Fifth Internat. Conf. on Principles of Knowledge Representation and Reasoning (KR'96), pp. 256-265

Computing approximate diagnoses by using approximate entailment

Annette ten Teije
SWI
University of Amsterdam
The Netherlands
annette@swi.psy.uva.nl

Frank van Harmelen
Dept. of Math and CS
Vrije Universiteit Amsterdam
The Netherlands
frankh@cs.vu.nl

Abstract

The most widely accepted models of diagnostic reasoning are all phrased in terms of the logical consequence relations. In work in recent years, Schaerf and Cadoli have proposed efficient approximations of the classical consequence relation. The central idea of this paper is to parameterise the notion of diagnosis over different approximations of the consequence relation. This yields a number of different approximations of classical notions of diagnosis. We derive results about the relation between approximate and classical notions of diagnosis. Our results are attractive for a number of reasons. We obtain more flexible notions of diagnosis, which can be adjusted to particular situations. Furthermore, we obtain efficient anytime algorithms for computing both approximate and classical diagnoses.

1 INTRODUCTION

The motivations of this paper come from two areas, namely from the theory of diagnosis and from recent work on approximate entailment. We will discuss each of these in turn.

Diagnosis. The current theories of diagnostic reasoning give a strict definition of the functionality of diagnosis in terms of the classical consequence relation. The current literature gives little or no indication of what to do when the results of diagnostic reasoning do not satisfy the goal of the computation (e.g. when no diagnosis is obtained, or too many, etc.). We propose a number of approximate notions of diagnoses that may be usefully exploited in such cases.

Even in cases where current methods compute good solutions, we may be interested in solutions which are not exact (too large, too few, etc) but cheaper to compute. An advantage of the method we use is that approximate diagnoses can be computed more efficiently than classical diagnoses. In fact, an accumulation of ever better approximate diagnoses is no more expensive to compute than the most precise diagnosis of this series. This can be exploited by iteratively computing ever better approximations. This iteration can stop at a point when either a sufficiently suitable diagnosis has been obtained or when no time for further computation is available, yielding attractive any-time interruptible algorithms.

Approximate entailment. The proposal for approximate deduction from Schaerf and Cadoli (1995) includes a parameter S (a set of predicate letters) which determines both the accuracy and the cost of the approximation. The correct choice of this set S is crucial for the usefulness of the approximation. In the worst case, when choosing S incorrectly, approximate deduction may end up as expensive as classical deduction. In their paper, Schaerf and Cadoli admit that they have very few general strategies for choosing S appropriately. In general of course, this choice must be heuristic, since a perfect choice for S would defeat known complexity boundaries. In this paper, we exploit properties of diagnostic reasoning to propose a number of informed strategies for choosing this crucial parameter.

The structure of this paper is as follows: section 2 summarises the results on approximate deduction from Schaerf and Cadoli. Section 3 repeats some basic definitions on diagnosis and introduces some notation. Section 4 constitutes the main body of this paper. We derive theorems that characterise the influence of approximate entailment relations on the diagnoses that can be obtained. Section 5 discusses ways in which the results from this paper can be put to practical use and section 6 concludes.

2 SUMMARISING APPROXIMATE ENTAILMENT

In this section we will summarise the work in (Schaerf and Cadoli 1995), which defines the approximate entailment relations that we will exploit for our work on diagnoses. Schaerf and Cadoli define two approximations of classical entailment, named $\vdash_1$ and $\vdash_3$ which are either unsound but complete ($\vdash_1$) or sound but incomplete ($\vdash_3$). By analogy, they sometimes write $\vdash_2$ for classical entailment. Both of these approximations are parameterised over a set of predicate letters S (written $\vdash_1^S$ and $\vdash_3^S$) which determines their accuracy. We repeat some of the basic definitions from (Schaerf and Cadoli 1995):

A 1-S-assignment is a truth assignment to all literals such that
• If $x \in S$ then x and $\neg x$ are given opposite truth values
• If $x \notin S$ then x and $\neg x$ are both given the value 0.

