Inductive-deductive systems: a mathematical logic and statistical learning perspective

Nicolas Baskiotis, Michèle Sebag, Olivier Teytaud

Inductive-deductive systems: a mathematical logic and statistical learning perspective

N. Baskiotis, O. Teytaud, M. Sebag
TAO-inria, LRI, UMR 8623(CNRS - Universite Paris-Sud),
bat 490 Universite Paris-Sud 91405 Orsay Cedex France
email: {nicolas.baskiotis, olivier.teytaud, michele.sebag}@lri.fr

October 31, 2010

Abstract

This paper has not been published with copyrighted proceedings and was only presented, in french, in a national conference.

The theorems about incompleteness of arithmetic have often been cited as an argument against automatic theorem proving and expert systems. However, these theorems rely on a worst-case analysis, which might happen to be overly pessimistic with respect to real-world domain applications. For this reason, a new framework for a probabilistic analysis of logical complexity is presented in this paper. Specifically, the rate of non-decidable clauses and the convergence of a set of axioms toward the target one when the latter exists in the language are studied, by combining results from mathematical logic and from statistical learning. Two theoretical settings are considered, where learning relies respectively on Turing oracles guessing the provability of a statement from a set of statements, and computable approximations thereof. Interestingly, both settings lead to similar results regarding the convergence rate towards completeness.

1 INTRODUCTION

Inspired by the "Learning to Reason" framework [Khardon and Roth, 1997], this paper investigates the conditions for a hybrid inductive-deductive system (IDS). This system is provided with a set of axioms or statements (e.g. examples), and its goal is to determine the truth value of any further statement e. We consider a framework for dealing with undecidable theories as well; this is a main difference with many previous works ([Shapiro, 1981]). We will often refer to arithmetic or set theory, but many other essentially undecidable theories could be considered instead of this. From a mathematical logic perspective, the question is whether i) the available set of statements is complete, and ii) the logical setting is complete. Under these assumptions, the truth value of e is determined using mathematical deduction; the algorithmic challenge is to provide an efficient search engine for constructing a proof of e or $\neg e$. When the set

of statements is not complete, by definition there exists statements e which can neither be proved nor refuted; the famous Gödel's theorem (1931) states that sufficiently powerful logical settings (e.g. including arithmetic) are incomplete. When the set of statements is not sufficient for deciding relevant statements, inductive reasoning is needed to find additional axioms, consistent with the available ones and sufficient for determining the truth value of e. The challenge here is to compare the different natural methods available for adding new axioms. From a hybrid inductive-deductive perspective, the logical setting considered must thus be examined with respect to both its completeness (deduction-oriented performances), and its VC-dimension or PAC learnability (induction-oriented performances, Appendix A.2). Typically, statements $C(1)$, $C(3)$, $C(5)$, $C(7)$, $\neg C(4)$, $\neg C(6)$, $\neg C(2)$, $C(217)$, $\neg C(200)$ do not allow deduction of $\forall n\ C(2n+1) \wedge \neg C(2n)$. In the meanwhile, inductive logic programming might learn the hypothesis $\forall n\ C(2n+1) \wedge \neg C(2n)$, which could in turn allow for many other deductions. This paper examines the convergence properties of an inductive-deductive system, i.e. the probability that the $n+1$-st example can be proved from the axioms learned from the previous n examples. The originality of the work is to propose a probabilistic analysis of logical decidability and completeness, contrasting with the worst-case analysis and undecidability results used in the literature. Indeed, a worst-case perspective does not account for the fact that many relevant statements can yet be proved in an undecidable setting. The rest of this section describes the proposed framework, discusses the relevant work and introduces the results reported in the paper.

Formalisation. This paper considers a first order logic language, where the initial set of axioms $\mathfrak{Z}$ is an essentially undecidable ([Kleijnen, 1992, p277], [Tarski, 1949])[1] set of axioms with finite description length[2] such as the Zermelo-Fraenkel set of axioms.

