617

# Identifying Languages From Stochastic Examples

Dana Angluin *
Yale University

March 1988

**Abstract**

We consider the problem of identifying an unknown formal language in the limit when the source of information about the language is a sequence of examples drawn independently according to some probability distribution.

## 1 Introduction

Gold [6] introduced the criterion of identification in the limit for successful learning of a formal language. He showed that there was a fundamental difference in what could be learned from positive versus complete samples. A positive sample presents all and only strings from the unknown language. A complete sample presents all strings, each classified as to whether it belongs to the unknown language. Gold showed that very simple classes of languages, including the class of regular sets, cannot be successfully identified from positive samples. By contrast, any recursively enumerable class of recursive languages, including the regular, context-free, and context-sensitive languages, can be successfully identified from complete samples.

In discussing the weakness of positive samples (which he termed "text"), Gold suggested the following approach.

> Perhaps this can be prevented by some reasonable probabilistic assumption concerning the generation of the text. In this case one would only require identification in the limit with probability one, rather than for every allowed text.

Horning [7] considered the case of stochastic context-free grammars, and assumed that sample sequences were generated from the unknown grammar according to the probabilities assigned to the productions. No negative examples were included in the sample sequences. He gave a learning algorithm, and proved that it converged in a certain sense to the correct grammar in the limit with probability one. Van der Mude and Walker [14] proved a similar type of convergence with probability one for an algorithm to learn stochastic regular grammars.

Wexler and Culicover [15], in their study of learnability of transformational grammars, also assume that a stochastic process generates positive examples. However, they require

*Supported by NSF grant IRI-8404226

that the learning procedure converge in the limit in Gold's sense to a correct grammar with probability one. This is a stronger requirement than the type of convergence used by Horning. They give a learning procedure that succeeds in this sense on a restricted class of grammars. Osherson, Stob, and Weinstein [10] define a notion of a uniformly measurable class of recursively enumerable sets, and show that any such class can be effectively learned in the limit in Gold's sense with probability one.

These positive results on learning large classes of languages from stochastically generated positive examples suggest that the assumption of stochastically generated samples is able to compensate for the lack of explicit negative information in the samples. These results also invite comparison with a new criterion of finite learnability proposed by Valiant [13].

In Valiant's setting, there is an unknown probability distribution on examples, and the learning algorithm draws random examples and attempts to construct and output a hypothesis that is "not too different" from the correct language "with high probability". Both "not too different" and "with high probability" are quantified with respect to the unknown probability distribution. Valiant and others have given a number of learning algorithms that succeed with respect to this criterion. This is a very strong criterion, since nothing is assumed about the unknown distribution – it need not be computable, or well-behaved in any sense.

Our study is motivated by the question of what has to be assumed about the probability distribution in order to achieve the kinds of positive results on language identification described above. We define an analog of Valiant's finite criterion for limit identification, and show that in this case, the assumption of stochastically generated examples does not enlarge the class of learnable sets of languages. This settles an open question in [10]. We also give a general definition of identifying or approximating a probability distribution in the limit, and shown that a fairly weak computability restriction is sufficient to guarantee identifiability. This generalizes the comparable results of Horning and Osherson, Stob, and Weinstein described above.

# 2 Preliminaries

## 2.1 Functions, languages, presentations

We take the universe of possible elements to be $N$, the set of all natural numbers. We assume that appropriate computable codings are chosen represent strings of symbols or pairs of natural numbers by natural numbers, as necessary. A *language* is a subset of $N$. The *characteristic function* of the language $L$, denoted $\chi_L$, is defined by $\chi_L(x) = 1$ if $x \in L$, and $\chi_L(x) = 0$ if $x \notin L$. A special symbol, $*$, will also be used in presenting examples.

The set of all total functions from $N$ to $N$ is denoted $F[N, N]$. Let $f \in F[N, N]$. A *complete presentation* of $f$ is an infinite sequence

$$\sigma = \langle x_0, y_0 \rangle, \langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots$$

such that for all $i \in N$, $y_i = f(x_i)$, and for every $x \in N$ there exists at least one $i$ such that $x_i = x$. Thus, a complete presentation of a function eventually exhibits every argument/value pair for that function.

Let $L$ be a language. A *complete presentation* of $L$ is a complete presentation of $\chi_L$, the characteristic function of $L$. A complete presentation of $L$ eventually classifies every element of $N$ as to its membership in $L$. The following is an initial segment of an uncountable number of complete presentations of the set of even elements of $N$:

$$\langle 2,1\rangle, \langle 7,0\rangle, \langle 16,1\rangle, \langle 7,0\rangle, \langle 0,1\rangle.$$

A *positive presentation* of a language $L$ is an infinite sequence

$$\sigma = x_0, x_1, x_2, \ldots$$

of elements of $L \cup \{*\}$ with the property that for every element $x \in L$, there exists at least one $i$ such that $x_i = x$. Thus a positive presentation includes all the elements of $L$, and no other elements of $N$. Note that if $L$ is the empty set, then it has only one positive presentation, consisting of an infinite sequence of $*$'s. The following is an initial segment of an uncountable number of positive presentations of the set of even elements of $N$:

$$*, 2, *, *, *, 16, *, 4, 0, 66, 0, 16, *.$$

Let

$$\phi_0, \phi_1, \phi_2, \ldots$$

be an acceptable numbering [9] of all the partial recursive functions from $N$ to $N$. If we let $W_i$ denote the set of inputs $x$ for which $\phi_i(x)$ is defined, then

$$W_0, W_1, W_2, \ldots$$

is an enumeration of all the recursively enumerable subsets of $N$. A *recursively enumerable language* is any recursively enumerable subset of $N$.

## 2.2 Inductive inference machines

An *inductive inference machine* is a Turing machine with an input tape, an output tape, a work tape, and four additional special states: input-requested, input-available, output-ready, output-recorded. We imagine running such a machine on an infinite sequence of natural numbers

$$\sigma = x_0, x_1, x_2, \ldots$$

as follows. The machine is started in its start state. If and when it ever enters its input-requested state, the input tape is erased, the code for the next value of $x_i$ is written on the input tape, the input head is positioned at the beginning of the input tape, and the machine is continued in its input-available state. If and when it ever enters its output-ready state, the number encoded by the initial non-blank portion of the output tape is appended to the end of the (initially empty) output sequence, and the machine is continued in its output-recorded state. If $\sigma$ is an infinite sequence of natural numbers, let $M[\sigma]$ denote the empty, finite, or infinite sequence of natural numbers output by $M$ with $\sigma$ as input.

We also consider *probabilistic* inductive inference machines, which in addition to the above have a one-way read-only coin tape filled with an infinite sequence of H's and T's.

The next-state function may depend on the value read on the coin tape, and the head is advanced one symbol to the right each time the coin tape is read.

Each coin tape determines a unique computation of the the probabilistic inductive inference machine. In particular, if $\sigma$ is an infinite sequence of natural numbers and $\tau$ is an infinite sequence of H's and T's, let $M[\sigma,\tau]$ denote the unique empty, finite, or infinite sequence of natural numbers output by $M$ when run with input $\sigma$ and coin tape $\tau$. Probabilities are assigned by assuming that the values on the coin tape are the results of an infinite sequence of tosses of a fair coin, with H representing "heads", and T representing "tails". Pitt [11] gives more detail on probabilistic inductive inference machines.

## 2.3 Criteria of identification: EX and TXTEX

A finite non-empty sequence of natural numbers *converges* to its last element. An infinite sequence of natural numbers *converges* to the value $i$ if all but finitely many of the elements of the sequence are equal to $i$.

The basic criterion of identification, EX, is defined as follows. Let $\sigma$ be an infinite sequence of natural numbers, let $M$ be an inductive inference machine, and let $f \in F[N,N]$. Then $M$ *EX-identifies* $f$ *on input* $\sigma$ if and only if $M[\sigma]$ is a non-empty sequence that converges to some $i$ such that $\phi_i = f$. $M$ *EX-identifies* $f$ if and only if for every complete presentation $\sigma$ of $f$, $M$ EX-identifies $f$ on input $\sigma$. Let

$$EX(M) = \{f \in F[N,N] : M \text{ EX-identifies } f\}.$$

Then $EX$ is the set of all $S \subseteq F[N,N]$ such that for some inductive inference machine $M$, $S \subseteq EX(M)$.

A large number of variants of this basic identification criterion have been studied. For languages we shall be interested primarily in the TXTEX criterion, in which hypotheses are interpreted as enumerators rather than decision-rules, and the inputs are positive presentations.

