A Probabilistic Identification Result

Eric McCreath

Basser Department of Computer Science
University of Sydney NSW 2006 Australia
ericm@cs.usyd.edu.au

Abstract. The approach used to assess a learning algorithm should reflect the type of environment we place the algorithm within. Often learners are given examples that both contain noise and are governed by a particular distribution. Hence, probabilistic identification in the limit is an appropriate tool for assessing such learners. In this paper we introduce an exact notion of probabilistic identification in the limit based on Laird's thesis. The strategy presented incorporates a variety of learning situations including: noise free positive examples, noisy independently generated examples, and noise free with both positive and negative examples. This yields a useful technique for assessing the effectiveness of a learner when training data is governed by a distribution and is possibly noisy. An attempt has been made to give a preliminary theoretical evaluation of the Q-heuristic. To this end, we have shown that a learner using the Q-heuristic stochastically learns in the limit any finite class of concepts, even when noise is present in the training examples. This result is encouraging, because with enough data, there is the expectation that the learner will induce a correct hypothesis. The proof of this result is extended to show that a restricted infinite class of concepts can also be stochastically learnt in the limit. The restriction requires the hypothesis space to be *g-sparse*.

1 Introduction

The type of training examples provided to a learner has a significant effect on the class of concepts that may be learnt. For example, in the identification in the limit framework, by restricting the training examples to positive only examples we severely restrict the class of concepts that may be identified. However, by attaching a distribution to the instance space, providing the positive examples to the learner according to this distribution, the class of concepts that may be learnt is extended [12]. Also, the environment in which we assess a learning system should reflect the environment in which we expect the learner to operate. We often expect learners to operate in domains that both contain noise and training examples which are governed by some distribution. This provides a strong motivation for probabilistic identification in the limit, introduced by Laird [7, 8], where training examples are possibly noisy. Laird's approach, although embracing noise in the training examples, assumes both positive and negative examples are provided to the learner. Whereas, the approach taken in this paper uses an oracle to determine if an example will be positive or negative. This generalizes the type of training examples given to a learner, permitting probabilistic identification results to encompass a larger variety of learning situations. The stochastic process used to generate example texts and the definition of probabilistic identification is presented in section 2.

The Q heuristic was designed for an ILP system, LIME. This system learns from possibly noisy data where the number of positive and negative training examples are fixed and independent from the concept provided[11, 10]. The heuristic simply uses Bayes rule[1] given the assumptions regarding the training examples. We show that a learner which employs the Q heuristic will stochastically learn in the limit:

- any finite class of concepts, and
- a restricted infinite classes of concepts.

Of course, a finite class of concepts is trivially learnable from positive only data in the identification in the limit setting [6]. Hence, it is also learnable in the stochastic identification in the limit setting. What keeps our result from being trivial is the presence of noise in the data. Having presented this result, we explore conditions under which the result can be extended to an infinite class of concepts. The proof techniques for the infinite case, which introduces the notion of *g-sparse* hypothises spaces, builds on that of the finite case. These results are presented in section 3.

Section 4 contains two example concepts classes which may be shown to be *g-sparse*. We finally discuss possible future direction in section 5.

2 Probabilistic Identification in the Limit

Probabilistic identification in the limit extends identification in the limit by replacing the teacher that presents all the examples to the learner with a teacher

[1] $P(H|E) = \frac{P(E|H)P(H)}{P(E)}$

that uses a distribution to present examples to the learner. The criterion of success is correspondingly altered requiring that with probability 1 the learner induces a correct hypothesis all but finitely many times.

Let X be the instance space and D_X be a probability measure over X. Note, D_X is a mapping from 2^X to $[0,1]$ and $D_X(\{x\})$ is simply written $D_X(x)$. We also assume X to be a countable set. Recall that members of 2^X are concepts. Let C be a class of concepts. The probability cover of a concept c, defined $\theta(c)$, is $D_X(c) = \sum_{x \in c} D_X(x)$.

