classes weights which are proportional to their size. Below we use the pac-Bayes framework to quantify this and show that this view indeed implies tighter generalization bounds.

We assume a countably infinite hypothesis class  $\mathcal{H}$  with known probability distribution  $Q$  over it and study the effects of coherency constraints. Let  $\mathcal{C}$  be the class of hypotheses consistent with the data;  $c \in \mathcal{C}$ . Using (McAllester, 1999), the generaliza-

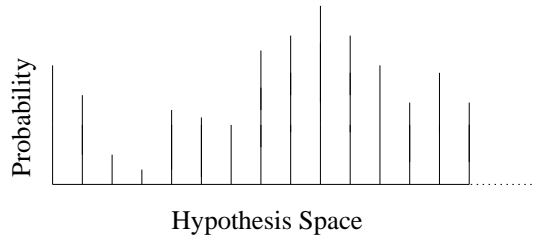


Figure 1. Probability distribution over the hypothesis space

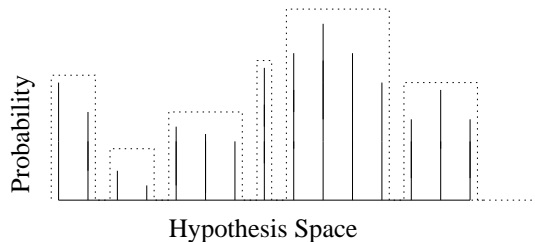


Figure 2. Probability distribution over the hypothesis space. Dotted lines show the equivalence class, i.e. all the hypotheses that fall in one group of the dotted lines are indistinguishable on the given data and are assigned probability equivalent to the sum of the probabilities of their equivalence class.

tion error is bounded by

$$\epsilon(c) \leq \frac{\ln \frac{1}{Q(c)} + \ln \frac{1}{\delta}}{m}. \quad (5)$$

Let  $\mathcal{H}_c \subseteq \mathcal{H}$ , such that  $\forall h \in \mathcal{H}_c$ , the constraint is satisfied with high probability. We can think of  $\mathcal{H}_c$  as the class of hypotheses that are representatives of the equivalence class, although the discussion that follows will apply to any projection (filtering) of the hypothesis class. We are interested in solving for  $\epsilon(c|c \in \mathcal{H}_c)$ , that is, the probability of a consistent hypothesis making an error given that it satisfies the constraints. To do that we need  $Q(c \in \mathcal{C}|c \in \mathcal{H}_c)$ . (I.e., we restrict our learning algorithm to consider only those hypotheses which satisfy some constraints). This is a more general case than the one discussed in the previous section. As discussed in (Roth & Zelenko, 2000), the constraints have the effect of reducing the size of the hypotheses space and this is what is observed here too. We first compute the term  $Q(c|c \in \mathcal{H}_c)$ .

$$Q(c|c \in \mathcal{H}_c) = \frac{Q(c, c \in \mathcal{H}_c)}{Q(c \in \mathcal{H}_c)} = \frac{Q(c)Q(c \in \mathcal{H}_c|c)}{Q(\mathcal{H}_c)} = \frac{\gamma Q(c)}{Q(\mathcal{H}_c)}, \quad (6)$$

where  $\gamma$  is the probability that a consistent hypothesis belongs to the subset of the hypotheses class satisfying the constraint. This leads to the following Lemma:

**Lemma 2**

$$\epsilon(c|c \in \mathcal{H}_c) \leq \frac{\ln \frac{1}{Q(c)} + \ln Q(\mathcal{H}_c) + \ln \frac{1}{\gamma} + \ln \frac{1}{\delta}}{m} \quad (7)$$

Note that here we are considering a weaker constraint; only with high probability a hypothesis consistent with the data satisfies the constraint. This can also be seen as modifying the probability distribution over the concept class which is governed by the presence of constraints.

**5. VC-Dimension Based Bounds**

In this section we consider the general case where the hypothesis class may contain infinite number of hypotheses. For the present analysis we will assume that the class has finite VC-dimension. We first introduce the basic principles of the VC theory and then develop related results under coherency constraints. Finally we discuss applications of these to some realistic learning scenario.