Let $M$ be an inductive inference machine, $L$ a subset of $N$, and $\sigma$ an infinite sequence of natural numbers. $M$ *TXTEX-identifies* $L$ *on input* $\sigma$ if and only if $M[\sigma]$ is non-empty and converges to some $i$ such that $W_i = L$. $M$ *TXTEX-identifies* $L$ if and only if for every positive presentation $\sigma$ of $L$, $M$ TXTEX-identifies $L$ on input $\sigma$. Let

$$TXTEX(M) = \{L \subseteq N : M \text{ TXTEX-identifies } L\}.$$

Then $TXTEX$ denotes the set of all classes $C$ of languages such that for some inductive inference machine $M$, $C \subseteq TXTEX(M)$.

## 2.4 Probabilistic criteria of identification

Pitt [11] has defined probabilistic versions of these identification criteria. If $M$ is a probabilistic inductive inference machine and $\sigma$ is an infinite sequence of inputs, then the set of coin tapes $\tau$ for which $M[\sigma,\tau]$ converges to any particular index $i$ is a measurable set. Hence, the probability that $M$ EX-identifies a function $f$ on input $\sigma$ is well-defined.

If $f$ is a function and $p$ is a real number then $M$ *EX-identifies* $f$ *with probability* $p$ if and only if for every complete presentation $\sigma$ of $f$, the probability that $M$ EX-identifies $f$

on input $\sigma$ is at least $p$. Let

$$EX_{prob(p)}(M) = \{f \in F[N,N] : M \text{ EX-identifies } f \text{ with probability } p\}.$$

Then $EX_{prob(p)}$ denotes the collection of all sets $S$ of functions such that for some probabilistic inductive inference machine $M$, $S \subseteq EX_{prob(p)}(M)$. The definitions for $TXTEX_{prob(p)}$ are analogous.

Pitt [11] shows that the $EX_{prob(p)}$ classes form a discrete hierarchy, with "breakpoints" at $p = 1/n$ for $n = 1, 2, 3, \ldots$. In particular, he proves

**Theorem 1** *For each $p > 1/2$, $EX_{prob(p)} = EX$. $EX$ is a proper subset of $EX_{prob(1/2)}$.*

(Wiehagen, Freivalds, and Kinber [16] also prove this.) For TXTEX-identification Pitt does not have a complete characterization, but he proves

**Theorem 2** *For each $p > 2/3$, $TXTEX_{prob(p)} = TXTEX$. $TXTEX$ is properly contained in $TXTEX_{prob(1/2)}$.*

The gap between 1/2 and 2/3 is an open problem. An improvement of this result would correspondingly improve Corollary 10.

These results show that if the probability of identification is required to be above some threshold, randomization is no advantage for EX-identification or TXTEX-identification. The results below show that if no assumption is made about the probability distribution, stochastic input gives no greater power than the ability to flip coins.

### 2.5 Probability distributions

The set of real numbers between 0 and 1 inclusive is denoted $[0,1]$. Let $X$ be a non-empty finite or countable set. A *distribution* on $X$ is a function $D$ mapping $X$ to $[0,1]$ such that the sum of $D(x)$ for all $x \in X$ is 1. (This is a "density function" in strict usage.) The *support* of $D$, denoted $S(D)$, is defined by

$$S(D) = \{x \in X : D(x) > 0\}.$$

Let $X$ be any non-empty finite or countable set, and let $\sigma = x_0, x_1, x_2, \ldots$ be an infinite sequence of elements of $X$. Then *range*$(\sigma)$ denotes the set of elements $x \in X$ such that $x_i = x$ for some $i \in N$.

**Example 3** *We define a specific distribution, $D_\sigma$, on $X$ such that*

$$S(D_\sigma) = range(\sigma).$$

*For each $x \in X$, let*

$$I(x) = \{i \in N : x_i = x\}.$$

*$I(x)$ is the set of indices of appearances of $x$ in $\sigma$. Define, for each $x \in X$,*

$$D_\sigma(x) = \sum_{i \in I(x)} 1/2^{i+1},$$

*with a sum over an empty set of indices interpreted as 0.*

Then it is clear that $D_\sigma(x) > 0$ if and only if $x \in range(\sigma)$. Moreover, since every index $i$ appears in exactly one set $I(x)$, it is clear that

$$\sum_{x \in X} D_\sigma(x) = \sum_{i \in N} 1/2^{i+1} = 1,$$

as required. Thus, $D_\sigma$ is a distribution on $X$ such that $S(D_\sigma) = range(\sigma)$.

### 2.6 The *DRAW*(*D*) oracle

If $X$ is any non-empty countable set and $D$ is any distribution on $X$, then *DRAW*($D$) is an oracle that is called with no input and returns an element of $X$ drawn according to $D$. Each call to the oracle is an independent event.

**Example 4** *We describe a probabilistic inductive inference machine $M$ such that for any infinite sequence $\sigma$ of elements of $N \cup \{*\}$, $M[\sigma]$ simulates an infinite sequence of calls to the oracle DRAW($D_\sigma$), where $D_\sigma$ is defined in Example 3.*

*$M$ keeps a table, initially empty, of the inputs it has read, in the order in which they were read. To simulate a call to DRAW($D_\sigma$), $M$ sets a counter $C$ to 1, and reads coin tosses from the tape, incrementing the counter $C$ for each $T$ read from the tape. At the first $H$ read, $M$ stops reading the coin tape. If the table of inputs has fewer than $C$ inputs, $M$ proceeds to read and store inputs until the table has $C$ inputs. $M$ then outputs the element at position $C$ in the table (numbering from 1.)*

Let

$$\sigma = x_0, x_1, x_2, \ldots$$

be an infinite sequence of elements of $N \cup \{*\}$. When $M$ is run with input $\sigma$, $M$ independently selects element $x_i$ with probability $1/2^{i+1}$ to be each output. Thus, for each $x \in N \cup \{*\}$, the probability that $x$ is selected is precisely $D_\sigma(x)$, as required. It is clear that each output is an independent event, and the probability that $M$ fails to produce an infinite sequence of outputs is 0.

## 3 The distribution-free case

In this section we give definitions analogous to Valiant's [13] for the cases of EX-identification and TXTEX-identification.

### 3.1 Complete distributions

Let $D$ be a distribution on $N$. We say that $D$ is *complete* if and only if $D(x) > 0$ for all $x \in N$. Let $f \in F[N,N]$ and let $D$ be a complete distribution on $N$. $D$ and $f$ determine a distribution $D[f]$ on $N \times N$, the set of ordered pairs of real numbers, as follows.

$$D[f](\langle x, y\rangle) = D(x) \text{ if } f(x) = y,$$

$$D[f](\langle x, y\rangle) = 0 \text{ if } f(x) \neq y.$$

The following is immediate.

**Lemma 5** *If $D$ is a complete distribution on $N$ and $f \in F[N,N]$, then the sequence of values returned by an infinite sequence of calls to $DRAW(D[f])$ is a complete presentation of $f$ with probability 1.*

If $M$ is an inductive inference machine, then for any complete distribution $D$ on $N$ and function $f \in F[N,N]$, the probability that $M$ correctly EX-identifies $f$ in the limit when run with oracle $DRAW(D[f])$ for input is well-defined, and will be denoted

$$\Pr(EX(M, DRAW(D[f]))).$$

$EX_{draw(p)}(M)$ denotes the class of all functions $f \in F[N,N]$ such that for every complete distribution $D$ on $N$,

$$\Pr(EX(M, DRAW(D[f]))) \geq p.$$

Consistent with the notation used by Pitt [11], we denote by $EX_{draw(p)}$ the collection of all classes $C \subseteq F[N,N]$ such that for some inductive inference machine $M$, $C \subseteq EX_{draw(p)}(M)$. The main theorem of this section is the following.

**Theorem 6** *For all $p \in [0,1]$, $EX_{draw(p)} = EX_{prob(p)}$.*

*Proof.* Suppose that $C \in EX_{prob(p)}$. Then for some probabilistic inductive inference machine $M$, $C \subseteq EX_{prob(p)}(M)$. We define from $M$ a deterministic inductive inference machine $M'$ such that $C \subseteq EX_{draw(p)}(M')$.

$M'$ works by simulating $M$. Whenever $M$ requests an input, $M'$ requests an input and gives it to $M$. Whenever $M$ supplies an output, $M'$ outputs the same value. The remaining problem is to simulate the coin tosses used by $M$, which is done by using the stochastic input.

Whenever $M$ would make a coin flip, $M'$ instead requests two inputs, say $\langle x_i, y_i\rangle$, and $\langle x_{i+1}, y_{i+1}\rangle$. If $x_i$ is 1 and $x_{i+1}$ is 2, $M'$ continues simulating $M$ as though the value read from the coin tape were H. If $x_i$ is 2 and $x_{i+1}$ is 1, $M'$ continues simulating $M$ as though the value read from the coin tape were T. For any other outcome, $M'$ repeats the process, requesting another two inputs, and checking them, until one of the two outcomes above occurs.