The error or difference between two concepts c_1 and c_2 with respect to the probability measure D_X is defined as $\text{error}(c_1, c_2) = \theta(c_1 \Delta c_2)$. By using *error* to evaluate a hypothesis the hypothesis only needs to be correct on instances which have nonzero probability in the instance space distribution. This is reasonable as the learner will never be presented with an instance with zero probability.

We let $\mathbb{E} = X \times \{\text{Pos}, \text{Neg}\}$ be the set of all labelled instances of the instance space X. We usually refer to labelled instances as examples. An example text $E \in \mathbb{E}^\infty$ is an infinite sequence of examples. The learner conjectures a hypothesis from an initial finite sequence of E. This initial finite sequence of E of length m is denoted $E[m]$. We let SEQ denote the set of all initial finite sequences, $\{E[m] | E \in \mathbb{E}^\infty \wedge m \in \mathbb{N}\}$.

Let h be a hypothesis. In the present work, h is a computer program. The extension of a hypothesis h, denoted $ext(h)$, is the concept which h represents. A hypothesis space is a sequence (usually infinite) of hypothesis. We assume that the hypothesis space H under consideration is enumerable. Let $h_0, h_1, \ldots$ be an enumeration of H. We further assume that H is uniformly decidable, i.e., there exists a computable function $f : \mathbb{N} \times X \to \{0,1\}$ defined below:

$$f(i,x) = \begin{cases} 1 & \text{if } x \in ext(h_i), \\ 0 & \text{otherwise.} \end{cases}$$

We say that a hypothesis space H is complete with respect to a concept class C if for each $c \in C$, there is a hypothesis h in the space H such that $c = ext(h)$.

We define a learner M to be a computable machine that implements a mapping from SEQ into H.

We also assume the learner is able to compute $\theta(ext(h))$ for any h in the hypothesis space H. Note that such a capability is unlikely to be available to any computable learner, however, $\theta(ext(h))$ may always be estimated and its exact value is not critical to induce the hypothesis with the largest Q-value[2].

Definition 1 (Convergence). *Learner M converges to hypothesis h on E just in case for all but finitely many $m \in \mathbb{N}$, $M(E[m]) = h$. This is denoted $M(E)\downarrow = h$.*

A stochastic process GEN is used to generate these example texts. This process may be formulated in a variety of ways depending on the kind of tests against which the learner is to be benchmarked. The example texts generated

[2] Note that, the Q-value is the value use to compare competing hypotheses.

will reflect the target concept, although it may not be an exact or complete representation of the target concept. As the text may contain examples which have opposite labelling to that which would reflect the concept. Also, there is no explicit requirement that the text contain a complete set of instances.

We now introduce a general stochastic process for generating example texts, this process is denoted $\mathrm{GEN}^{O}_{\langle \mu_p, \mu_n \rangle}$. The parameters $\langle \mu_p, \mu_n \rangle$ governs the amount of noise in the texts generated. μ_p gives the level of noise in the positive examples and correspondingly μ_n for the negative examples. In most cases $\mu_p = \mu_n$, however, it is useful to allow these parameters to be different in some cases. By setting $\mu_p = \mu_n = 0$ the process will generate noise free example texts. The parameter $O \in \{\mathrm{Pos}, \mathrm{Neg}\}^{\infty}$ is an oracle which determines which elements will be positive and negative in the sequence generated by $\mathrm{GEN}^{O}_{\langle \mu_p, \mu_n \rangle}$ prior to any instance being selected. The n'th element in the oracle O is denoted $O(n)$. By using an oracle we may model a variety of situations. For example, the oracle may determine all examples in the example text to be negative, hence we will model learning from only negative examples. We show the stochastic convergence results for any oracle, thus proving the result for a variety of situations. We may also place a probability measure over $\{\mathrm{Pos}, \mathrm{Neg}\}^{\infty}$ and assume O is stochastically generated by such a measure. As the stochastic learning result is shown for any $O \in \{\mathrm{Pos}, \mathrm{Neg}\}^{\infty}$ the result will be also true for an oracle generated by any stochastic process.