Let us consider a sequence of examples or statements e_i, independently and identically distributed from a probability distribution M. We further assume that M is consistent with $\mathfrak{Z}$, in the sense that $\mathfrak{Z} \cup \{e\ s.t. M(e) > 0\}$ is consistent. From each set of examples $\mathfrak{E}_n = \{e_1, ..., e_n\}$, the system extracts a recursive set of axioms noted A_n, which together with $\mathfrak{Z}$ allows for proving every example in $\mathfrak{E}_n$. Three types of induction are distinguished:

- in *deduction*, A_n includes all examples in $\mathfrak{E}_n$, except those examples e_i which could be proved from the theory learned from the previous examples (i.e. all e_i except those such that $A_{i-1}, \mathfrak{Z} \vdash e_i$). A_n thus is an independent axiom set.
- in *pruned-deduction*, A_n is a minimal subset of $\mathfrak{E}_n$, sufficient to prove every example in $\mathfrak{E}_n$ ($A_n, \mathfrak{Z} \vdash \mathfrak{E}_n$).
- in *induction-deduction*, A_n is a set of axioms, minimal wrt its description length, such that every example in $\mathfrak{E}_n$ can be proved from A_n and $\mathfrak{Z}$. Contrasting with deduction and pruned-deduction, A_n is no longer necessarily included in $\mathfrak{E}_n$.

The theoretical case where (deductive, pruned-deductive and inductive-deductive) learning is based on Turing oracles will first be considered in sections 3 and 4, respec-

[1] A set of axioms is essentially undecidable if any recursive extension of this set is undecidable.

[2] In all the paper, the description length refers to any classical mathematical notation of statements or proofs. Note that a set of axioms with finite description length can include an infinite axiom schema.

tively devoted to the cases where the target set of axioms is finite and infinite.
Section 5 extends the analysis, considering Turing-computable-approximations of Turing oracles, based on finite-length proofs.

Goals of the study. The behavior of an inductive-deductive system is examined with respect to three stochastic variables, modeling respectively the completeness, the accuracy and the compactness of the current axiom set (noted A_n in the following instead of $A_n, \mathfrak{Z}$ for simplicity of notation):

- the incompleteness of A_n refers to the probability L_n that A_n does not allow for deciding on further examples ($L_n = M(\{e \; s.t. \; A_n \not\vdash e \wedge A_n \not\vdash \neg e\})$); in order to distinguish this incompleteness from the standard logical one, it will be refered to as the *relative* incompleteness;

- the error or falsity of A_n is the probability that A_n decides wrongly on further examples[3];

- the compactness is measured as the description size $DL(A_n)$ of the current axiom set (this is not related to other definitions of compactness, but we keep this notion as no ambiguity arises).

It must be noted that the behavior of the relative incompleteness rate L_n is known in some specific cases:

- In case $\mathfrak{Z}$ is a complete set of axioms, no learning is required ; $A_n = \mathfrak{Z}$ leads to $L_n = 0$ as for any e, $\mathfrak{Z} \vdash e$ or $\mathfrak{Z} \vdash \neg e$.

- Otherwise, if M is modified at each time step n by a malign adversary with unrestricted computational power, then after Gödel's first theorem, $L_n = 1$ for all n, as the adversary can choose M concentrated on some undecided e.

Thus, our framework lies between the (too simple, unrealistic) complete case, and the (too difficult, pessimistic) straightforward application of essential undecidability.

The goal is to examine the practical limitations of learning in a powerful language (e.g. including the axioms of set theory or arithmetic). The limitations of such languages regarding completeness and decidability issues are well known; these limitations have significant impact on inductive learning as well. For example, in languages including arithmetic there are always infinitely many theories proving a finite consistent set of statements; discussions around this fact are referred to as Quine's underdetermination thesis [List, 1999, Norton, 2003, Shook, 2002]. However, it might be the case that the problems entailed by incompleteness and undecidability, though certain, are actually *not* frequent. And if there are an infinite number of solutions to a finite learning problem (Quine's underdetermination thesis), then it might be interesting to assess the average quality of these solutions. Therefore, our goal is to provide a statistical study of the relative incompleteness, compactness and falsity of learned theories, applying the statistical learning methodology and body of results to other learning criteria, namely

[3]Note that in the deduction or pruned-deduction cases, A_n cannot be inconsistent with e_{n+1} since $A_n \subseteq \mathfrak{E}_n$ and distribution M is assumed to be consistent with $\mathfrak{Z}$.

the probability of facing an undecided example or introducing inconsistencies in the theory.