A 3-S-assignment is a truth assignment to all literals such that
• If $x \in S$ then x and $\neg x$ are given opposite truth values
• If $x \notin S$ then x and $\neg x$ are *not* both given the value 0.

In other words: for letters in S, these assignments behave as classical truth assignments, while for letters $x \notin S$ they make either all literals false (1-S-assignments) or make one or both of x and $\neg x$ true (3-S-assignments).

The names of the assignments can be explained as follows. In 1-S-assignments means: there is *one* possible model for letters outside S, namely false for both x and $\neg x$. In 2-S-assignments: there are *two* possible models for letters namely x true and $\neg x$ false, or x false and $\neg x$ true. In 3-S-assignments: there are *three* possible models for letters outside S namely x true and $\neg x$ false, or x false and $\neg x$ true or true for both x and $\neg x$.

Satisfaction of a clause by a 1-S- or 3-S-assignment is defined in the usual way: a formula ϕ is satisfied by an interpretation σ if σ evaluates ϕ into true using the standard rules for the connectives.

The notions of 1-S-entailment and 3-S-entailment are now defined in the same way as classical entailment: a theory T 1-S-entails a formula ϕ written ($T \vdash_1^S \phi$) *iff* every 1-S-assignment that satisfies T also satisfies ϕ, and similarly for 3-S-entailment ($T \vdash_3^S \phi$). Similarly, we can speak of 1-S- and 3-S-consistency.

The following syntactic operations[1] can be used to clarify these definitions. For a theory in clausal form, 1-S-entailment corresponds to classical entailment, but after removing from every clause any literals with a letter outside S. When this results in an empty clause, the theory becomes the inconsistent theory $\perp$. Similarly, 3-S-entailment corresponds to classical entailment, but after removing every clause from the theory that contains a literal with a letter outside S. This may result in the empty theory $\top$. These intuitions lead to the main result of (Schaerf and Cadoli 1995):

Theorem 1 (Approximate entailment)

$$\vdash_3^{\emptyset} \Rightarrow \vdash_3^{S} \Rightarrow \vdash_3^{S'} \Rightarrow \vdash_2 \Rightarrow \vdash_1^{S'} \Rightarrow \vdash_1^{S} \Rightarrow \vdash_1^{\emptyset}$$

where $S \subseteq S'$. Everywhere in this paper we will use primed letters to represent sets that are a superset of the unprimed letter.

This states that $\vdash_3^S$ is a sound but incomplete approximation of the classical $\vdash_2$. The counterpositive of the second halve of the theorem (reading $\not\vdash_1^S \Rightarrow \not\vdash_1^{S'} \Rightarrow \not\vdash_2$) states that $\not\vdash_1^S$ is a sound but incomplete approximation of $\not\vdash_2$.

Example 1 [Illustrating $\vdash_3^S$ and $\not\vdash_1^S$] We illustrate these notions with the example theory given in figure 1. We shall call this theory BM, for reasons that will become apparent later. We will use this simple theory throughout the paper to illustrate our results.

We can see that $\vdash_3^S$ is incomplete with respect to $\vdash_2$, since in the theory BM of figure 1 we have that classically $BM \cup \{H_3\} \vdash_2 O_1$, but if we restrict S to $\mathrm{LET}(BM) \setminus \{H_3\}$, where $\mathrm{LET}(BM)$ stands for all the predicate letters in BM, we do not have that $BM \cup \{H_3\} \vdash_3^S O_1$. Similarly, $\not\vdash_1^S$ is incomplete with respect to $\not\vdash_2$ since, for example, if $S = \mathrm{LET}(BM) \setminus \{H_0\}$, then $BM \cup \{H_1\} \not\vdash_2 O_1$, but not $BM \cup \{H_1\} \not\vdash_1^S O_1$.