The algorithm for $\mathrm{GEN}^{O}_{\langle \mu_p, \mu_n \rangle}(c, X, D_X)$ works as follows. In each cycle of the main loop the next example in the example text is generated. The oracle O is used to determine if the next example will be positive or negative. If the oracle decides that the next example will be positive, the following process is used: a biased coin is flipped where the probability of the coin coming up "Heads" is μ_p and "Tails" is $1-\mu_p$; if the coin comes up "Heads" then any instance is randomly selected from X using D_X and output as a positive example, if the coin is "Tails" then any instance is randomly selected from c using the distribution J^c_X where:

$$J^c_X(x) = \begin{cases} D_X(x)/\theta(c) & \text{if } x \in c, \\ 0 & \text{otherwise.} \end{cases}$$

A similar process is used if the oracle decides that the next example will be negative. This algorithm generates a text which reflects the concept c, where the sign of each example in the text matches the sign of the corresponding element in O and the parameters $\langle \mu_p, \mu_n \rangle$ determine the levels of noise introduced into the example text.

We now calculate the probability measure over $\mathbb{E}$ for each example generated by $\mathrm{GEN}^{O}_{\langle \mu_p, \mu_n \rangle}$. There are two possible probability measures an example may have, either G^+ or G^-. The n'th element of the example text will have probability measure G^+ if $O(n) = \mathrm{Pos}$, otherwise it will have probability measure G^- when $O(n) = \mathrm{Neg}$.

So when the oracle O determines the n'th example to be labelled "Pos", that is $O(n) = \text{Pos}$, example $\langle x, s \rangle$ is governed by:

$$G^+(\langle x, s \rangle) = \begin{cases} \mu_p D_X(x) + (1 - \mu_p)(J^c_X(x)) & \text{if } s = \text{Pos}, \\ 0 & \text{if } s = \text{Neg}. \end{cases}$$

Correspondingly, for the examples where $O(n) = \text{Neg}$:

$$G^-(\langle x, s \rangle) = \begin{cases} \mu_n D_X(x) + (1 - \mu_n)(J^{\overline{c}}_X(x)) & \text{if } s = \text{Neg}, \\ 0 & \text{if } s = \text{Pos}. \end{cases}$$

As D_X, $J^{\overline{c}}_X$, and J^c_X are probability measures on X it is straightforward to show that G^+ and G^- are probability measures on $\mathbb{E}$.

Note $G^+(\mathbb{E}) = 1$ and $G^-(\mathbb{E}) = 1$. These measures are used to define the probability measure $\text{Prob}_{\text{GEN}^O_{\langle \mu_p, \mu_n \rangle}(c, X, D_X)}$ on the σ-field $\mathcal{F} \subset 2^{(\mathbb{E}^\infty)}$, where $\mathcal{F}$ is the σ-field generated from the prefix sets of $2^{(\mathbb{E}^\infty)}$. Note that for every prefix set $B_\sigma = \{E \in \mathbb{E}^\infty | \sigma = E[|\sigma|]\}$ where $\sigma = \langle e_0, e_1, \ldots, e_n \rangle$ we have $\text{Prob}_{\text{GEN}^O_{\langle \mu_p, \mu_n \rangle}(c, X, D_X)}(B_\sigma) = \prod_{n < |\sigma|} f(e_n, O(n))$, where

$$f(\langle x, s \rangle, o) = \begin{cases} G^+(\langle x, s \rangle) & \text{if } o = \text{Pos}, \\ G^-(\langle x, s \rangle) & \text{if } o = \text{Neg}. \end{cases}$$

We refer the reader to *Measure Theory and Probability* by Adam and Guillemin [1] or *Probability and Measure* by Billingsley [3] for further information on measure theory.

Using $\text{GEN}^O_{\langle \mu_p, \mu_n \rangle}$ provides a flexible way of modelling different forms of training data. We now provide a list of common models for training data and show how these are specializations of $\text{GEN}^O_{\langle \mu_p, \mu_n \rangle}$.