Related work. As far as we know, the simultaneous use of deduction and induction has not been studied yet in a statistical perspective though the three domains involved (automatic deduction, inductive and statistical learning, mathematical logic) have some intersections[4]. Along an inductive-deductive setting, the works related to Quine's underdetermination thesis [List, 1999, Norton, 2003, Shook, 2002] focused on a worst case analysis; they do not integrate the statistical learning and generalization aspects. Other studies deal with some kinds of incomplete frameworks, e.g. involving recursion theory and referring to Turing machines with oracles or infinite-time Turing machines [Hamkins, 2002]. However, this work focuses on extending the set of decidable statements and the role of induction is not considered.

Overview of the paper. As an alternative to the worst-case analysis, the framework proposed in this paper is based on a logically consistent probability distribution M over the set of statements. In each step n, the system outputs a set of axioms A_n from the first n statements, and one examines whether this set allows for proving further statements. As noted earlier on, if these further statements are selected in a worst-case manner, A_n does not allow for deciding their truth value even with unbounded computational resources. However, a worst-case perspective often leads to overly pessimistic conclusions [Cheeseman et al., 1991]. The probabilistic setting proposed is inspired by the standard Probably Approximately Correct (PAC) framework [Valiant, 1984], and the study borrows the standard statistical learning tools (VC-dimension [Vapnik and Chervonenkis, 1974]) in order to bound the relative incompleteness expectation $L_n = M(\{e \; s.t. \; A_n \nvdash e \wedge A_n \nvdash \neg e\})$.

The paper is organized as follows. Section 2 introduces general definitions and lemmas used in the rest of the paper. Section 3 presents results about the induction of a target theory with bounded description length, comparing the *deductive*, *pruned-deductive* and *inductive-deductive* learning settings. It is shown that (corollaries 1-4): i) in all cases, non-asymptotic performance depends on the underlying distribution M and it might be arbitrarily bad (as in the worst-case setting); ii) *induction-deduction*, and more generally restrictions on the description length entails faster convergence rates than *deduction* ; iii) for any algorithm with a faster completeness convergence rate than *deduction*, there exists a distribution such that the error or falsity is not almost surely zero ($\exists M, e \; s.t. \; \forall n, P(A_n \vdash \neg e) > 0$ and $M(e) > 0$); iv) *pruned* learning can behave arbitrarily badly in the sense of an infinite asymptotic description length. Section 4 considers the case of a target theory with infinite description length, and presents negative results (corollaries 5-8): i) arbitrarily slow convergence rates can occur; ii) the length of the axiom set can increase fast. However, the completeness rate goes to 1 as the number of examples goes to infinity. While results presented in sections 3 and 4 are based on an oracle (axiomatic optimization or theorem proving with unbounded computational power), section 5 considers the case of Turing-computable approximations of such an oracle. Results similar to those of the oracle case are presented (with, unfortunately, a huge computational complexity). The paper ends with a discussion

[4]Some advances in mathematical logic have been exploited for automatic theorem proving, for instance Craig's interpolation theorem is used to design a "partition-based" logic [Amir and McIlraith, 2003].

of the presented results; a short introduction to the terminology and state of the art is given in Appendix A.

For space limitations, some proofs are omitted in the paper, and can be found in [Baskiotis et al., 2007].

2 FORMAL BACKGROUND

Let $\mathfrak{Z}$ denote a consistent essentially undecidable set of axioms (e.g. Zermelo-Fraenkel). We note $DL(A)$ the description length of an axiom set A, and by abuse we use also $DL(e) = DL(\{e\})$. Let T' denote the set of consistent theories including $\mathfrak{Z}$, and let $T \subset T'$ be the set of consistent theories defined from an axiom set with finite description length ($T = \{t \in T', \exists A \; s.t. \; A \vdash t, DL(A) < \infty\}$). T and T' can be viewed as boolean mappings from the set of well-formed statements ($t(e) = 1$ iff $t \vdash e$).

This section examines the VC-dimensions and shattering properties of both T and T' spaces.