Suppose $D$ is any complete distribution on $N$, and $f$ is any function in $C$. What happens when we run $M'$ with oracle $DRAW(D[f])$? Lemma 5 shows that $M$ will be simulated with a complete presentation for $L$ with probability 1. Since $D$ is complete, $D(1) > 0$ and $D(2) > 0$, so the probability is 0 that $M'$ will get "stuck" trying to determine a coin flip for $M$. Moreover, separate calls to $DRAW(D[f])$ are independent events, so the probability of 1 followed by 2 is the same as 2 followed by 1.

Thus, with probability 1 we will have a correct simulation of the probabilistic machine $M$ on a complete presentation of $f$. Hence, the probability that $M'$ correctly identifies $f$ in the limit is at least $p$. Thus, $C \in EX_{draw(p)}$, and $EX_{prob(p)} \subseteq EX_{draw(p)}$.

Conversely, suppose $C$ is in $EX_{draw(p)}$. Then there is some inductive inference machine $M$ such that $C \subseteq EX_{draw(p)}(M)$. We use $M$ to construct a probabilistic inductive inference machine $M'$ such that $C \subseteq EX_{prob(p)}(M')$.

$M'$ simulates $M$, using the procedure in Example 4 to simulate a call to the oracle $DRAW(D_\sigma)$ whenever $M$ requests an input, where $\sigma$ is the sequence of inputs to $M'$. Whenever $M$ produces an output, $M'$ outputs the same value.

Let $f$ be any function from $C$, and let

$$\sigma = \langle x_0, y_0\rangle, \langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots$$

be any complete presentation of $f$. It is not difficult to check that the distribution $D_\sigma$ is equal to the distribution $D_{\sigma'}[f]$, where

$$\sigma' = x_0, x_1, x_2, \ldots .$$

Thus, with probability 1, $M'$ succeeds in simulating $M$ with oracle $DRAW(D_{\sigma'}[f])$. Moreover, $D_{\sigma'}$ is a complete distribution on $N$ because $\sigma$ is a complete presentation of $f$.

Since $\Pr(EX(M, DRAW(D[f]))) \geq p$ for any complete distribution $D$ on $N$, the probability that $M'$ identifies $f$ is at least $p$. Hence $C \in EX_{prob(p)}$, and $EX_{draw(p)} \subseteq EX_{prob(p)}$, concluding the proof of Theorem 6. □

An immediate corollary of this and Theorem 1 is:

**Corollary 7** *For each $p > 1/2$, $EX_{draw(p)} = EX$. $EX$ is a proper subset of $EX_{draw(1/2)}$.*

That is, if an inductive inference machine can identify in the limit a class of sets with probability greater than 1/2 from complete stochastic input, then there is another inductive inference machine that can deterministically identify that class in the limit from complete (non-stochastic) presentations. Hence, the assumption of complete stochastic rather than complete input is no help, if we require correct convergence with any probability exceeding 1/2. However, if we are willing to accept probabilities less than or equal to 1/2, there is an enlargement of the collection of identifiable classes.

The above construction incidentally shows that probabilistic inductive inference machines with stochastic input have no greater power than deterministic ones with stochastic input, since the stochastic input can be used to simulate coin tosses as indicated.

### 3.2 Positive samples

Now we turn to the case of TXTEX-identification and show an analogous result. A distribution $D$ on $N \cup \{*\}$ is defined to be *admissible for* a language $L$ if and only if $S(D) - \{*\} = L$. That is, the language consists of just those elements of $N$ with positive probability.

Again, the following is immmediate.

**Lemma 8** *If $L$ is any language and $D$ is any distribution on $N \cup \{*\}$ admissible for $L$, then the values returned by an infinite sequence of calls to $DRAW(D)$ will be a positive presentation of $L$ with probability 1.*

If $M$ is an inductive inference machine, then the probability that $M$ TXTEX-identifies $L$ when run with inputs drawn from $DRAW(D)$ is well-defined and will be denoted

$$\Pr(TXTEX(M, DRAW(D))).$$

Then $M$ *TXTEX-identifies L in the limit with probability p from positive stochastic input* if and only if for every distribution $D$ on $N \cup \{*\}$ that is admissible for $L$,

$$\Pr(TXTEX(M, DRAW(D))) \geq p.$$

$TXTEX_{draw(p)}(M)$ is the collection of all languages $L$ such that $M$ TXTEX-identifies $L$ in the limit with probability $p$ from positive stochastic input. $TXTEX_{draw(p)}$ is the set of all classes $C$ such that for some inductive inference machine $M$, $C \subseteq TXTEX_{draw(p)}(M)$.

The main theorem of this section is:

**Theorem 9** *For all $p \in [0,1]$, $TXTEX_{draw(p)} = TXTEX_{prob(p)}$.*

*Proof.* The proof is essentially the same as the proof of Theorem 6, except that a slightly more complicated method is required to simulate coin tosses.

Suppose $C$ is in $TXTEX_{prob(p)}$. Let $M$ be a probabilistic inductive inference machine such that $C \subseteq TXTEX_{prob(p)}(M)$. We define a deterministic inductive inference machine $M'$ such that $C \subseteq TXTEX_{draw(p)}(M')$.

$M'$ reads the first input, $x_0$. If $x_0 = *$, then $M'$ outputs any index for the empty set. Otherwise, $M'$ outputs any index for the set $\{x_0\}$. $M'$ continues to read inputs until (if ever), some input $x_n \neq x_0$ is read. If and when this happens, $M'$ proceeds as follows.

$M'$ begins simulating $M$. When $M$ requests an input, $M'$ reads an input and gives it to $M$. When $M$ outputs a value, $M'$ outputs the same value. When $M$ requests a coin-flip, $M'$ reads pairs of inputs until some pair consists of $x_0$ followed by $x_n$ or vice versa. $M'$ then continues the simulation as though H were read from the coin tape in the first case, T in the second case.

Let $L$ be any element of $C$, and $D$ any distribution on $N \cup \{*\}$ admissible for $L$. Consider what happens when $M'$ is run using calls to $DRAW(D)$ as input. If $L$ is the empty set, then all the calls will return $*$, and $M'$ will correctly converge to an index for the empty set. If $L$ is a singleton set $\{x\}$, then it may happen that $D$ is the distribution that assigns 1 to $x$ and 0 to all other elements of $N \cup \{*\}$. In that case, $M'$ will correctly converge to an index for the singleton set $\{x\}$.

In all other cases, there are at least two different values of $x$ with $D(x) > 0$, so with probability 1, the search for $x_n \neq x_0$ will succeed, and the simulation of $M$ will commence. Since $D(x_0) > 0$, $D(x_n) > 0$, and successive calls to $DRAW(D)$ are statistically independent, the simulation of the coin-tosses will succeed with probability 1, and the two outcomes will have independent probability 1/2 at each flip, as required. Also, if $M$ reads an infinite sequence of inputs, it will be a positive presentation of $L$ with probability 1, by Lemma 8. Hence, with probability 1 the simulation of $M$ will be correct, so $M'$ will TXTEX-identify $L$ with probability $p$. Thus, $C \in TXTEX_{draw(p)}$, and $TXTEX_{prob(p)} \subseteq TXTEX_{draw(p)}$.

For the converse, suppose $C$ is in $TXTEX_{draw(p)}$. Let $M$ be a deterministic inductive inference machine such that $C \subseteq TXTEX_{draw(p)}(M)$. We define a probabilistic inductive inference machine $M'$ such that $C \subseteq TXTEX_{prob(p)}(M')$.

$M'$ simulates $M$, using the procedure in Example 4 to simulate calls to $DRAW(D_\sigma)$ when $M$ requests inputs, where $\sigma$ is the sequence of inputs to $M'$. When $M$ produces an output, $M'$ outputs the same value.

Suppose $L$ is in $C$ and $\sigma$ is any positive presentation of $L$. On input $\sigma$, $M'$ will succeed in simulating $M$ with oracle $DRAW(D_\sigma)$ with probability 1. Since $D_\sigma$ is a distribution on $N \cup \{*\}$ that is admissible for $L$, $M$ must TXTEX-identify $L$ with probability $p$. Hence $M'$ must TXTEX-identify $L$ with probability $p$, so $C \in TXTEX_{prob(p)}$ and $TXTEX_{draw(p)} \subseteq TXTEX_{prob(p)}$, concluding the proof of Theorem 9. □

An immediate corollary of this theorem and Theorem 2 is:

**Corollary 10** *For each $p > 2/3$, $TXTEX_{draw(p)} = TXTEX$. $TXTEX$ is a proper subset of $TXTEX_{draw(1/2)}$.*

That is, if an inductive inference machine can identify in the limit a class of sets with probability greater than 2/3 from positive stochastic input, then there is another inductive inference machine that can deterministically identify that class in the limit from positive (non-stochastic) presentations. Hence, the assumption of positive stochastic rather than positive input is no help, if we require correct convergence with any probability exceeding 2/3. If we are willing to accept probabilities less than or equal to 1/2, more classes of sets become identifiable. The open problem of the gap between 1/2 and 2/3 was mentioned in connection with Theorem 2.