Noise free, positive examples: If we set $\mu_p = \mu_n = 0$ and set $O = \langle \text{Pos}, \text{Pos}, \text{Pos}, \ldots \rangle$ the training data will be noise free and positive. The distribution of this training data will reflect a normalized version of the instance space distribution, where elements outside the target concept have probability zero of appearing in a text. This is identical to the assumption about the training data used by Montagna and Simi [12] who showed that whatever may be learnt in the limit from both positive and negative data may also be stochastically learnt in the limit from only positive data. This result assumes D_X is approximately computable. This is also similar to the model used by Angluin [2] when she considered TXTEX-identification. Angluin allows a null or empty element, denoted $\star$, to be part of the text, to facilitate modelling a text for the empty language.

Noisy, independently generated examples: Laird's [7, 8] classification noise process assumes that instances are chosen according to some distribution and then correctly labelled according to the target concept. After this a demon with probability ξ flips the class label from positive to negative or from negative to

positive, thereby creating noise in the training data[3]. This process generates an example text where each example is independent and has the following distribution:

$$P_{\text{Laird}}(\langle x, s\rangle) = \begin{cases} (1-\xi)D_X(x) & \text{if } s = \text{Pos} \wedge x \in c, \\ \xi D_X(x) & \text{if } s = \text{Pos} \wedge x \notin c, \\ \xi D_X(x) & \text{if } s = \text{Neg} \wedge x \in c, \\ (1-\xi)D_X(x) & \text{if } s = \text{Neg} \wedge x \notin c. \end{cases}$$

Now let us see how this distribution can be modelled in our framework. We now place a probability measure over $\{\text{Pos}, \text{Neg}\}^\infty$ such that each element is the sequence is independent and is "Pos" with probability w and "Neg" with probability $1-w$. We denote an oracle produced by such a distribution O_w.

Now, each element in the example text produced by $\text{GEN}^{O_w}_{\langle \mu_p, \mu_n \rangle}$ will be independent and have the following distribution:

$$P(\langle x, s\rangle) = \begin{cases} w(\mu_p + (1-\mu_p)/\theta(c))D_X(x) & \text{if } s = \text{Pos} \wedge x \in c, \\ w\mu_p D_X(x) & \text{if } s = \text{Pos} \wedge x \notin c, \\ (1-w)\mu_n D_X(x) & \text{if } s = \text{Neg} \wedge x \in c, \\ (1-w)(\mu_n + (1-\mu_n)/(1-\theta(c)))D_X(x) & \text{if } s = \text{Neg} \wedge x \notin c. \end{cases}$$

Now, let $w = \theta(c) + \xi - 2\theta(c)\xi$, $\mu_p = \frac{\xi}{\theta(c)+\xi-2\theta(c)\xi}$, and $\mu_n = \frac{\xi}{1-\theta(c)-\xi+2\theta(c)\xi}$. Then the distribution for each example in the example text generated by $\text{GEN}^{O_w}_{\langle \mu_p, \mu_n \rangle}$ will be identical to P_{Laird}. It follows, their probability measures over $\mathbb{E}^\infty$ will also be identical. Hence, by showing a result for stochastic learning with $\text{GEN}^{O}_{\langle \mu_p, \mu_n \rangle}$ we correspondingly show the result for Laird's model of training data.

Noise free, with both positive and negative examples: Learning with both positive and negative examples is the same as EX-identification where the functions in question have range restricted to either "Pos" or "Neg". Angluin [2] when considering EX-identification in a probabilistic setting assumes that each example is independent in the text and the probability of an example appearing is based on a distribution from the range of the function. This gives us the following distribution over the examples:

$$P(\langle x, s\rangle) = \begin{cases} D_X(x) & \text{if } s = \text{Pos} \wedge x \in c, \\ 0 & \text{if } s = \text{Pos} \wedge x \notin c, \\ 0 & \text{if } s = \text{Neg} \wedge x \in c, \\ D_X(x) & \text{if } s = \text{Neg} \wedge x \notin c. \end{cases}$$

If the oracle O_w, as defined in the previous model, where $w = \theta(c)$ and $\mu_p = \mu_n = 0$, then $\text{GEN}^{O_w}_{\langle \mu_p, \mu_n \rangle}$ gives the same distribution over each of the generated examples in the text.