Theorem 1 [Baskiotis et al., 2007]: *T has infinite VC-dimension.*

Although T and therefore T' both have infinite VC-dimensions, they differ by their shattering properties :

Theorem 2 [Baskiotis et al., 2007]: *T' shatters an infinite set.* The above theorem implies significant differences about learning in the search spaces T and T'. Specifically, in the case of T' there exists distributions leading to arbitrarily slow convergence rates, such as $C/\log(\log(\log(n)))$ where n is the number of examples. In contrast, we shall see that a reasonable convergence rate is obtained within T, although the convergence can be delayed due to adverse distributions.

3 THE FINITE DESCRIPTION LENGTH CASE

Let M denote a probability distribution on the well-formed statements, such that the mass of M is restricted to a consistent theory t in T ($\exists t \in T \; s.t. \; M(t) = 1$). A first negative result concerns the incompleteness convergence rate. We show that there exists a distribution M such that the incompleteness rate is bounded from below.

In the corollary below, as in corollary 5, we will use a link between supervised learning (i.e. learning statements with their truth values) and unsupervised learning (i.e. finding theories covering true statements). This link is based on the fact that stating e is exactly equivalent to stating $\neg e$. A distribution M on couples $(x, y) = (e, true)$ or $(x, y) = (e, false)$, where e is a statement, can be replaced by a distribution M such that $M(e)$ is the probability of $(e, true)$ *plus* the probability of $(e, false)$, as well as we can identify sets of axioms with classifiers (the associated classifier separates theorems and non-theorems). This allows the use of counter-examples from learning in the framework of this paper. A family of classifiers (for supervised learning) such that for any learning algorithm $\mathbb{E}L_n \geq c$ for some distribution on these classifiers, is identified with a family of sets such that for any algorithm, for some distribution $\mathbb{E}L_n \geq c$.

Corollary 1 : *for any $n > 0$, for any $\delta > 0$, for any method generating A_n, there exists a generator of examples (distribution M) such that $\mathbb{E}[L_n] \geq \frac{1}{2exp(1)} - \delta$.*

Remark : This result can be reformulated as: for any $n > 0$, for any $\delta > 0$, there exists M such that after n examples the learned theory is at distance at least $\frac{1}{2exp(1)} - \delta$ from the target one. Note that the above distribution M depends on n.

Proof of corollary 1 : Follows from theorem 1 and the lower bound cited in Appendix A.2.

■

Let us first consider the *fine* learning case, restricting the description length of the induced axiom set. By abuse of notation, in the following the description length of a theory derived from an axiom set A is set to $DL(A)$.

Theorem 3 [Baskiotis et al., 2007]: *Let T_s denote the set of theories in T which can be generated from a set of axioms with description length less than s, and let V_s denote the VC-dimension of T_s. V_s is finite as T_s is finite. For each theory t, let $V(t)$ be defined by $V(t) = inf\{V_s | t \in T_s\}$.*

Let V denote $V(t^*)$ where t^* is the target theory, assuming it is finite. Let t_n denote the theory extracted along induction-deduction *after n examples.*

Then the following results hold :

1. **Convergence rate :** *if $n > V$,*

$$P(L_n > \epsilon) \leq 2(2\exp(1)n/V)^V 2^{-n\epsilon/2}$$

2. **Asymptotic behavior :** *almost surely, there exists n_0 such that*

$$n \geq n_0 \Rightarrow L_n = 0$$

3. *T does not shatter any infinite set.*

4. *$V(t_n) \leq V$.*

Let us now consider the *deduction* and *pruned-deduction* learning settings.

Theorem 4 [Baskiotis et al., 2007]: *Consider the deduction or pruned-deduction settings. For any decreasing sequence a_n bounded by $1/2$ and converging to 0, there exists a probability distribution M such that i) $\forall n, L_n \geq a_n$; ii) there exists $t \in T$ (i.e. $DL(t) < \infty$) such that $M(t) = 1$.*

Corollary 2 : *In the* deduction *or* pruned-deduction *framework, for any decreasing sequence a_n upper bounded by $1/2$, there exists M such that for any n the relative incompleteness of A_n is bounded from below by a sum of n independent binary random variables X_i, where X_i takes value $1/0$ with probability $(a_i, 1 - a_i)$. In particular, under distribution M, the relative incompleteness of A_n is greater than $\sum_{i \leq n} a_i$.*

Corollary 3 : *Theorem 4 and corollary 2 can be extended to any learning method producing a minimal (wrt set inclusion) theory such that it covers examples $e_1, \ldots, e_n$.*

Proof : This is a direct corollary of the proof of theorem 4. ■

Corollary 4 : *Any method which does not incur the limitations stated by theorem 4 or corollary 2, can with non-zero probability select a theory A_n which is strictly larger (wrt set inclusion) than the minimal theory generated from $\{e_1, ..e_n\}$. In particular,*

for some distribution, there is a positive probability of generating a theory inconsistent with some statement of non-zero measure.