### 3.3 Remarks on the distribution-free case

It is clear that the techniques of Theorems 6 and 9 extend to many other reasonable criteria of identification, though it does not seem worth the trouble of formalizing such a result in this paper. In particular, the extension to behaviorally correct identification (denoted BC-identification) is immediate, and Pitt [11] has results analogous to the $EX_{prob(p)}$ case characterizing the $BC_{prob(p)}$ classes.

Likewise, the extension of Theorem 9 to TXTBC-identification is immediate. For TXTBC-identification, Pitt [11] has the stronger result

**Theorem 11** *For each $p > 1/2$, $TXTBC_{prob(p)} = TXTBC$. $TXTBC$ is a proper subset of $TXTBC_{prob(1/2)}$.*

In fact, if $C$ is the class consisting of all finite subsets of $N$ together with $N$ itself, then it is easy to see that $C \in TXTEX_{prob(1/2)}$. However, Case and Lynes [3] show that $C \notin TXTBC$.

If we restrict the functions considered in Theorem 6 to have only values of 0 and 1, then this is the case of identifying recursive languages by means of complete presentations, where the hypotheses are required to be decision-rules. It is clear that the proof is unchanged in this case, and Corollary 7 also holds in this case.

There are two more possible combinations of input and hypotheses: complete presentations and enumerators, and positive presentations and decision-rules. It is clear that theorems analogous to Theorems 6 and 9 hold in this case also, since the form of the hypothesis is not critical in either case. However, Pitt does not consider these cases.

It may be possible to extend Pitt's results to these two cases, since the key lemma, that an inductive inference machine that outputs a finite set of indices at least one of which is correct can be converted to a machine that does EX-identification, holds in these two cases as well. (This lemma is false in the case of positive presentations and enumerators, even if all the hypotheses that are output infinitely often are correct, as shown by Case [2].)

In the case of complete presentations and enumerators, the new machine outputs a hypothesis that represents the union of the sets enumerated by the finite set of hypotheses, removing any hypothesis that enumerates an element $x$ such that $\langle x, 0\rangle$ has appeared in the input sequence. Eventually all the remaining hypotheses will enumerate subsets of the correct language, and at least one of them is correct, so the union of them is correct and will not change.

In the case of positive presentations and decision-rules, consider first the problem of identifying an enumerator for the complement of the language. Construct from the finite set of hypotheses an enumerator that outputs $x$ whenever at least one hypothesis $\phi_i$ in the set gives $\phi_i(x) = 0$. Continue reading inputs, and discard from the finite set any hypothesis $\phi_i$ such that for some $x$ in the positive presentation, $\phi_i(x) = 0$. In the limit, all the hypotheses $\phi_i$ such that $\phi_i(x) = 0$ for some $x$ in the language being presented will be eliminated. Since at least one of the remaining hypotheses is correct, this process converges to a correct enumerator for the complement of the language being presented.

This enumerator, together with the positive presentation, give complete information about the language being presented. Thus, the final procedure assumes that the input, together with the enumerator constructed, furnish complete, correct information about the language, and executes the procedure for complete information. Every time the hypothesis for the enumerator of the complement is modified, the whole procedure for complete information is restarted. (Saving past positive information, of course.) Eventually the enumerator of the complement stabilizes correctly, and the procedure for complete information then synthesizes a correct decision-rule for the language in the limit.

As a final remark, if in the *EX* case the distributions are permitted to be incomplete, that is, to have $D(x) = 0$ for one or more elements of $x$, then examples may be constructed of sets of functions $S$ and distributions $D$ such that $S \in EX$, but no inductive inference machine $M$ can EX-identify $S$ with any probability $p > 0$ under the distribution $D$. As an example, consider the set of total self-identifying functions:

$$S = \{\phi_i : \phi_i(0) = i\},$$

and any distribution $D$ on $N$ such that $D(0) = 0$.

# 4 Identifying or approximating distributions

The results in the preceding section show that for EX and TXTEX-identification, the assumption of stochastic input is no more helpful than assuming the ability to toss coins, if no essential restrictions are placed on the distributions. However, in the cases of the positive results on identifying languages from stochastic positive presentations described in the introduction, strong assumptions are made about the possible distributions. For example, Horning [7] assumes that the possible distributions are given by stochastic context-free grammars with rational probabilities on the productions. Thus, they are enumerable and computable in a useful sense.

The primary result in this part of the paper is an algorithm for identifying or approximating distributions using a general definition of a computable sequence of distributions. This will be shown to generalize the results described in the introduction.

## 4.1 Motivation: Identifying coins

Before we begin on the general problem, we illustrate our identification task for biases of coins. Suppose there is a coin with a fixed unknown probability $p$ of coming up heads, and probability $q = 1 - p$ of coming up tails. Then $p$ is called the *bias* of the coin.

The usual procedure of statistical estimation is to toss the coin $n$ times, record the number, $S_n$, of times it comes up heads, and then to estimate $p$ by $S_n/n$. Bounds on the tails of the binomial distribution can then be used to estimate the probability of an error of a certain magnitude in this estimate.

Suppose instead that the unknown coin is one of two coins, $C_1$ and $C_2$, with known biases $p_1$ and $p_2$, where $p_1 \neq p_2$. Suppose also that our task is to make an infinite sequence of guesses of whether the unknown coin is $C_1$ or $C_2$, and all but a finite number of these guesses are required to be correct.

One simple algorithm is as follows. Initialize $H = 0$. $H$ is a running total of the number of heads seen so far. At the $n^{th}$ trial, flip the unknown coin once and set $H = H + 1$ if it comes up heads. Guess $C_1$ if $|p_1 - H/n| \leq |p_2 - H/n|$, and guess $C_2$ otherwise.

To see that this algorithm solves the limit identification problem, we note that the law of the iterated logarithm implies that for any $\lambda > 1$, with probability 1 there are only finitely many values of $n$ for which [1]

$$|p - S_n/n| > \lambda\sqrt{(2pq \log\log n)/n},$$

where $S_n$ is the number of heads in the first $n$ of an infinite sequence of tosses of a coin with bias $p$ [5]. Since $pq$ is at most $1/4$, and $\lambda = \sqrt{2}$ is sufficient, with probability 1 there are only finitely many $n$ for which

$$|p - S_n/n| > \sqrt{(\log\log n)/n}.$$

Thus if the unknown coin is $C_1$, then with probability 1

$$|p_1 - H/n| \leq \sqrt{(\log\log n)/n} < |p_1 - p_2|/2$$

for all sufficiently large $n$. That is, with probability 1, $C_1$ will be the guess for all sufficiently large $n$. (Similarly if the unknown coin is $C_2$.)

To push the problem a little further, suppose now we have an infinite set of possible choices for the unknown bias, say $1/2$ and all numbers of the form $1/2 - 1/2^i$ or $1/2 + 1/2^i$ for $i = 1, 2, \ldots$. That is, the possible biases are

$$1/2,\ 0,\ 1,\ 1/4,\ 3/4,\ 3/8,\ 5/8,\ 7/16,\ 9/16,\ \ldots\ .$$

The goal is still to output an infinite sequence of guesses of the bias, of which all but a finite number are correct.

The approach of choosing the bias closest to the empirical estimate $H/n$ for samples of increasing size breaks down. For example, if the unknown bias is $1/2$, then with probability 1 the estimate of $p$ will be different from $1/2$ infinitely often, and the approach of choosing the closest bias will be incorrect each time.

In this case, we number the possible biases in some order, $p_1, p_2, p_3, \ldots$, and proceed as follows. Initialize $H = 0$. At the $n^{th}$ trial, flip the unknown coin once and set $H = H + 1$ if it comes up heads. Let $i \leq n$ be the least positive integer such that

$$|p_i - H/n| \leq \sqrt{(\log\log n)/n}.$$

[1] "log" denotes the logarithm to the base $e$ in this paper.

If such an $i$ is found, output $p_i$; otherwise, output any $p_j$.

To see that this process identifies the unknown bias $p$ in the limit with probability 1, we argue as follows. Let $k$ be the least positive integer such that $p_k = p$. (We are guaranteed that there is at least one such $k$.) Then the law of the iterated logarithm implies that with probability 1,

$$|p_k - H/n| \leq \sqrt{(\log \log n)/n}$$

for all but finitely many $n$. Thus, once $n$ exceeds $k$, $p_k$ will be a candidate for output all but finitely many times, with probability 1.