The VC-dimension based bounds can be intuitively viewed as extensions of the bounds for the case of finite hypothesis class (Sec. 3) only that the size of the hypothesis class is replaced by *annealed entropy*. The *annealed entropy* is a distribution dependent concept which is then bounded from above by a function of the VC-dimension (using Sauer’s lemma) and thus gives distribution free bounds. The annealed entropy is given as

$$H_{ann} = \int \Delta^\wedge(x^1, x^2, \dots, x^L) dF(\mathbf{x}) \quad (8)$$

where  $\Delta^\wedge(x^1, x^2, \dots, x^L)$  is the maximum number of dichotomies the sample  $x^1, x^2, \dots, x^L$  can have when using a given set of hypothesis. The integral is taken over all possible samples of size  $L$  thus giving the expected number of dichotomies that are possible for a given distribution over the data and a given hypothesis class. Given the annealed entropy, the bound on the generalization error is given by

$$P \left\{ \sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| > \epsilon \right\} < 4 \exp \left\{ \left( \frac{H_{ann}(2l)}{l} - \left( \epsilon - \frac{1}{l} \right)^2 \right) l \right\}, \quad (9)$$

where  $R(R_{emp})$  is the expected (empirical, resp.) risk associated with the target function  $\alpha$ . The

bound is developed in (Vapnik, 1998) (Theorem 4.1). It gives the explicit dependence of the true error bound on the annealed entropy. The smaller the annealed entropy, the tighter the bound is. This can also be thought of as the capacity of the hypothesis class as a function of the distribution over the data.

Indeed, this is a much better bound than the most commonly used VC-dimension bound. Consider a hypothesis class  $\mathcal{H}$  which consists of all hyperplane in  $\mathbb{R}^n$ , and assume that the distribution that governs the data generation supports only data points on the  $x$ -axis. For this case, while the VC-dimension of the hypothesis class is directly related to the dimensionality of the space ( $n + 1$ ) and is not effected by the distribution of the data, the annealed entropy of  $\mathcal{H}$  is independent of the space in which the data lies, yielding a much better and more realistic bound. The VC-dimension can thus be thought of as a function of annealed entropy for the worst case probability distribution on the data. However, since, in general, computing the annealed entropy is not feasible, the VC-dimension bound is commonly used.

Next we show that one can use limited information on the data distribution to obtain the *effective annealed entropy*. Our goal is related in spirit to the one studied in (Vapnik et al., 1994). While they give a method of calculating the effective VC-dimension given observed data we, instead, use similar techniques to bound the effective annealed entropy  $H_{ann}^{eff}$  within the coherency constraints framework.

**5.1 A General Framework for Constraints**

This section develops a general framework for modeling the effect of the constraints in terms of the effective annealed entropy; this can then be mapped to the effective VC-dimension of the data. Recall the coherency constraints definition (1). Denote

$$\Gamma_1 \subseteq \Gamma = \{x|g_x(\bar{c}) = 1\} \quad \Gamma_2 = \neg\Gamma_1. \quad (10)$$

This generalizes the discussion in Sec. 2. We assume that the constraint is satisfied only by a particular labeling of the data for  $\Gamma_1$  instances, however, for the case when  $x \in \Gamma_2$ , data can take any label.

Eqn. 8 can be written in terms of  $\Gamma_1, \Gamma_2$ . The expected value is taken over the space of all samples.

Since one is looking at the number of dichotomies that can be achieved for the given set of samples, this integral is over the  $L$  fold distribution. The annealed entropy can therefore be written as:

$$H_{ann} = \int_{\Gamma_1} \Delta^\wedge(x^1, \dots, x^L) dF(\mathbf{x}) + \int_{\Gamma_2} \Delta^\wedge(x^1, \dots, x^L) dF(\mathbf{x}) \quad (11)$$

Denote  $P(x \in \Gamma_1) = \alpha$ , the probability that a sample satisfies the constraint. The probability that not all the samples came from  $\Gamma_1$  is  $1 - \alpha^L$ . When all the samples are from  $\Gamma_1$ , then only one labeling of samples is possible; otherwise a large number of dichotomies are possible. We get the following bound on the *effective* annealed entropy:

$$H_{ann} \leq H_{ann}^{eff} = (1 - \alpha^L) H_{ann}^\wedge, \quad (12)$$

where  $H_{ann}^\wedge$  gives the maximum number of dichotomies of any set of  $L$  samples using the hypothesis from the given class. This bound gives a much smaller value of the effective annealed entropy for small values of  $L$ . However, for large values of  $L$ ,  $\alpha^L$  goes to zero as does the effect of constraints (in the formulation given). Thus, as the number of samples grow, the effect of the constraints as played in the simple argument above, on the generalization performance, goes down. To understand this, note that, intuitively, with infinite amount of data, if the constraints affect only a small portion of the instance space ( $\alpha$  is small) the number of ‘‘observed’’ dichotomies will be almost as large as the number of possible dichotomies. However, the more interesting case might be that of small values of  $L$ , and of large  $\alpha$ . Note that in the natural examples alluded to earlier in the paper,  $\alpha$  is large. Also, this analysis is still very general and makes no assumptions on the structure of the constraints. Recall, for example, the case discussed at the beginning of this section, where all the instances lie on the  $x$ -axis. Our future work will address exploiting general structural constraint to explicitly bound the annealed entropy. We exemplify these ideas in two concrete cases.