Furthermore, with increasing S, the accuracy of these approximations improves, until the approximate versions coincide with classical entailment when all letters are included in S. Conversely, the approximations trivialise when $S = \emptyset$: any formula is $\vdash_3^{\emptyset}$ satisfiable, and no formula is $\vdash_1^{\emptyset}$ satisfiable.

Without proof, we give a number of simple lemma's on basic properties of $\vdash_i^S$ that we will use in this paper. In the following lemma's, γ and δ are clauses, ϕ is any formula, l is a literal, and x is a proposition letter.

The first lemma shows that a familiar property of $\vdash$ also holds for all approximations:

Lemma 1 (Monotonicity of $\vdash_i^S$)

$$\textit{If} \quad T \vdash_i^S \phi \quad \textit{then} \quad T \wedge x \vdash_i^S \phi$$

[1] It is because of this close correspondence between 1,3-S-entailment and these syntactic operations that we write $\vdash_i^S$ instead of $\models_i^S$, which was the notation used in (Schaerf and Cadoli 1995).

$$
\begin{array}{lcl}
H_1 \wedge H_0 & \rightarrow & O_1 \\
H_1 \wedge H_0 & \rightarrow & O_2 \\
H_2 & \rightarrow & O_2 \\
H_3 & \rightarrow & O_1 \\
H_0 & \rightarrow & O_3 \\
H_4 & \rightarrow & O_3 \\
H_4 & \rightarrow & \neg O_2
\end{array}
\qquad
\begin{array}{l}
\neg H_1 \vee \neg H_0 \vee O_1 \\
\neg H_1 \vee \neg H_0 \vee O_2 \\
\neg H_2 \vee O_2 \\
\neg H_3 \vee O_1 \\
\neg H_0 \vee O_3 \\
\neg H_4 \vee O_3 \\
\neg H_4 \vee \neg O_2
\end{array}
$$

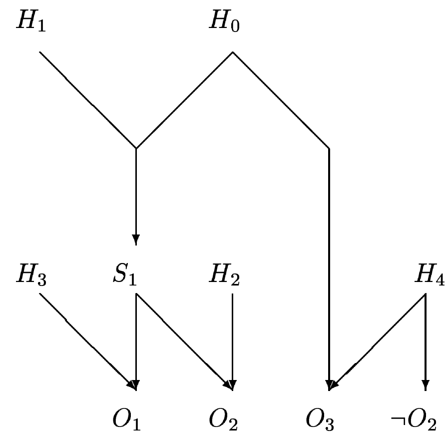

Figure 1: A simple example theory BM (non-clausal, clausal, and as causal network)

The second lemma again shows that a simply property of $\vdash$ continues to hold for $\vdash_i^S$:

Lemma 2 (Extendibility of $\vdash_i^S$)

$$
\begin{array}{llll}
\textit{If} & T \not\vdash_1^S \bot \textit{ and } x \in S & & \\
 & \textit{then} & T \wedge x \not\vdash_1^S \bot & \textit{or } T \wedge \neg x \not\vdash_1^S \bot \\
\textit{If} & T \not\vdash_3^S \bot & \textit{then} \quad T \wedge x \not\vdash_3^S \bot & \textit{or } T \wedge \neg x \not\vdash_3^S \bot
\end{array}
$$

While the previous lemma said that of a consistent theory at least one extension with x or $\neg x$ remained consistent under the same S, the following lemma says that when moving from S to $S + x$, both extensions remain consistent:

Lemma 3 (Extendability of $\vdash_1^S$ with increasing S)

$$\textit{If } T \not\vdash_1^S \bot \textit{ then } T \wedge x \not\vdash_1^{S+x} \bot \textit{ and } T \wedge \neg x \not\vdash_1^{S+x} \bot$$

The following lemmas are stated informally in (Schaerf and Cadoli 1995). They make precise the syntactic intuitions behind $\vdash_1^S$ and $\vdash_3^S$ discussed above.