Now consider any $p_i$ such that $i < k$. Since $p_i \neq p$, let $\epsilon_i = |p_i - p| > 0$. Then

$$\epsilon_i = |p_i - p_k| \leq |p_i - H/n| + |p_k - H/n|,$$

so

$$|p_i - H/n| \geq \epsilon_i - \sqrt{(\log \log n)/n}$$

for all but finitely many $n$, with probability 1. Since the right-hand side exceeds

$$\sqrt{(\log \log n)/n}$$

for all sufficiently large $n$, we see that $p_i$ can be a candidate for output only finitely many times, with probability 1. Hence, with probability 1, the procedure will output $p_k$ for all sufficiently large $n$.

Suppose the bias of the unknown coin is not in the set of possible biases, e.g., $p = 4/9$ in the example above. Then this procedure will, with probability 1, produce an arbitrary sequence of guesses. But in this case there is a single closest hypothesis, namely 7/16, and it would be reasonable to require the identification procedure to converge to this value. We show how to accomplish this stronger requirement in the general case.

## 4.2 Approximately computable distributions

In general we are interested in identifying or approximating one of a countable sequence $D_0, D_1, D_2, \ldots$ of distributions. For the application to language identification, we use distributions on $N \cup \{*\}$. It will simplify notation if we consider distributions on $N$ in this and subsequent sections, except as noted. In fact, by a straightforward coding, the results are the same for distributions on $N \cup \{*\}$ or any other domain recursively related to $N$.

We need a computability condition on this sequence of distributions. For this purpose, we assume that the rational numbers are coded in some appropriate way (e.g., as pairs of integers), so that we can speak of computations with rational numbers as inputs and outputs.

A distribution $D$ on $N$ is said to be *approximately computable* if and only if there is a total recursive function $f$ such that for every $x \in N$ and every positive rational number $\epsilon$, $f(x, \epsilon)$ is a rational number $r$ such that

$$|D(x) - r| \leq \epsilon.$$

That is, $f(x, \epsilon)$ is a rational approximation of $D(x)$ to within $\epsilon$. Note that the values of $D(x)$ need not be rational.

The sequence of distributions $D_0$, $D_1$, $D_2$, ... is said to be *uniformly approximately computable* if and only if there is a total recursive function $C(i, x, \epsilon)$ such that for every $i \in N$, every $x \in N$, and every positive rational number $\epsilon > 0$, the value of $C(i, x, \epsilon)$ is a rational number $p$ such that

$$|D_i(x) - p| \leq \epsilon.$$

That is, $C(i, x, \epsilon)$ is a rational approximation of $D_i(x)$ to within distance $\epsilon$.

This generalizes the computability condition used by Osherson, Stob, and Weinstein [10], which requires the value of $D_i(x)$ to be exactly computable from $i$ and $x$. Under their definition it is decidable from $i$ and $x$ whether $x \in S(D_i)$.

**Example 12** *We define a uniformly approximately computable sequence of distributions $E_0$, $E_1$, $E_2$, ... on $N \cup \{*\}$ such that for every $i \in N$, $L(E_i)$ is the recursively enumerable set $W_i$. There is an effective process $P$ such that on input $i$, $P(i)$ enumerates a positive presentation $\sigma_i$ of $W_i$. In Example 3 we defined a specific distribution $D_\sigma$ from any infinite sequence $\sigma$. For each $i \in N$, and each $x \in N$, let*

$$E_i(x) = D_{\sigma_i}(x).$$

By the definition of $D_\sigma$, we have that $L(E_i) = W_i$ for all $i \in N$. To see that this sequence of distributions is uniformly approximately computable, we indicate how to compute an appropriate $C(i, x, \epsilon)$. Given a positive rational number $\epsilon$, let $n$ be sufficiently large that $1/2^n \leq \epsilon$. Use $P(i)$ to enumerate the first $n$ elements

$$x_0, x_1, \ldots x_{n-1}$$

of the positive presentation $\sigma_i$ of $W_i$. Let

$$I_n(x) = \{i : 0 \leq i \leq n - 1 \text{ and } x_i = x\}.$$

This is the set of indices of $x$ among the first $n$ elements of $\sigma_i$. Then return the rational number

$$E_{i,n}(x) = \sum_{i \in I_n(x)} 1/2^{i+1},$$

where the sum over an empty set of indices is interpreted as 0. The probabilities attached to all occurrences of elements past the $n^{th}$ sum to $1/2^n$, so $E_{i,n}(x)$ is an approximation within distance $1/2^n \leq \epsilon$ of $E_i(x)$, as required.

### 4.3 Distance measures for distributions

To generalize the approach described above for identifying the bias of a coin, we need an appropriate generalization of the measure $|p_1 - p_2|$ of the "difference" between two coins. A variety of functions would work; the one we choose is

$$d_*(D_1, D_2) = \sup\{|D_1(x) - D_2(x)| : x \in N\}.$$

We observe that the value $d_*(D_1, D_2)$ is attained for some $x \in N$. It is obvious that

$$d_*(D_1, D_2) = d_*(D_2, D_1)$$

for all distributions $D_1$ and $D_2$. Also,

$$d_*(D_1, D_2) = 0 \text{ if and only if } D_1 = D_2.$$

From the corresponding triangle inequality for $|x - y|$ it is easy to show that

$$d_*(D_1, D_3) \leq d_*(D_1, D_2) + d_*(D_2, D_3).$$

Taken together, these three properties show that $d_*(D_1, D_2)$ is a metric on the space of distributions on $N$.

We now formulate a computability condition on such metrics. An oracle $X$ *represents* a distribution $D$ on $N$ if and only if whenever $X$ is called with $x \in N$ and a positive rational number $\epsilon$, the output of $X$ is a rational number $p$ such that

$$|p - D(x)| \leq \epsilon.$$

Let $d(D_1, D_2)$ be a metric on the space of distributions on $N$. Then $d(D_1, D_2)$ is *approximately computable* if and only if there is a Turing machine $M^{X,Y}(\epsilon)$ that calls on two oracles $X$ and $Y$, and is such that whenever $\epsilon$ is a positive rational number and $D_1$ and $D_2$ are distributions on $N$ and $X$ represents $D_1$ and $Y$ represents $D_2$, the output of $M^{X,Y}(\epsilon)$ is a rational number $r$ such that

$$|r - d(D_1, D_2)| \leq \epsilon.$$

Then we have

**Lemma 13** *$d_*(D_1, D_2)$ is approximately computable.*

Before we begin the proof of this lemma, we show that given a positive rational number $\epsilon$ and an oracle $X$ representing a distribution $D$ on $N$, we can compute a finite approximation *MIN* of $D$ to within $\epsilon$.

*MIN* will be a finite function with domain $S$. Initially $S$ is the empty set. We dovetail calls of the oracle $X$ with inputs $(x, \epsilon)$ for all $x \in N$ and all $\epsilon = 1/2^i$ for $i \in N$. Whenever a value $p$ is returned for $(x, \epsilon)$ such that

$$p - \epsilon > 0,$$

do the following. If $x \notin S$, add $x$ to $S$ and set $MIN(x) = p - \epsilon$. If $x \in S$, then set $MIN(x)$ to the maximum of $MIN(x)$ and $p - \epsilon$. Continue this process until

$$\sum_{x \in S} MIN(x) \geq 1 - \epsilon.$$

It is not difficult to show that if $X$ represents the distribution $D$ on $N$, then this process will eventually halt, and for all $x \in S$,

$$MIN(x) \leq D(x) \leq MIN(x) + \epsilon,$$

and for all $x \notin S$,

$$D(x) \leq \epsilon.$$

*Proof of Lemma 13.* Given $\epsilon$ and oracles $X$ and $Y$ representing distributions $D_1$ and $D_2$, we first use the procedure above to compute finite functions $MIN_1$ and $MIN_2$, defined on finite domains $S_1$ and $S_2$, such that for all $x \in S_1$,

$$MIN_1(x) \leq D_1(x) \leq MIN_1(x) + \epsilon/2,$$

and for all $x \notin S_1$,

$$D_1(x) \leq \epsilon/2,$$

and similarly for $MIN_2$ and $D_2$.