## 5.2 Highly Biased Class Probability

In many realistic learning problems the probability of observing positive examples is very small relative to that of the negative examples. Consider the

problem of face detection in Computer Vision. One may see only a few 10’s of positive examples and may see millions of negative examples. Similar phenomena occurs in many natural language and information extraction learning situations. The considerations developed earlier can be used to show that the generalization performance in these cases is better than predicted by current theories. To show that, we will compute the effective annealed entropy.

Denote by  $\alpha$  the probability of the positive class is  $\alpha$ ,  $1 - \alpha$  is the probability of the negative class. Without loss of generality, we will assume that  $\alpha \ll 1$ . To model the highly biased class probability as a coherency constraint, one can think of the equality constrain  $g(\bar{c})$  with  $c_1$  as the target function and  $c_2(x) \equiv 0$ . And, we assume that this constraint holds with high probability  $(1 - \alpha)$ . Using the analysis given for Eqn. 12, we obtain:

**Corollary 1** *Assume a highly biased class probability case, with the probability of the positive class being  $\alpha \ll 1$ . Then the effective annealed entropy for a data set of  $L$  samples is*

$$H_{ann}^{eff} \leq (1 - (1 - \alpha)^L) H_{ann}^\wedge \quad (13)$$

where  $H_{ann}$  is the annealed entropy (no assumptions).

For small values of  $L$ , we see that  $H_{ann}^{eff} \ll H_{ann}^\wedge$ . Although as  $L \rightarrow \infty$ ,  $H_{ann}^{eff} \rightarrow H_{ann}^\wedge$ , we argue that the interesting case is when  $L$  is not too large, since,

$$\lim_{L \rightarrow \infty} \frac{H_{ann}^\wedge(L)}{L} \rightarrow 0 \quad (14)$$

(This is a simple consequence of uniform convergence as the number of samples observed approaches infinity.)

We note that in this case one can observe the effect of the constraints not only as a consequence of the smaller effective annealed entropy but also directly by looking more closely on the form of Chernoff bound. In general, the binary classification problem is modeled as the convergence of the observed frequencies, in a Bernoulli experiment with mean  $p$ , to the true frequencies. The standard formulation used for the bound is that of Hoeffding Bound:

$$P(S > (p + \epsilon)m) \leq e^{-2m\epsilon^2},$$

which gives a bound which is independent of  $p$ . However, an exact analysis of the Chernoff bound for the Bernoulli case results in the tighter bound:

**Lemma 3**

$$P(S > (p + \epsilon)m) \leq e^{-n(\epsilon+p) \log \frac{\epsilon+p}{p} - n(1-\epsilon-p) \log \frac{1-\epsilon-p}{1-p}} \quad (15)$$

This can be easily verified using the standard definition of the Chernoff bound. Fig. 3 compares the standard bound given above to the tighter one given by Eqn. 15 as a function of the class probability. It is evident that the bound is significantly better for small values of  $p$ .

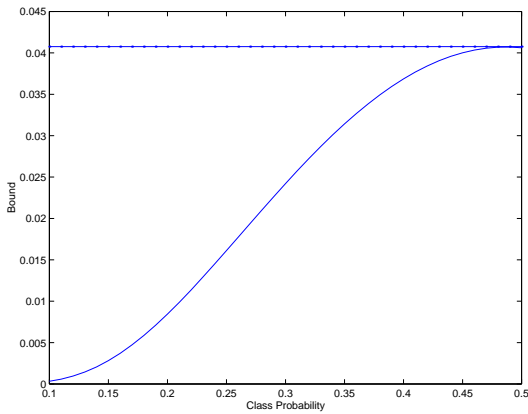


Figure 3. The dotted line gives the bound for the standard Chernoff bound; the solid line is the tighter version of the bound.