Lemma 4 ($\vdash_1^S$ amounts to removing literals)
If $\mathrm{LET}(l) \notin S$ *then*

$$T \wedge (l \vee \delta) \vdash_1^S \gamma \textit{ iff } T \wedge \delta \vdash_1^S \gamma$$

Lemma 5 ($\vdash_3^S$ amounts to removing clauses)
If $\mathrm{LET}(l) \notin S$ *and* $\mathrm{LET}(\gamma) \subseteq S$ *then*

$$T \wedge (l \vee \delta) \vdash_3^S \gamma \textit{ iff } T \vdash_3^S \gamma$$

Lemma 5.1 in (Schaerf and Cadoli 1995) shows that the condition $\mathrm{LET}(\gamma) \subseteq S$ is no restriction, since $T \vdash_3^S \gamma$ iff $T \vdash_3^{S \cup \mathrm{LET}(\gamma)} \gamma$.

Schaerf and Cadoli also present incremental algorithms for computing $\vdash_1^S$ and $\vdash_3^S$ when S increases. They have obtained attractive complexity results which state that even when computing $\vdash_2$ through iterative computation of $\vdash_3^S$, the total cost of the iterated computation is not larger than the direct computation of $\vdash_2$ (and similarly for $\not\vdash_1^S$ to compute $\not\vdash_2$). However, the iterative computation of the approximate entailment has an important advantage that the iteration may be stopped when a confirming answer is already obtained for a smaller value of S. This yields a potentially drastic reduction of the computational costs. The size of these savings depend on the appropriate choice for S.

In this paper, we use Schaerf and Cadoli's results on approximating the propositional entailment relation. However, in (Schaerf and Cadoli 1995) they show how these results can be extended to first-order theories in a straightforward way: instead of a set of propositional letters, S becomes a set of ground atoms from the Herbrand base.

3 BASIC DEFINITIONS

In this section we introduce the definition of a diagnostic problem and a diagnostic solution that we will use in this paper, and we introduce some notation. We use a common definition of diagnosis that is widespread in the literature. We adopt the mainstream approach of using entailment to characterise the relation between the observations, behavioural model and computed diagnoses. In this paper we present the results on the standard abductive notion of diagnosis. However, we have results on consistency based notion of diagnosis, and results on the combination of abductive based and consistency based notion of diagnosis proposed in (Console and Torasso 1991).

Definition 1 (Diagnostic problem and its solution)
Given a behaviour model BM as a logical theory in

clausal form, and observations O as a set of literals (read as a conjunction), a solution to a diagnostic problem $\langle BM, O\rangle$, is a set of literals E ("E" for explanation, again read as a conjunction), which satisfies the following:

$$BM \cup E \quad \vdash \quad O \tag{1}$$
$$BM \cup E \quad \not\vdash \quad \perp \tag{2}$$

These two formulae state that the explanation E must explain the observations O in a non-trivial way.

For technical reasons, we restrict ourselves to causal theories consisting of only two layers. The theory from figure 1 can be interpreted as the causal network from the same figure, where any intermediate layers have been removed in the translation from causal network to theory (in this case the removal of S_1). This restriction does not affect the expressive power of causal theories: any causal theory T containing intermediate layers can be trivially transformed into a theory T' where the intermediate layers have been removed (by resolution on their positive and negative occurrences in the clausal form). T is equivalent to T' in the sense that for any sentence ϕ consisting of only causes and effects, we have

$$T \vdash \phi \text{ iff } T' \vdash \phi$$

A further restriction is that causes in the network may only be positive literals. Observations in the network are allowed to be either positive or negative. Although made for technical reasons, this restriction is observed by many of the causal models found in the diagnostic literature. The work in (Bylander *et al.* 1991) shows that even for such reduced causal theories, the computational complexity of abductive reasoning is sufficiently bad to require the kind of computationally attractive approximations that we investigate in this paper.