Extend $MIN_1$ to domain $N$ by defining

$$MIN'_1(x) = 0 \text{ if } x \notin S_1,$$

and similarly extend $MIN_2$ to $MIN'_2$ with domain $N$. Then output

$$\max\{|MIN'_1(x) - MIN'_2(x)| : x \in N\}.$$

This is clearly computable, since $S_1 \cup S_2$ is finite, and both $MIN'_1$ and $MIN'_2$ are 0 outside $S_1 \cup S_2$. To see that this value is an approximation of $d_*(D_1, D_2)$ to within $\epsilon$, note that for all $x \in N$,

$$|D_1(x) - MIN'_1(x)| \leq \epsilon/2,$$

and

$$|D_2(x) - MIN'_2(x)| \leq \epsilon/2.$$

Thus, for all $x \in S$,

$$||D_1(x) - D_2(x)| - |MIN'_1(x) - MIN'_2(x)|| \leq \epsilon.$$

Hence, the maximum value of $|MIN'_1(x) - MIN'_2(x)|$ will be within $\epsilon$ of $d_*(D_1, D_2)$. This concludes the proof of Lemma 13. □

The final property that we require of $d_*(D_1, D_2)$ is a bound on how well the empirical distribution after $n$ samples approximates the true distribution. Let $D$ be any distribution on $N$. The experiment we consider is drawing an infinite sequence of examples from $DRAW(D)$, say

$$x_0, x_1, x_2, \ldots .$$

We'll use $D\langle n \rangle$ to denote the empirical distribution after drawing $n$ samples, for $n \geq 1$. That is, let

$$I_n(x) = \{0 \leq i \leq n-1 : x_i = x\},$$

and let

$$D\langle n \rangle(x) = |I_n(x)|/n,$$

for all $x \in N$. Then $D\langle n \rangle(x)$ is the frequency of occurrence of $x$ among the first $n$ samples.

**Lemma 14** *Let $D$ be any distribution on $N$. Let $a > 1$ and let*

$$I(n) = \sqrt{6a(\log n)/n}.$$

*Then, with probability 1,*

$$d_*(D, D\langle n\rangle) \leq I(n)$$

*for all but finitely many values of $n$.*

This lemma is proved in the Appendix, where we also discuss its relationship to the Kolmogorov-Smirnov test. To summarize the properties of $d_*(D_1, D_2)$ we shall need for our identification algorithm, it suffices if $d(D_1, D_2)$ is an approximately computable metric on the space of all distributions on $N$ such that there exists a non-zero, approximately computable bound $b(n) \to 0$ as $n \to \infty$ such that for any distribution $D$ on $N$, the probability is 1 that $d(D, D\langle n\rangle) \leq b(n)$ for all but finitely many values of $n$. (For clarity we omit the details of dealing with the approximation to $b(n)$, and assume it is exactly computable.)

### 4.4 Our criteria of identification

Let

$$\Delta = D_0, D_1, D_2, \ldots$$

be a sequence of distributions on $N$. If $D$ is any distribution on $N$, then let

$$\mathit{approach}_\Delta(D) = \inf\{d_*(D, D_i) : i \in N\}.$$

If $D = D_i$ for some $i \in N$, then $\mathit{approach}_\Delta(D) = 0$, but the converse is not necessarily true.

If for $D$ there exists some $D_i$ such that $d_*(D, D_i) = \mathit{approach}_\Delta(D)$, then we'll say that $D$ is *finitely approachable by* $\Delta$. If $\mathit{approach}_\Delta(D) = 0$ and $D$ is finitely approachable by $\Delta$, then $D$ is equal to some $D_i$.

Let $M$ be an inductive inference machine. If $D$ is any distribution on $N$, let

$$M[\mathit{DRAW}(D)]$$

denote the output of $M$ run with calls to the $\mathit{DRAW}(D)$ oracle as input.

We will say that $M$ *EX-identifies* the distribution $D$ if and only if the probability is 1 that $M[\mathit{DRAW}(D)]$ is a non-empty sequence of indices that converges to some $i$ such that $D = D_i$. Note that $D$ must be in the original sequence of distributions in this case.

We will say that $M$ *EX-approaches* the distribution $D$ if and only if the probability is 1 that $M[\mathit{DRAW}(D)]$ is a non-empty sequence of indices that converges to some $i$ such that $d_*(D, D_i) = \mathit{approach}_\Delta(D)$. Note in particular that $D$ must be finitely approachable by $\Delta$ in this case. Note also that if $D$ is in the original sequence of distributions and $M$ EX-approaches $D$, then $M$ must EX-identify $D$.

The third definition does not require that the sequence of indices converge. We will say that $M$ *$\Delta$-approaches* the distribution $D$ if and only if with probability 1, $M[\mathit{DRAW}(D)]$ is an infinite sequence of indices $j_0, j_1, j_2, \ldots$ such that

$$\lim_{n\to\infty} d_*(D, D_{j_n}) = \mathit{approach}_\Delta(D).$$

### 4.5 EX-identifying distributions

**Theorem 15** *Let $\Delta = D_0, D_1, D_2, \ldots$ be a uniformly approximately computable sequence of distributions on $N$. There exists an inductive inference machine $M$ that EX-identifies any $D_i$ from $\Delta$.*

*Proof.* We describe and analyze an inductive inference machine $M$ that uses $\Delta$. Recall the bound $I(n)$ defined in Lemma 14.

$M$ works in stages $n = 1, 2, 3, \ldots$. At stage $n$, $M$ requests one more input, and forms the empirical distribution $D\langle n\rangle$. For each $i$, $0 \leq i \leq n-1$, $M$ approximates $d_*(D_i, D\langle n\rangle)$ to within $I(n)$. Let $e_i\langle n\rangle$ denote this approximation.

$M$ outputs the least $i < n$ such that

$$e_i\langle n\rangle \leq 2I(n)$$

if there is any such $i$. Then $M$ goes to stage $n+1$.

$\Delta$ is uniformly approximately computable by hypothesis, $d_*(D_1, D_2)$ is approximately computable by Lemma 13, and $D\langle n\rangle$ is exactly computable, therefore there is a recursive procedure to approximate $d_*(D_i, D\langle n\rangle)$ to within the bound $I(n)$. Thus, $M$ is an effective procedure.

To see that $M$ behaves correctly, suppose that $D$ is any distribution from $\Delta$. Let $i$ be the least index such that $D_i = D$. Then for all $j < i$, $d_*(D_j, D) > 0$. Let $N_0$ be sufficiently large that

$$d_*(D_j, D) \geq 5I(n)$$

for all $n \geq N_0$ and for all $j < i$.

Suppose the sequence of inputs drawn from $DRAW(D)$ is such that $d_*(D, D\langle n\rangle) \leq I(n)$ for all but finitely many $n$. (This occurs with probability 1 by Lemma 14.) Let $N_1$ be sufficiently large that

$$d_*(D, D\langle n\rangle) \leq I(n)$$

for all $n \geq N_1$.

Let $N_2$ be the maximum of $i+1$, $N_0$, and $N_1$. Then for all $n \geq N_2$,

$$|e_i\langle n\rangle - d_*(D_i, D\langle n\rangle)| \leq I(n),$$

and because $D_i = D$,

$$|d_*(D_i, D\langle n\rangle)| \leq I(n),$$

so

$$e_i\langle n\rangle \leq 2I(n).$$

Thus, $i$ will be a candidate to be output by $M$ for all $n \geq N_2$.

Suppose $j < i$ and $n \geq N_2$. By the triangle inequality, we have

$$d_*(D_j, D) \leq d_*(D_j, D\langle n\rangle) + d_*(D\langle n\rangle, D),$$

so

$$d_*(D_j, D\langle n\rangle) \geq d_*(D_j, D) - d_*(D\langle n\rangle, D).$$

However, since $n \geq N_0$,

$$d_*(D_j, D) \geq 5I(n),$$

and since $n \geq N_1$,

$$d_*(D, D\langle n\rangle) \leq I(n).$$

Thus,

$$d_*(D_j, D\langle n\rangle) \geq 4I(n).$$

Because $e_j\langle n\rangle$ is an approximation of $d_*(D_j, D\langle n\rangle)$ to within $I(n)$, we have

$$e_j\langle n\rangle \geq 3I(n),$$

so for all $n \geq N_2$ and all $j < i$, $j$ will not be a candidate for output at stage $n$ of $M$.

Thus $i$ will be the output of $M$ for all stages $n \geq N$. Since this case occurs with probability 1 by Lemma 14, $M$ EX-identifies $D$. □

This result improves on the limit identification results of Osherson, Stob, and Weinstein [10] in that our requirement of computability for distributions is weaker. In particular, Example 12 shows that there is a uniformly approximately computable sequence of distributions whose associated languages are all the recursively enumerable sets. However, the class of all recursively enumerable sets is not "uniformly measurable" in Osherson, Stob, and Weinstein's definition.

The following corollary concerns the class of stochastic context-free grammars, considered by Horning [7].

**Corollary 16** *Let $\Sigma$ be any fixed finite alphabet. The class of distributions over $\Sigma^*$ represented by non-blocking stochastic context-free grammars with terminal alphabet $\Sigma$ and rational probabilities can be EX-identified.*

*Proof.* Construct some computable listing of all such grammars, $G_0$, $G_1$, $G_2$, … . (Note that the property of being non-blocking is recursively decidable.) We need to see that the probability that grammar $G_i$ generates some string $y$ can be approximated to within any $\epsilon$.