Proof : Reformulation of corollary 3. ■

4 THE INFINITE DESCRIPTION LENGTH CASE

Removing the finite length assumption has significant impact on the convergence results obtained in the previous section, notably in relation to the lower bounds on the convergence rate (see Appendix A.2). In the proof of the following corollary, we use the same correspondence as explained before corollary 1.

Corollary 5 : *for any decreasing sequence* (a_n) *running to* 0 *and upper-bounded by* $1/16$*, there exists a distribution of examples such that* $\mathbb{E}(L_n) \geq a_n$*.*

Proof : Direct consequence of the lower bound on the convergence rate in Appendix A.2. ■

Corollary 6 : *for the* deduction *or* pruned-deduction *settings,* $\mathbb{E}[DL(A_n)] \geq \sum_{i \leq n} a_i$*.*

Corollary 7 : *In all learning cases such that the empirical error rate is null,* L_n *goes to* 0 *(in the sense that for any* $\epsilon > 0$*,* $P(L_n > \epsilon) \to 0$*).*

Corollary 8 : *For all learning methods such that* A_n *is consistent and proves all statements* $e_1, \ldots, e_n$*, there exists a distribution* M *such that the compactness* $DL(A_n)$ *goes to infinity.*

Proof : Consider M a distribution which support is a consistent non-recursively axiomatizable set of statements. For any axiom set A let $M(A)$ be the measure for M of all statements proved from A ($M(A) = \sum_{e \ s.t. \ A \vdash e} M(e)$). Consider $K(\epsilon)$, the minimal description length over all axiom sets A such that $M(A)$ greater than $1 - \epsilon$. $K(\epsilon)$ is non-decreasing, and $lim_{\epsilon \to 0} K(\epsilon) = \infty$. It is sufficient to see that $L_n < \epsilon$ implies that $DL(A_n) > K(\epsilon)$. ■

These results show that although the error rate goes to 0 as the number of examples increases, the convergence rate can be arbitrarily low. Moreover, the description length of the induced theory cannot be bounded, as showed above.

5 TURING-COMPUTABLE ALGORITHMS : PROOFS OF BOUNDED LENGTH

By definition, deduction, pruned-deduction and induction-deduction all rely on Turing machines with oracles. As a first step toward a practical analysis, this section considers instead approximate learning, based on Turing machines without oracles and bounded length reasoning[5]. The approximation is considered from an algorithmic complexity perspective.

Section 5.1 is devoted to a complexity analysis of axiomatic optimization. This result is used in section 5.2 to provide a bound on the convergence of pruned-deductive

[5]Since recursion theory provides negative results in the case of proofs with arbitrary length (ie, the set of statements that can be proved, in many cases, is not recursive but only recursively enumerable), we restricted this study to proofs with bounded length.

and deductive-inductive learning toward the target theory, in the case where the latter is finite.

5.1 Algorithmic complexity

Let us consider as hypothesis space the sets of axioms with finite description length (possibly including axiom schemas; proofs using axiom schemas are allowed as well). Let us define the k-deduction as follows: statement e is k-proved from the set of axioms A, noted $A \vdash_k e$, if there exists a proof of e from A with description length less than k. The algorithmic complexity of k-deduction (i.e. the complexity required to decide the fact that a statement e is k-proved from A) is upper bounded by a function noted $Complexity(DL(A), k, DL(e))$. In the arithmetic setting considered, $Complexity(DL(A), k, DL(e))$ is dominated by a 2^k term (considering all 2^k strings of length k and determining whether they are proofs of $A \vdash e$). Along the same lines, the k-consistency of a set of axioms is defined as follows: A is k-consistent, noted $A \nvdash_k \bot$ if there is no proof with length smaller than k that A is inconsistent. Similarly, the algorithmic complexity of k-consistency is upper bounded by $Complexity(DL(A), k, DL(\bot))$. Therefore, the generality and consistency tests (respectively, $B \vdash_k A$ and $B \nvdash_k \bot$) can be performed with complexity $\sum_{e \in A} Complexity(DL(B), k, DL(e))$, over all statements or axiom schemas e in A. The following proposition is then straightforward.