Example 2 [Illustrating diagnosis] In figure 1, if we take $O = \{O_1, O_2\}$, then $\{H_0, H_1\}$ and $\{H_0, H_1, \neg H_4\}$ are both diagnoses according to definition 1. The difference between these alternatives is that the the second says that H_4 is false, whereas the first is uncommitted to the states of H_4. A set like $\{H_0, H_1, H_4\}$ is *not* a solution, since it violates (2) via H_4.

Note that we have not mentioned a notion of minimal diagnosis. The literature contains a large number of proposals for different notions of minimality, each of which are appropriate in different contexts. Our theorems are largely independent of particular notions of minimality. We have therefore not included it in our definition, thereby making our results more widely applicable.

The central idea of this paper is to parameterise the notion of diagnosis over different approximations of the entailment relation. In particular, we will want to use $\vdash_1^S$ and $\vdash_3^S$ in our definition of diagnosis. We will write ABD_3^S for formulae (1) and (2) but using $\vdash_3^S$ instead of $\vdash$, and ABD_1^S for using $\vdash_1^S$. We will write ABD_i^S when we intend both ABD_1^S and ABD_3^S. Remember that we write $\vdash_2$ for the classical deduction relation $\vdash$, and we will therefore also write ABD_2 for (1)–(2).

We want to emphasise that our particular definition of a diagnostic problem and its solution, while very common in the literature, is not of crucial importance to our work. It would be well possible to adopt another definition, as long as this definition is again based on logical entailment. The details of our theorems would change, but our central message (namely that approximate entailment can be usefully exploited for diagnostic reasoning to obtain interesting and efficient results) would still hold.

4 APPROXIMATING DIAGNOSES

This section constitutes the main body of work in this paper. We derive theorems that characterise the influence of approximate entailment relations on the diagnoses that can be obtained. The choice for a particular approximation

is characterised by selecting either $\vdash_1^S$, $\vdash_2$ or $\vdash_3^S$, and additionally choosing a value for the set of predicate letters S.

The main intuitions behind using $\vdash_1^S$ and $\vdash_3^S$ in diagnosis are as follows. By using $\vdash_1^S$, candidate solutions more easily satisfy part (1) of our definition of diagnosis, because $\vdash_2 \Rightarrow \vdash_1^S$. Similarly, by using $\vdash_3^S$, candidate solutions more easily satisfy part (2) of our definition of diagnosis, since $\not\vdash_2 \Rightarrow \not\vdash_3^S$.

Example 3 [Illustrating ABD_1^S and ABD_3^S diagnoses] Again in figure 1, if we take $O = \{O_1, O_2\}$, then $\{H_2, H_3\}$ is an ABD_2 diagnosis, i.e. satisfying (1)–(2). $\{H_2\}$ is not an ABD_2 diagnosis, but if we take $S = \text{LET}(BM) \setminus \{H_3\}$, then $\{H_2\}$ is an ABD_1^S diagnosis. In this case, H_3 is no longer required for explaining O_1. Because $H_3 \notin S$, H_3 and $\neg H_3$ are both false in all 1-S-models, and since $(\neg H_3 \vee O_1) \in BM$, O_1 must be true in all 1-S-models satisfying BM, i.e. $BM \vdash_1^S O_1$. In other words, O_1 requires no further explanation. We will sometimes say that in such a case O_1 is explained "for free".

Similarly, $\{H_0, H_1, H_4\}$ is not a classical ABD_2 diagnosis, since it violates (2). However, if we take $S = \text{LET}(BM) \setminus \{H_4\}$, then $\{H_0, H_1, H_4\}$ is an ABD_3^S diagnosis, showing that it is easier to satisfy (2) through using ABD_3^S.