**5.3 Linear Mapping to Higher Dimensional Space**

As a second example to the ideas discussed in this section we consider a problem of learning a linear classifier for a data lying in a high dimensional space (say  $M$ ). Due to the high dimensionality it is very likely that the training data is linearly separable. In fact, in many natural language and visual learning problems the dimensionality is larger than the number of training instances. The basic question is to understand the generalization properties of the resulting classifier, given that it was learned based on a small number of examples relative to the dimensionality.

For simplicity, we assume that there exists a one-to-one mapping of the  $M$  dimensional data to a lower

dimensional space,  $N$ , through a linear transformation. In this case, we show that the data is linearly separable in the  $N$  dimensional space. (E.g., think of a case in which the the data is originally in a lower dimensional space  $N$  but is being observed in a higher dimensional space.) and the generalization performance is thus governed by it.

Notice that even in this case, the problem of recovering the transformation matrix and using it to map the data back to the lower dimensional space is intractable. However, our claim is that this is not necessary and learning in the high dimensional space does not require to see more data. To see that, let  $x = (x_1, x_2, \dots, x_M)$  be a training example and  $h = (h_1, h_2, \dots, h_M)$  the linear classifier in the higher dimensional space. Denote  $z = (z_1, z_2, \dots, z_N)$  the data point in the  $N$  dimensional space such that  $x = Az$  where  $A$  is the (unknown)  $M \times N$  transformation matrix. The outcome of the classifier is  $y = h^t x$ , which can also be written as

$$y = h^t x = h^t A z = (A^t h)^t z = (h')^t z$$

That is, there exists a linear classifier  $h'$  in the lower dimensional space that will achieve the same performance as  $h$  in the higher dimensional space. The idea is that one doesn't need to know either  $A$  or  $h'$ .

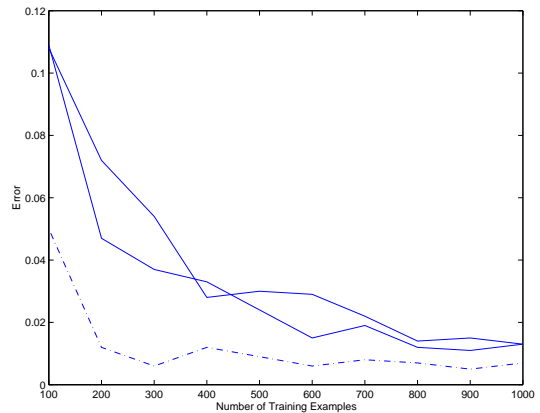


Figure 4. Learning Curve for the noiseless case. The dotted curve shows the learning case in the presence of the constraint and the solid lines show the learning curves of the individual classifiers in the absence of constraints.

This scenario is also directly representable in the coherency constraint framework, as in Def. 1. To do

that, let  $c_1 \equiv h$ , the target function in the  $M$  dimensional space,  $c_2 \equiv h' \circ A^{-1}$ , and let the constraint be the equality constraint. Clearly, this simple scenario can be relaxed to the full generality of the definition, but the outcome is essentially the same. That is, there is no need to recover the transformation, but rather the fact that it exists implies that the generalization properties are as good as they could be in the lower dimensional case. The constraints can also be used directly to show that the VC-dimension in this case is actually  $N + 1$  and not  $M + 1$  as was originally thought. We note that this case has also been discussed using the notion of random-projection in (Arriaga & Vempala, 1999).

## 6. Experiments

In this section we describe some preliminary experiments we use to exhibit and evaluate the implications of the insights gained in this work. We considered the problem of learning a half space in the presence of another, constraining, half space. Specifically, data was sampled from an  $n$  dimensional space, but the (randomly chosen) classifiers  $c_1, c_2$  actually depend only on  $n/2$  dimensions:  $c_1$  depends on  $x_1, \dots, x_{n/2}$  and  $c_2$  on  $x_{n/2+1}, \dots, x_n$ . We show learning curves for learning  $c_1$  given data sampled uniformly from  $\mathbb{R}^n$ , and also of learning  $c_1$  when the data observed is filtered to satisfy the equality constraint, that is, for all input instances  $x$ ,  $c_1(x) = c_2(x)$ . (For completeness we also show the curves for  $c_2$ .)

Figure 4 shows the learning rate, for the noise free case, with and without the constraints. We used data in  $\mathbb{R}^{24}$ , and tested on 1000 examples. The curves give the errors as a function of the number of training examples; the solid curve – for the individual half-spaces and the dot-dash curve for learning in presence of the equality constraint.