Initially let $W$ be the empty set. $MIN$ will be a function whose domain is $W$. Enumerate all leftmost derivations from the start symbol in $G_i$ in order of increasing length. Whenever a derivation has as a last element a terminal string $w$, do the following. Let $p$ be the probability of the derivation, and if $w \in W$ then set $MIN(w) = MIN(w) + p$. If $w \notin W$, then add $w$ to $W$ and set $MIN(w) = p$.

The sum of $MIN(w)$ for $w \in W$ is a non-decreasing function whose limit is 1 because $G_i$ is non-blocking. Continue the process above until $MIN(w) \geq 1 - \epsilon$. Then return the value $MIN(y)$ if $y \in W$, and the value 0 otherwise. It is clear that this will be an approximation to the probability of generating $y$ from $G_i$ to within $\epsilon$. Applying Theorem 15 concludes the proof of Corollary 16. □

Note that this improves on Horning's limiting result [7] in that the criterion of convergence is stronger and we permit ambiguous grammars. Ambiguous grammars do not cause problems for our procedure because we require that the probability of a string be only approximately computable.

## 4.6 EX-approaching distributions

In this section we show that a slightly more complex procedure can EX-approach any $D$ that is finitely approachable by $\Delta$. The key idea is that we must approximate $approach_\Delta(D)$ using the sample and the sequence of distributions, instead of assuming that $approach_\Delta(D)$ is zero.

**Theorem 17** *Let $\Delta = D_0, D_1, D_2, \ldots$ be a uniformly approximately computable sequence of distributions on $N$. Then there exists an inductive inference machine $M'$ such that for every distribution $D$ on $N$ that is finitely approachable by $\Delta$, $M'$ EX-approaches $D$. In particular, $M'$ EX-identifies every $D_i$ in $\Delta$.*

*Proof.* We describe an inductive inference machine $M'$ that uses $\Delta$. $M'$ works in stages $n = 1, 2, 3, \ldots$ . At stage $n$, $M'$ requests one more input, and forms the empirical distribution $D\langle n\rangle$. For each $i$, $0 \leq i \leq n-1$, $M'$ approximates $d_*(D_i, D\langle n\rangle)$ to within $I(n)$. Let $e_i\langle n\rangle$ denote this approximation.

Let $m\langle n\rangle$ denote the minimum value of $e_i\langle n\rangle$ for $0 \leq i \leq n-1$. $M'$ outputs the least $i < n$ such that

$$e_i\langle n\rangle \leq m\langle n\rangle + 4I(n).$$

Note that such an $i$ is guaranteed to exist in this case. Then $M'$ goes to stage $n+1$.

The argument that $M'$ is a computable procedure is the same as in the proof of Theorem 15. Let $D$ be any distribution on $N$ that is finitely approachable by $\Delta$. Consider what happens when $M'$ is run with inputs drawn from $DRAW(D)$.

Let $m = approach_\Delta(D)$ and let $i$ be the least index such that $d_*(D_i, D) = m$. Then for all $j < i$, $d_*(D_j, D) > m$. Let $N_0$ be sufficiently large that

$$d_*(D_j, D) \geq m + 9I(n)$$

for all $n \geq N_0$ and for all $j < i$.

Suppose the sequence of inputs drawn from $DRAW(D)$ is such that $d_*(D, D\langle n\rangle) \leq I(n)$ for all but finitely many $n$. (This occurs with probability 1 by Lemma 14.) Let $N_1$ be sufficiently large that

$$d_*(D, D\langle n\rangle) \leq I(n)$$

for all $n \geq N_1$.

Then for all $n \geq N_1$ and for all $0 \leq j \leq n-1$,

$$|e_j\langle n\rangle - d_*(D_j, D\langle n\rangle)| \leq I(n).$$

Moreover,

$$d_*(D, D\langle n\rangle) \leq I(n),$$

so

$$|e_j\langle n\rangle - d_*(D_j, D)| \leq 2I(n).$$

Thus, $e_j\langle n\rangle$ is an approximation of $d_*(D_j, D)$ to within $2I(n)$. Hence the value of $m\langle n\rangle$ will be an approximation to the minimum value of $d_*(D_j, D)$ for $0 \leq j \leq n-1$ to within $2I(n)$.

Let $N_2$ be the maximum of $i+1$, $N_0$, and $N_1$, and let $n \geq N_2$. Since $n > i$, $m$ is the minimum value of $d_*(D_j, D)$ for $0 \leq j \leq n-1$. Then $m\langle n\rangle$ is an approximation to $m$ within $2I(n)$. Likewise, $e_i\langle n\rangle$ is an approximation of $m = d_*(D_i, D)$ to within $2I(n)$, so $e_i\langle n\rangle$ is within $4I(n)$ of $m\langle n\rangle$. Thus, $i$ is a candidate for output at stage $n$.

Consider any $j < i$. Since $n \geq N_0$,

$$d_*(D_j, D) \geq m + 9I(n).$$

However, $e_j\langle n\rangle$ is an approximation of $d_*(D_j, D)$ to within $2I(n)$, so

$$e_j\langle n\rangle \geq m + 7I(n).$$

Since $m\langle n\rangle$ is an approximation to $m$ within $2I(n)$,

$$e_j\langle n\rangle \geq m\langle n\rangle + 5I(n).$$

Hence, $j$ is not a candidate for output at stage $n$.

Thus, for all stages $n \geq N_2$, $i$ is the output of $M'$. This case occurs with probability 1 by Lemma 14, so $M'$ EX-approaches $D$. Since $D$ is an arbitrary distribution on $N$ that is finitely approachable by $\Delta$, this concludes the proof of Theorem 17. It is not difficult to see that for an arbitrary distribution $D$ on $N$, $M'$ $\Delta$-approaches $D$. □

# 5 Remarks

We have investigated the effect of assuming randomly drawn examples on various types of limiting identification. In the distribution-free case, stochastic input reduces to the case of probabilistic algorithms. For several types of identification criteria (e.g., EX, BC, TXTEX, TXTBC), Pitt has shown that assuming probabilistic identification algorithms does not enlarge the identification type if the algorithm is required to be correct with probability 1.

However, if the distributions are assumed to come from a uniformly approximately computable sequence of distributions, there is an algorithm to EX-identify them in the limit. This general result explains the success of Horning [7], Van der Mude and Walker [14], and Osherson, Stob, and Weinstein [10] in finding algorithms to identify languages from positive stochastic data.

It is interesting to compare this result for distributions with Gold's result [6] that all the recursively enumerable sets are identifiable in the limit from positive presentations if those positive presentations are required to be generated by primitive recursive functions. The key to this result is to concentrate on modelling the functions presenting the text, which, being primitive recursive, are an enumerable class of total functions. Similarly, our result on identifying distributions concentrates on an enumerable, computationally tractable class of "generators", namely, uniformly approximately computable sequences of distributions. The analogy between these results suggests that there is great power in attempting to model "how" a behavior is produced, as well as "what" behavior is produced.

Rudich's result on identifying the structure of certain types of Markov chains in the limit [12] is of a different character from our results on distributions, since he has to overcome the problem of (in effect) not being told when the machine re-enters its start state.

What are the implications of these limiting results for the problems of finite, computationally tractable identification or learning? Valiant's paradigm is appropriate in general for positive and negative examples, but only for a few cases of positive-only examples, namely, domains in which the correct target concept can effectively be approached by a sequence of hypotheses each of which is a subset of the target concept. In this case, the "error" consists exclusively of examples in the target concept but not in the hypothesis.

It is an open problem to find a satisfactory generalization of Valiant's ideas to the general case of positive-only examples. Our results suggest that an appropriate generalization may have to abandon the "distribution-free" aspect of Valiant's paradigm, and explicitly model the relevant distributions. The central problem in such an approach is to define a useful notion of the "difference" between two distributions. Here the limiting case provides little guidance, since a wide variety of measures have appropriate limiting behavior.

## 6 Appendix

In this section we provide a proof of Lemma 14, and discuss its relation to the Kolmogorov-Smirnov test.

### 6.1 Proof of Lemma 14

First we need a technical lemma. Recall that "log" denotes the logarithm to the base $e$.