Proposition : Complexity of axiomatic optimization

Assume that there are at most 2^n sets of axioms with description length less than n. Given a set A of statements, axiomatic optimization aims at a set of axioms B with minimal description length such that it entails all statements in A : Find $Arg\ min_B\{DL(B) | B \vdash_k A,\ B \nvdash_k \bot\}$. Its complexity is upper bounded by

$$O\left(2^{DL(A)} \times DL(A) \times Complexity(DL(A), k, DL(A))\right)$$

5.2 Axiomatic optimization

Given a set $\mathfrak{E}_n$ of statements, the point here is to find a minimal set A_n of axioms, using a finite number δ_n of computation steps to check A_n consistency and completeness wrt $\mathfrak{E}_n$. We show that if δ_n increases sufficiently fast, there exists an algorithm with essentially same convergence results as in the oracle-based analysis (section 3). Practically, let $\delta_1, \ldots, \delta_n$ denote a sequence of integers. Then:

- Let T_{n+1} be the set of statements that can be proved with proofs of length at most δ_n from A_n ;
- A_n is a[6] minimal description length set of axioms such that: i) A_n proves all examples in $\mathfrak{E}_n$ with proof of length at most δ_n; ii) A_n does not prove $\bot$ with proof of length at most δ_n.

[6] In case of equality, the first axiom set in lexicographic order is retained.

Note that A_n is not necessarily a minimal set of axioms in the usual sense, e.g. one of the axioms could be proved from the others (but its presence makes it feasible to prove k-completeness). We note A^* the shortest axiom set capable of proving any e in the target theory (i.e. such that $M(e) > 0$). Let L_n here denote $L_n = M(\{e; e \notin T_{n+1} \vee (\neg e) \in T_{n+1}\})$.

Theorem 5 [Baskiotis et al., 2007]: *Consider e a random variable on statements with probability law M and assume that the mean and the variance of the shortest proof of e from A^* are finite, and assuming further that $\delta_n = \Omega(n^3)$, then ,*

1. $P(DL(A_n) \geq DL(A^*) + \epsilon) \leq p_{n,\epsilon}$ *with* $p_{n,\epsilon}$ *is* $O(1/n^2)$*, as soon as* $n = \Omega(1/\sqrt{\epsilon})$.
2. A_n *reaches* A^* *almost surely, and thus is consistent for* n *sufficiently large.*
3. $L_n \leq \epsilon'$ *with probability at least* $1 - p_{n,\epsilon} - 2\,S\,2^{-n\epsilon'/2}$ *for any* $\epsilon, \epsilon' > 0$*, where* S *is the number of axioms sets with description length bounded by* $DL(A^*) + \epsilon$.

This result shows that the theory extracted by induction-deduction (using Turing machines with no oracle and bounded-deduction) is consistent, for sufficiently large number of examples; that its description length converges toward the optimal one, and finally, that its relative incompleteness goes to 0 as $O(1/\sqrt{n})$.

In summary, Theorem 5 shows that a Turing-machine algorithm with no oracle can implement an inductive-deductive system, with essentially the same performances and limitations regarding consistency and completeness as in the theoretical case. Indeed the complexity of this algorithm is exponential in n.