[3] Laird [7] uses μ for the noise parameter, however, as it different to the noise parameter used here, we use ξ to refer to Laird's noise parameter.

By showing a learner to stochastically identify a class of concepts C when examples are provided by $\mathrm{GEN}^0_{\langle \mu_p, \mu_n \rangle}$, we also show that the learner will stochastically identify C when examples are provided by distributions used in the other models.

Definition 2, of probabilistic identification in the limit, is based on the definition given in Laird's thesis [7, 8].

Definition 2 (Probabilistic identification in the limit). *Given an instance space X and a probability measure D_X over X. A learner M is said to identify the class of concepts C stochastically in the limit, with respect to a hypothesis space H, if and only if*

(a) *examples are provided by GEN, and*

(b) $(\forall c \in C)\ Prob_{GEN(c,X,D_X)}\left\{ E \in \mathbb{E}^\infty \,\middle|\, M(E)\downarrow = h \wedge error(ext(h), c) = 0 \right\} = 1.$

This setting has the expected property that any subset of a class that is stochastically learnable in the limit is also stochastically learnable in the limit with respect to the same hypothesis space.

Laird [7] shows that any class of concepts that has a recursively enumerable set of hypotheses may be stochastically identified in the limit.[4] This assumes both positive and negative examples are presented to the learner according to the distribution. This result is then extended by Laird to include noise in the training examples. Both the Borel-Cantelli[5] and Hoeffding's probability inequality [5], used in the proofs by Laird, are also central to the results given in this paper.

3 Probabilistic Identification with the Q-heuristic

Let m be the total number of examples presented to the learner, so $m = n + p$, where p is the number of positive examples and n is the number of negative examples. Let $\mathrm{GEN}^O_{\langle \mu_p, \mu_n \rangle}$ generate the example text E. The learner M, given initial sequence $E[m]$ induces the hypothesis $M(E[m])$.

The order of presentation of examples, the sign of examples, and the proportion of positive and negative examples is dependent on the choice of oracle O. Since these aspects of the example presentation are not crucial for the learning algorithm, we assume that the learner is provided with a multiset of positive examples (of cardinality p) a multiset of negative examples (of cardinality n).

The algorithm simply works by choosing the hypothesis with the maximum[6] Q value [7] given the current examples. In general there may be a set of hypotheses with equal Q values. To stop the algorithm alternating between them, the

[4] Note that, if hypotheses are total Turing programs then a recursively enumerable set of hypotheses is the same as a uniform recursive set of hypotheses.

[5] The reader is directed to an introductory text on measure theory such as, *Measure Theory and Probability* [1] for more information.

[6] The notation $\mathrm{argmax}_{h \in H} Q(h)$ denotes the set $\{h \in H \mid (\forall h' \in H)\ Q(h) \geq Q(h'))\}$.

[7] $Q_\sigma(h) = \lg(P(h)) + |\mathrm{TP}_\sigma| \lg\left(\frac{1-\epsilon}{\theta(ext(h))} + \epsilon\right) + |\mathrm{TN}_\sigma| \lg\left(\frac{1-\epsilon}{1-\theta(ext(h))} + \epsilon\right) + |\mathrm{FPN}_\sigma| \lg(\epsilon)$ where TP_σ, TN_σ, and FPN_σ are respectively the true positive, true negatives, and

hypothesis with the minimum index[8] is chosen. If this minimum index selection is removed then the algorithm will still learn stochastically in the limit, although only in the behaviourally correct sense. Note that the algorithm is computable as it only must consider a finite initial portion of the possibly infinite hypothesis space [9, proposition 4.3.2].