Fig. 5 depicts the results of the same experiment for noisy data (this time, data lied in  $\mathbb{R}^{10}$ ). It is clearly evident from the learning curves shown that the classifier is able to learn much faster in the presence of the constraint. As we have pointed out throughout the paper, the learning algorithm is unaware of the existence of form of the constraints.

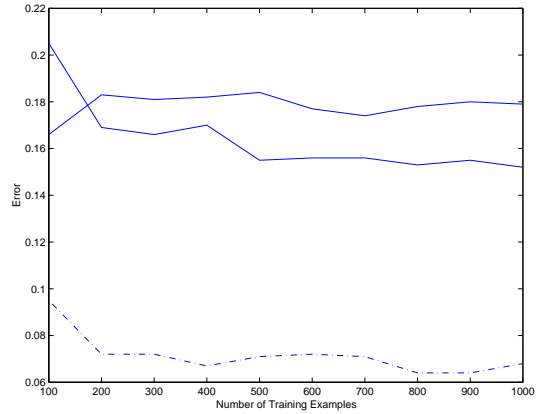


Figure 5. Learning Curve for the noisy case. The dotted curve shows the learning case in the presence of the constraint and the solid lines show the learning curves of the individual classifiers in the absence of constraints.

## 7. Conclusions

The power of existing models of learning (Valiant, 1984; Vapnik, 1998) stems from the distribution-free nature of the model. The underlying assumption is that the probability distribution governing the occurrences of instances is too complex to model and a theory should be developed without making explicit assumptions on it. The resulting theories, however, cannot explain well a wide range of phenomena, in which learning can be done robustly from a relatively small number of examples.

In this work we have developed a learning model within which we attempt to explain these phenomena. The key observation underling this model is that, in many situations, learning problems do not occur in isolation. Our model is therefore concerned with learning scenarios where multiple learners co-exist but there are mutual coherency constraints on their outcomes. Within this model, we have developed generalization bounds and have shown that in the presence of coherency constraints the learning problem indeed becomes easier than predicted by the general theories. This could explain the ability to generalize well from a fairly small number of examples and can help in understanding several realistic learning situations.

While several works (e.g., (Amsterdam, 1988)) have criticized the distribution free pac learning model as being too restrictive this work still pur-

sues the distribution free approach (see discussion in Sec. 2). In some sense, our model can be viewed as an intermediate model between the worst case distribution free model that is commonly studied in learning theory and the simpler, but unrealistic, distribution specific model (in which one assumes a complete knowledge of the distribution, and can utilize it when learning). We assume, instead, a distribution free model in which the distribution could be constrained in natural, but intricate ways. The learner is unaware of this. This view opens up a number of questions; in particular, an interesting direction could be to understand generalization under specific families of constraints.

## References

- Amsterdam, J. (1988). Some philosophical problems with formal learning theory. *National Conference on Artificial Intelligence* (pp. 580–584).
- Arriaga, R. I., & Vempala, S. (1999). An algorithmic theory of learning: Robust concepts and random projection. *Proc. of the 40th Foundations of Computer Science*.
- Benedek, G., & Itai, A. (1991). Learnability with respect to fixed distributions. *Theoret. Comput. Sci.*, 86, 377–389.
- Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proc. of the Annual ACM Workshop on Computational Learning Theory* (pp. 92–100).
- Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's razor. *Information Processing Letters*, 24, 377–380.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.
- Druker, H., Schapire, R., & Simard, P. (1993). Improving performance in neural networks using a boosting algorithm. *Neural Information Processing Systems 5* (pp. 42–49). Morgan Kaufmann.
- Golding, A. R., & Roth, D. (1999). A Winnow based approach to context-sensitive spelling correction. *Machine Learning*, 34, 107–130. Special Issue on Machine Learning and Natural Language.
- McAllester, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, 37, 355–363.
- Roth, D., & Zelenko, D. (2000). Towards a theory of coherent concepts. *National Conference on Artificial Intelligence* (pp. 639–644).
- Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27, 1134–1142.
- Vapnik, V., Levin, E., & Cun, Y. L. (1994). Measuring the VC-dimension of a learning machine. *Neural Computation*, 6, 851–876.
- Vapnik, V. N. (1998). *Statistical learning theory*. New York: John-Wiley and Sons Inc.