6 DISCUSSION

A probabilistic relational setting has been proposed in this paper to study inductive-deductive systems (IDS). Precisely, from a random generator providing statements and their truth values, the IDS extracts a set of axioms via one among three settings: the *deduction* one corresponds to a purely deductive algorithm; the *pruned-deduction* one extracts a minimal excerpt of the statements, sufficient to prove all seen statements; and the *inductive-deductive* setting selects the set of axioms with minimal description length such that it proves all seen statements. Two cases are distinguished: the "finitely describable" (FDR) and "non-finitely describable" (NFDR) realities respectively correspond to the case where the target set of axioms has a finite (resp. infinite) description length. FDR and NFDR cases are confronted to *deduction*, *pruned-deduction* and *induction-deduction* settings, considering two criteria: relative incompleteness (proportion of statements which cannot be proved from the current theory) and compactness (description length of the current theory). Though relative incompleteness always goes to 0 as the number of examples goes to infinity, its convergence rate can be arbitrarily low in all cases, except when reality is finitely describable and in a *inductive-deductive* setting, in other words, when the system actually performs induction. In this favorable case, the target concept is reached almost surely in finite time. Along the same lines, the description length of the extracted theory is unbounded (for adverse distributions) in all

cases, except again when reality is finitely describable and in a *inductive-deductive* setting. This result provides additional precisions related to Quine's under-determination thesis. Despite the multiplicity of theories consistent with a finite set of statements, if the IDS system extracts the theory consistent with the statements already seen, that is minimal wrt its description length (as opposed to, wrt its set inclusion), then a fast convergence in terms of both incompleteness and length can occur granted that the reality is "finitely describable". Interestingly, there exists some distributions in the latter case which entail errors and not only undecidabilities; i.e. there are cases such that the event $\exists n; A_n \vdash \neg e$ has strictly positive probability. This result can be interpreted in the light of Popper's notion of falsifiability, central to the history of science; as shown by the very general corollary 4, if one abstains from producing hypotheses which can be falsified by examples, the convergence rate of the IDS is not better than that of rote learning. The last part of this paper has shown that the above theoretical results, obtained for Turing machines with oracles, essentially hold for Turing machines *without* oracles – although the considered algorithms are indeed of limited use due to their huge computational complexity. In summary, the main ambition of this paper is to contribute to a less pessimistic view of inductive-deductive systems in relational logic, than allowed by a worst-case analysis and based on undecidability results.

A STATE OF THE ART AND DEFINITIONS

This appendix briefly summarizes the notations and learnability results used in the paper. Notation $\sup X$, where X is a real-valued random variable, denotes the (possibly infinite) supremum of the x such that $P(X > x) > 0$.

A.1 Logical notations

A theorem is a statement which can be proved, which depends on both the logical setting and the axiom set considered. The paper only considers classical logic. Each axiom set A includes $\mathfrak{Z}$ and has finite description length. After Gödel's theorem, there exists thus e such that neither e nor $\neg e$ can be proved from A. For the feasibility of the study, it is assumed that $\mathfrak{Z}$ is consistent, although in many cases of interest, this has not been proved (and cannot be, e.g. for Zermelo Fraenkel, after Gödel's theorem). Notation $A \vdash e$ (respectively $A \vdash_k e$) denotes the fact that e can be proved from A (resp. with proof of description length less than k). In the whole paper, the description length $DL(\cdot)$ (of sets of axioms or proofs) refers to a standard logic coding (with no compression).

A.2 Statistical learning theory

The interested reader is referred to [Devroye et al., 1997], [Vidyasagar, 1997] for an exhaustive presentation. Let Z denote the example space, and let F denote the hypothesis space, where each hypothesis is viewed as a subset of Z. A set X of examples is said to be shattered by F if for any subset X' of X there exists $f \in F$ such that $f \cap X = X'$. The **VC-dimension** of F is the cardinal of the largest finite set that is

shattered by F. If arbitrarily large such sets exists, then the VC-dimension is said infinite. Hypothesis h is consistent with a set of examples X iff $h \cap X = h^* \cap X$, where h^* denotes the target concept. A learning algorithm associates a consistent hypothesis h_n to each training set X_n made of n iid examples drawn according to some probability distribution M. Accordingly, the loss variable L_n stands for the error expectation of h_n ($M\{z, z \in Z, h_n(z) \neq h^*(z)\}$). Fundamental results in the statistical learning theory can be summarized as: if the VC-dimension is finite, then the error expectation goes to 0 reasonably fast as the number of examples goes to infinity.