Input :
  An indexed hypothesis space H.
  A prior probability distribution over H.
  A function θ for evaluating the theta value of a hypothesis.
  A sequence $\sigma = E[m]$ of m examples from the example text E.
  The noise parameter $\epsilon \in [0, 1)$ such that $\mu_p \leq \epsilon$ and $\mu_n \leq \epsilon$.
Output :
  A hypothesis h.
$h :=$ minindex(argmax$_{h \in H} Q_\sigma(h)$)
output h

Algorithm 1: Stochastic Identification using the Q heuristic

3.1 A finite concept class

Theorem 1. *Let C be any finite concept class and let H be any hypothesis space which is complete for C. Then for any noise parameter ϵ, there exists a learning algorithm that stochastically identifies C in the limit with respect to H when examples are provided by $GEN^O_{\langle \mu_p, \mu_n \rangle}$ for any oracle O and any $\mu_p \leq \epsilon$ and $\mu_n \leq \epsilon$.*

Proof. Due to space limitations we only briefly outline the proof here, a full version may be obtained in the authors thesis [9]. The proof compares h_t, a hypothesis that correctly classifies the target concept, with h_δ, a hypothesis that is in error. The value of $Q(h_t) - Q(h_\delta)$ is partitioned into three parts: a fixed constant, a sum of a list of random variables each corresponding to a positive example, and a sum of a list of random variables each corresponding to a negative example. The expected value for each of these random variables is shown to be positive. Assuming that the sum of these random variable is at least half the expected sum, we will have $Q(h_t) > Q(h_\delta)$ at some point, even when the fixed constant is negative. Applying Hoeffding's inequality, we compute a bound on the failure of this assumption. This bound is then used in conjunction with the Borel-Cantelli lemma to show that the class of concepts can be stochastically identified in the limit. □

false negatives or positives where the initial sequence is evaluated using hypothesis h.

[8] The notation minindex(S) denotes the hypothesis $h \in S$ such that $(\forall h' \in S - \{h\})\ h < h'$, where $<$ is a total ordering on H.

3.2 A restricted infinite concept class

The problem with extending the above result to an infinite concept class is when the hypotheses, with respect to their priors, converge on the target concept too quickly. When this occurs over an infinite set of concepts the bound on inducing an incorrect hypothesis is not finite. To address this problem a restriction is placed on the rate any hypothesis may be converged on.

Definition 3 (g-sparse with respect to a concept). *Let $g : \mathbb{N} \to \mathbb{R}$. Let c be a concept. A hypothesis space, $H = \{h_i | i \in \mathbb{N}\}$, is said to be* g-sparse *with respect to concept c if there exists $m_c \in \mathbb{N}$ and $w_c \in \mathbb{R}$ such that for all $j > m_c$, we have:*

$$error(c, ext(h_j)) \neq 0 \Rightarrow error(c, ext(h_j)) \geq w_c g(j).$$

Definition 4 (g-sparse). *A hypothesis space H is said to be* g-sparse *if H is* g-sparse *with respect to concepts $\emptyset$, the instance space X, and $ext(h_i)$ for all $h_i \in H$.*

Theorem 2. *Let C be any concept class and H be any hypothesis space which is complete for C. Let $\epsilon \in [0, 1)$ be the noise parameter. Assuming H is* g-sparse *where $g(i) = \frac{1}{i^\alpha}$ for $\alpha < 1$, there exists an algorithm that stochastically identifies C in the limit with respect to H when examples are provided by $GEN^O_{\langle \mu_p, \mu_n \rangle}$ for any oracle O and any $\mu_p \leq \epsilon$ and $\mu_n \leq \epsilon$.*

Proof. Similarly we only briefly outline the proof here, see [9] for the full version. This proof extends the previous proof. The learner once again uses Algorithm 1.

Given the g-spares constraint we may apply Hoeffdings inequality to find a bound on the probability of inducing an incorrect hypothesis. This bound is then used in conjunction with the Borel-Cantelli lemma to show that the class of concepts can be stochastically identified in the limit. □

4 Example Concept Classes

The learnability results presented in the previous two sections are interesting because our model incorporates noise in the data and a stochastic criterion of success. We feel that our approach is more realistic because although the classes discussed previously are learnable in the limit (in the traditional Gold [4] sense), they are not learnable in the Gold setting if noise is present. We next consider a class that is not learnable in the limit from positive only data in Gold's setting, but is learnable in our stochastic setting from only positive data even in the presence of noise.