**Lemma 18** *For all $a$, $n$, and $p$ such that $a \geq 1$, $n \geq 2$, and $0 < p \leq 1$,*

$$e^{(-2a/p)\log n} < p^2.$$

*Proof.* For all $p$ such that $0 < p \leq 1$,

$$p\log(1/p) \leq 1/e.$$

For all $a \geq 1$ and $n \geq 2$,

$$a\log n > 1/e,$$

therefore,

$$a\log n > p\log(1/p).$$

Thus,

$$(-2a/p)\log n < 2\log p,$$

and therefore,

$$e^{(-2a/p)\log n} < p^2.$$

This concludes the proof of Lemma 18. □

*Proof of Lemma 14.* Let $D$ be any distribution on $N$. Let $a > 1$ and let

$$I(n) = \sqrt{6a(\log n)/n}.$$

We show that with probability 1,

$$d_*(D, D\langle n\rangle) \leq I(n)$$

for all but finitely many values of $n$, where $D\langle n\rangle$ is the empirical distribution defined by the first $n$ elements in an infinite sequence of examples drawn from $DRAW(D)$.

First, consider any $x$ such that $D(x) = 0$. This $x$ never occurs in any example sequence, so for all $n$, $D\langle n\rangle(x) = 0$. Hence, for all $n \geq 1$ and all $x \in N$ such that $D(x) = 0$, $|D(x) - D\langle n\rangle(x)| \leq I(n)$.

We use the following result, drawn from [1]. For each $x$ such that $D(x) > 0$, and any $\alpha$,

$$\Pr(|D(x) - D\langle n\rangle(x)| \geq \alpha) \leq 2e^{-\alpha^2 n/3D(x)}.$$

Substituting $\alpha = I(n)$, we have

$$\Pr(|D(x) - D\langle n\rangle(x)| \geq I(n)) \leq 2e^{(-2a/D(x))\log n}.$$

If $D(x) > 0$, then since $D(x) \leq 1$,

$$\Pr(|D(x) - D\langle n\rangle(x)| \geq I(n)) \leq 2n^{-2a}.$$

Let

$$B_n = \{x \in N : D(x) \geq n^{-a}\}.$$

Then $B_n$ contains at most $n^a$ elements, so the probability that $|D(x) - D\langle n\rangle(x)| > I(n)$ for some $x \in B_n$ is bounded by $2n^{-a}$.

Let

$$C_n = \{x \in N : D(x) < n^{-a}\}.$$

Applying Lemma 18 with $p = D(x)$, we have for every $n \geq 2$ and every $x$ such that $D(x) > 0$,

$$e^{(-2a/D(x))\log n} < (D(x))^2,$$

and therefore,

$$\Pr(|D(x) - D\langle n\rangle(x)| > I(n)) \leq 2(D(x))^2.$$

Also,

$$\sum_{x \in C_n} (D(x))^2 \leq n^{-a},$$

because $D(x) < n^{-a}$ for each $x \in C_n$, and $\sum_{x\in N} D(x) = 1$. Thus, for each $n \geq 2$, summing over all the elements in $C_n$, the probability that $|D(x) - D\langle n\rangle(x)| > I(n)$ for some $x \in C_n$ is bounded by $2n^{-a}$.

Hence, for each $n \geq 1$, the probability that $|D(x) - D\langle n\rangle(x)| > I(n)$ for some $x \in N$ is bounded by $4n^{-a}$. Since $d_*(D, D\langle n\rangle)$ is the maximum value of $|D(x) - D\langle n\rangle(x)|$ for all $x \in N$, this means that the probability that

$$d_*(D, D\langle n\rangle) > I(n)$$

is bounded by $4n^{-a}$. Because $a > 1$, the sum of these probabilities for all $n \geq 1$ is finite. By the Borel-Cantelli lemma, we conclude that with probability 1 there are only finitely many values of $n$ for which

$$d_*(D, D\langle n\rangle) > I(n),$$

which concludes the proof of Lemma 14. □

### 6.2 Comparison with the Kolmogorov-Smirnov test

In this section we must distinguish correctly between what we have called distributions above, which are really density functions, and the usual notion of distribution in probability theory. Let $F(x)$ be a continuous distribution function mapping the real numbers to $[0,1]$. $F(x)$ denotes the probability of drawing a number from the set of real numbers $\{y : y \leq x\}$.

Kolmogorov formulated a statistical test for whether a sequence of samples come from some known distribution $F(x)$. Smirnov formulated a similar test for whether two sequences of samples come from the same distribution. Our description of one kind of Kolmogorov-Smirnov test is drawn from Chung [4].

Consider an infinite sequence of random variables $x_0, x_1, x_2, \ldots$ each independently drawn according to the distribution $F(x)$. Let

$$F_n(x) = |\{i < n : x_i \leq x\}|/n.$$

Thus, $F_n(x)$ is the proportion of samples among the first $n$ that do not exceed $x$. That is, $F_n(x)$ is the empirical distribution defined from the first $n$ samples.

Let

$$k_n = \sup\{|F_n(x) - F(x)| : -\infty < x < \infty\}.$$

Theorem 2 from [4] shows that if $\lambda(n) \to \infty$ as $n \to \infty$ then the probability that

$$k_n > n^{-1/2}\lambda(n)$$

for infinitely many values of $n$ is 0 or 1 according to whether

$$\sum_{n=1}^{\infty}(\lambda^2(n)/n)e^{-2\lambda^2(n)}$$

is finite or infinite.

In particular, for $\lambda(n) = \sqrt{\log\log n}$ the sum converges, so with probability 1,

$$k_n > \sqrt{(\log\log n)/n}$$

for only finitely many values of $n$.

This suggests defining a metric on two density functions $D$ and $D'$ on $N$ as follows. Let

$$F(n) = \sum_{0 \leq m \leq n} D(m),$$

and

$$F'(n) = \sum_{0 \leq m \leq n} D'(n),$$

where $m$ and $n$ are natural numbers. Then define

$$d_K(D, D') = \sup\{|F(n) - F'(n)| : n \in N\}.$$

It is not difficult to see that $d_K(D, D')$ is an approximately computable metric on the space of all density functions on $N$.

We can extend $F(n)$ to be continuous distribution function on the real numbers by defining $F(x) = 0$ for all $x \leq -1$, and then making it piecewise linear from the point $(-1, 0)$ to the point $(0, F(0))$, piecewise linear from there to the point $(1, F(1))$, and so on. This corresponds to a probability density function that first selects the interval $(n-1, n]$ with probability $D(n)$, and then selects uniformly from that interval.

The extension agrees with $F(n)$ on all $n \in N$. Moreover, for any sequence of samples, the values of the empirical distribution for $F(n)$ and the empirical distribution for the extension agree on all $n \in N$. Hence, the Kolmogorov-Smirnov test gives us an analog of Lemma 14 for the metric $d_K$ (with a better bound in place of $I(n)$.) Other metrics, based on other statistical tests, are also possible.

## 7 Acknowledgements

The support of the National Science Foundation, through grant IRI-8404226, is gratefully acknowledged. The results on EX-identifying distributions were originally inspired by Phil Laird's results on identification in the limit in the presence of noise [8].

## References

[1] D. Angluin and L. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *J. Comput. Syst. Sci.*, 18:155–193, 1979.

[2] J. Case. The power of vacillation. Extended abstract, University of Rochester, 1987.

[3] J. Case and C. Lynes. Inductive inference and language identification. In *Proc. ICALP 82*, pages 107–115, Springer Verlag, 1982.

[4] K. Chung. An estimate concerning the Kolmogorov limit distribution. *Transactions of the American Mathematical Society*, 67:36–50, 1949.

[5] W. Feller. *An Introduction to Probability Theory and its Applications, Vol. I, Third Edition.* John Wiley & Sons, Inc., New York, 1968.

[6] E. M. Gold. Language identification in the limit. *Inform. Contr.*, 10:447–474, 1967.

[7] J. J. Horning. *A Study of Grammatical Inference.* PhD thesis, Stanford University, 1969.

[8] P. Laird. *Learning From Good Data and Bad.* PhD thesis, Yale University, 1987. Computer Science Dept. TR-551.

[9] M. Machtey and P. Young. *An Introduction to the General Theory of Algorithms.* North-Holland Publishing, 1978.

[10] D. Osherson, M. Stob, and S. Weinstein. *Systems That Learn.* MIT Press, Cambridge, MA, 1986.

[11] L. Pitt. *Probabilistic Inductive Inference.* PhD thesis, Yale University, 1985. Computer Science Dept. TR-400.

[12] S. Rudich. Inferring the structure of a Markov chain from its output. In *Proc. 26th IEEE Symposium on Foundations of Computer Science*, pages 321–326, IEEE, 1985.

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

[14] A. Van der Mude and A. Walker. On the inference of stochastic regular grammars. *Inform. and Contr.*, 38:310–329, 1978.

[15] K. Wexler and P. Culicover. *Formal Principles of Language Acquisition.* MIT Press, Cambridge, Mass., 1980.

[16] R. Wiehagen, R. Freivalds, and E. Kinber. On the power of probabilistic strategies in inductive inference. *Theoretical Computer Science*, 28:111–133, 1984.