Theorem, case of null empirical error *(see [Vapnik and Chervonenkis, 1974], [Devroye et al., 1997, Th. 12.7, p202]) :*

Define $\hat{L}(P) = \frac{1}{s}\sum_{i=1}^{s} \chi_{P(x_i)\neq y_i}$ and $L(P) = \mathbb{E}\chi_{P(X)\neq Y}$, with the (x_i, y_i) a sample of size s iid according to the law of the random variable (X, Y).

Consider $\mathfrak{F}$ a family of boolean functions on a domain X and let V be its VC-dimension. Then, for any $\epsilon > 0$ if $s > V$,

$$P(\sup_{P\in\mathfrak{F};\hat{L}(P)=0} |L(P) - \hat{L}(P)| \geq \epsilon) \leq 2\left(2exp(1)s/V\right)^V 2^{-s\epsilon/2}$$

where $(2\exp(1)s/V)^V$ can be replaced by the $2s$-shattering coefficient of $\mathfrak{F}$.

Lower bound : *([Devroye et al., 1997, p239], theorem 14.3). Assume that the VC-dimension of F is infinite. Then for any $n > 0$, for any $\delta > 0$, for any classification rule, there exists at least one distribution such that $\mathbb{E}L_n \geq \frac{1}{2\exp(1)} - \delta$ and F contains at least a function f such that $L(f) = 0$.*

Lower bound on the convergence rate : *Assume that F shatters an infinite set. Then for any sequence (a_n) decreasing to 0 and upper bounded by $\frac{1}{16}$, for any classification rule, there exists at least one distribution such that $\forall n\ \mathbb{E}L_n \geq a_n$ whereas F contains f such that $L(f) = 0$.*

Mainly, the difference with the previous result is that the distribution does not depend upon n.

Lemma : learning on countable domains *Consider learning on a countable domain with a distribution and an algorithm ensuring that the empirical error $\hat{L}$ is zero. Then, the generalization error almost surely converges toward 0.*

References

[Amir and McIlraith, 2003] Amir, E. and McIlraith, S. (2003). Partition-based logical reasoning for first-order and propositional theories. *Artificial intelligence*, 162(1,2):49:88.

[Baskiotis et al., 2007] Baskiotis, N., Sebag, M., and Teytaud, O. (2007). Unpublished draft.

[Cheeseman et al., 1991] Cheeseman, P., Kanefsky, B., and Taylor, W. (1991). Where the really hard problems are. In *proceedings of IJCAI91*, pages 331–337.

[Devroye et al., 1997] Devroye, L., Györfi, L., and Lugosi, G. (1997). *A probabilistic Theory of Pattern Recognition*. Springer.

[Hamkins, 2002] Hamkins, J. (2002). Infinite time turing machines. *Minds and Machines (special issue on hypercomputation)*, 12(4):521–539.

[Khardon and Roth, 1997] Khardon, R. and Roth, D. (1997). Learning to reason. *Journal of the ACM*, 44(5):697–725.

[Kleijnen, 1992] Kleijnen, J. (1992). Sensitivity analysis of simulation experiments: regression analysis and statistical design. *Mathematics and Computers in Simulation*, 34(3-4):297–315.

[List, 1999] List, C. (1999). Craig's theorem and the empirical underdetermination thesis reassessed. *Disputatio 7*, pages 28–39.

[Norton, 2003] Norton, J. (2003). Must evidence under-determine theory ? In *First Notre Dame-Bielefeld Interdisciplinary Conference on Science and Values, Zentrum fr Interdisziplinre Forschung, Universitt Bielefeld*.

[Shapiro, 1981] Shapiro, E. Y. (1981). Inductive inference of theories from facts. *Research Report 192*.

[Shook, 2002] Shook, J. (2002). Dewey and quine on the logic of what there is. In Tom Burke, D. M. H. and Talisse, R., editors, *Dewey's Logical Theory: New Studies and Interpretations*. Vanderbilt University Press.

[Tarski, 1949] Tarski, A. (1949). On essential undecidability. *Journal of Symbolic Logic*, 14:75–76.

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

[Vapnik and Chervonenkis, 1974] Vapnik, V. and Chervonenkis, A. (1974). *Theory of Pattern Recognition*. Nauka, Moskow. (in Russian).

[Vidyasagar, 1997] Vidyasagar, M. (1997). A theory of learning and generalization. In *Springer*.