## 8. Appendix

**Theorem 1** *Let  $c_1$  be a target concept and  $h$  a learned hypothesis that has true error  $\epsilon_1$  relative to it, based on Eqn. 1. Assume that  $h$  has true error  $\epsilon_2$  with respect to  $c_2$ . Then, the true error of  $h$  with respect to  $c_1$  on the data satisfying the constraint  $g$  (Eqn. 2,) and under the conditions given, is given by*

$$\epsilon = \frac{\epsilon_1 \epsilon_2 \alpha}{(1 - \epsilon_1)(1 - \epsilon_2) + \epsilon_1 \epsilon_2} + \frac{\epsilon_1(1 - \epsilon_2)(1 - \alpha)}{\epsilon_1(1 - \epsilon_2) + (1 - \epsilon_1)\epsilon_2}. \quad (16)$$

**Proof of Thm. 1:** Given the discussion, it is sufficient to prove the theorem for the case of  $g$  being the equality constraint. We denote by  $G$  the constraint set. I.e.  $x \in G$  implies that the sample  $x$  follows the constraint. Also denote  $C = \{x | c_1(x) = c_2(x)\}$  and  $\neg C$  its compliment. Similarly  $H_c^i = \{x | h(x) = c_i(x)\}$ .

The true error of  $h$  w.r.t.  $c_1$  on a sample satisfying the constraint is given by

$$\begin{aligned} P(h(x) \neq c_1(x) | x \in G) &= P(\neg H_c^1 | x \in G) \\ &= P(\neg H_c^1, C | x \in G) + P(\neg H_c^1, \neg C | x \in G) \\ &= P(\neg H_c^1 | C, x \in G)P(C | x \in G) \\ &\quad + P(\neg H_c^1 | \neg C, x \in G)P(\neg C | x \in G) \\ &= P(\neg H_c^1 | C)P(C | x \in G) + P(\neg H_c^1 | \neg C)P(\neg C | x \in G) \\ &= \frac{P(\neg H_c^1, C)\alpha}{P(C)} + \frac{P(\neg H_c^1, \neg C)(1-\alpha)}{P(\neg C)} \\ &= \frac{P(\neg H_c^1, \neg H_c^2)\alpha}{P(C)} + \frac{P(\neg H_c^1, H_c^2)(1-\alpha)}{P(\neg C)} \\ &= \frac{\epsilon_1 \epsilon_2 \alpha}{(1 - \epsilon_1)(1 - \epsilon_2) + \epsilon_1 \epsilon_2} + \frac{\epsilon_1(1 - \epsilon_2)(1 - \alpha)}{(1 - \epsilon_1)\epsilon_2 + \epsilon_1(1 - \epsilon_2)} \equiv \epsilon \end{aligned}$$

The fourth equality follows from the fact that conditioned upon the  $C$  or  $\neg C$ , whether  $h(x)$  agrees with  $c_1(x)$  or not is independent of whether  $x \in G$ . The reason is that the effect of the constraint is simply in determining the probability of  $C$ . The sixth equality is due to set equality (e.g.,  $\neg H_c^1 \cap C = \neg H_c^1 \cap \neg H_c^2$ ). The seventh equality uses the fact that, decisions made by concepts  $c_1, c_2$ , are independent of each other given  $h$ . To see it more specifically, one has to go through a series of probabilistic inequalities. Let  $H$  refers to the set of all  $x$ , such that  $h(x) = 1$  and  $\neg H$  refers to the set of  $x$  such that  $h(x) = 0$ .

$$\begin{aligned} P(\neg H_c^1, \neg H_c^2) &= \\ &= P(c_1(x) \neq h(x), c_2(x) \neq h(x) | H)P(H) \\ &\quad + P(c_1(x) \neq h(x), c_2(x) \neq h(x) | \neg H)P(\neg H) \\ &= P(c_1(x) \neq h(x) | H)P(c_2(x) \neq h(x) | H)P(H) \\ &\quad + P(c_1(x) \neq h(x) | \neg H)P(c_2(x) \neq h(x) | \neg H)P(\neg H) \\ &= P(c_2(x) \neq h(x) | H)\{P(c_1(x) \neq h(x) | H)P(H) \\ &\quad + P(c_1(x) \neq h(x) | \neg H)P(\neg H)\} \\ &= \epsilon_1 \epsilon_2 \end{aligned}$$

Where we have used the fact that if  $P(c_2(x) = 1 | h(x) = 0) = P(c_2(x) = 0 | h(x) = 1)$ , then  $P(c_2(x) = 1 | h(x) = 0) = \epsilon_2$ . Using the similar argument, one can derive the other two term  $P(C)$  and the term  $P(\neg H_c^1, H_c^2)$ . We have also used the fact that  $P(x : c_1(x) = c_2(x) | x \in G) = \alpha$  (Eqn. 2). This proves the theorem.  $\square$