The proofs of Propositions 1 and 2 work by showing that the hypothesis spaces in question are *g-sparse* with respect to a instance space distribution and then applying Theorem 2. The reader is directed to [9] for these proofs.

Proposition 1. *Let* $H = \{h_1, h_2, h_3, \dots\} = \{\mathbb{N}, \emptyset, \{1\}, \{2\}, \{1,2\}, \{3\}, \{1,3\}, \dots\}$. *The concept class consisting of all the finite subsets of* $\mathbb{N}$ *together with* $\mathbb{N}$ *is stochastically learnable in the limit with respect to* H.[9]

The g-sparse constraint is not a strong restriction as most enumerations of a hypotheses would generally not "target" a particular hypothesis "quickly".

We now consider the classes of concepts that consists of the empty set, the set of naturals and sets of the form $\{1, 2, \dots, k\}$, this class is a subset of the class shown to be stochastically learnable in the limit in the previous proposition, hence, the class will also be stochastically learnable in the limit.[10] However, we include this result as it may be proved using a restricted hypothesis space and a different instance space distribution which forms a tighter bound on the g-sparse restriction, and hence a more difficult concept to learn.

Proposition 2. *Let the instance space* X *be* $\mathbb{N}$. *Let the instance space distribution* $D_X(x) = \frac{s_1}{x^{3/2}}$ *where* s_1 *is the normalizing constant. Let* $H = \{h_1, h_2, h_3, \dots\} = \{\mathbb{N}, \emptyset, \{1\}, \{1,2\}, \{1,2,3\}, \{1,2,3,4\}, \dots\}$. *The concept class consisting of* $\mathbb{N}$ *and* $\emptyset$ *together with* $\{\{1, 2, \dots, k\} | k \in \mathbb{N}\}$ *is stochastically learnable in the limit with respect to* H.

5 Discussion

The results of stochastic identification in the limit in this paper are preliminary. An open question is whether these results could be extended to take into account complexity issues. This would give some idea of the the expected number of training examples provided to the learner, before the correct hypothesis is induced. In this case both the distribution of concepts presented to the learner and the prior probability used become critical. Another open question is what are the characteristics of *g-sparse* hypothesis spaces.

6 Acknowledgements

I thank Arun Sharma for his insight and advice. I also thank him for his encouragement to extend the initial finite result. I would also like to thank the reviewers for their helpful comments.

References

1. M. Adams and V. Guillemin. *Measure Theory and Probability.* Birkhäuser Boston, 1996.
2. D. Angluin. Identifying languages from stochastic examples. Technical Report TR–614, University of Yale, 1988.

[9] Since the data presentation could be guided by any oracle, this result holds for positive only data, too.

[10] See Proposition 7.1.1 of [9].

3. P. Billingsley. *Probability and Measure.* John Wiley & Sons, 1995.
4. E. M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.
5. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58:13–30, 1963.
6. S. Jain, D. Osherson, J. Royer, and A. Sharma. *Systems That Learn: An Introduction to Learning Theory.* MIT Press, second edition edition, 1999.
7. P. Laird. *Learning from Good Data and Bad.* PhD thesis, Yale University, 1987.
8. P. Laird. *Learning from Good and Bad Data.* Kluwer Academic Publishers, Boston, MA, 1988.
9. E. McCreath. *Induction in First Order Logic from Noisy Training Examples and Fixed Example Set Sizes.* PhD thesis, The University of New South Wales, 1999.
10. E. McCreath and A. Sharma. ILP with noise and fixed example size: A Bayesian approach. In *Fifteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1310–1315, 1997.
11. E. McCreath and A. Sharma. Lime: A system for learning relations. In *The 9th International Workshop on Algorithmic Learning Theory.* Springer-Verlag, October 1998.
12. F. Montagna and G. Simi. Paradigms in measure theoretic learning and in informant learning. Unpublished Draft.