The organisation of this section is as follows: we start with some observations that restrict the choice of S; we then investigate the effects of using $\vdash_1^S, \vdash_3^S$ in (1–2) (in other words: using ABD_i^S)

4.1 RESTRICTIONS ON CHOOSING S

The first restriction on the choice of S follows from the following theorem. It states that we must demand $O \subseteq S$ if we are to find any diagnoses at all. (Unless stated otherwise, variables are universally quantified):

Theorem 2 (S must contain O)

$$[\exists E : \mathrm{ABD}_1^S(E)] \rightarrow O \subseteq S$$

Proof Suppose that there would be an $o \in O$ with $o \notin S$, then o would have valuation false in all 1-S-models (by definition of $\vdash_1^S$), and then formula (1) can only hold if $BM \cup E$ also has valuation false in all 1-S-models, in other words, $BM \cup E$ would be inconsistent, but this would contradict requirement (2). □

The next theorem states that for ABD_1-diagnoses to exist, at least one letter from each clause in BM must occur in S.

Theorem 3 (S must meet all clauses in BM)

$$[\exists E : \mathrm{ABD}_1^S(E)] \rightarrow \forall \gamma \in BM : \mathrm{LET}(\gamma) \cap S \neq \emptyset$$

Proof If the conclusion of the theorem were not the case, BM would collapse into a 1-S-inconsistent theory, and would never satisfies formula (2). □

Example 4 [Illustrating theorem 3] If we would take $S = \mathrm{LET}(BM) \setminus \{H_3, O_1\}$, we would indeed have a clause with all letters outside S. This would make BM inconsistent, since $(\neg H_3 \vee O_1) \in BM$, but since both H_3 and O_1 are outside S, $\neg H_3$ and O_1 must evaluate to false in all 1-S-models (by definition of $\vdash_1^S$), thereby making BM inconsistent.

Whereas the above two observations restricted S in relation to either O or BM, the following restrict S in relation to the explanation E:

Theorem 4 (S must contain E)

$$\begin{array}{rcl} \mathrm{ABD}_1^S(E) & \rightarrow & \mathrm{LET}(E) \subseteq S \\ \mathrm{ABD}_3^S(E) \wedge \subseteq\text{-}min(E) & \rightarrow & \mathrm{LET}(E) \subseteq S \end{array}$$

where $\subseteq$ -$min(E)$ means that E is a subset-minimal explanation.

Proof For ABD_1, suppose that $x \in E$ and $x \notin S$, then x evaluates to false in all 1-S-models, so E must evaluate to false in all 1-S-models, making $BM \cup E$ inconsistent, thereby violating part (2) of our definition of diagnosis.

For ABD_3^S: suppose there would be an $x \in E$ and $x \notin S$. We will prove that $E \setminus \{x\}$ would then also be an ABD_3^S diagnosis, and this would contradict the assumption that E is subset-minimal. To prove that $\mathrm{ABD}_3^S(E \setminus \{x\})$, we must prove $BM \cup E \setminus \{x\} \not\vdash_3^S \perp$ and $BM \cup E \setminus \{x\} \vdash_3^S O$. The first follows from $BM \cup E \not\vdash_3^S \perp$ (which holds since $\mathrm{ABD}_3^S(E)$) plus using monotonicity (lemma 1), the second follows from $BM \cup E \vdash_3^S O$ (again by $\mathrm{ABD}_3^S(E)$) plus using lemma 5 which applies since $x \notin S$. □

Example 5 [Illustrating theorem 5] For ABD_3^S, take as an example $S = \{H_0, H_1, O_1, O_2\}$, and $O = \{O_1, O_2\}$, then $\{H_0, H_1\}$ is a subset-minimal ABD_3^S diagnosis. Other classical subset-minimal diagnoses (such as $\{H_2, H_3\}$) are not ABD_3^S diagnoses, since their letters are not in S. Non-subset-minimal ABD_3^S diagnoses are allowed to have their letters outside S, for example $\{H_0, H_1, H_3\}$.

On the basis of these initial restrictions on possible choices of S, we are now ready to derive results on the behaviour of ABD_i^S.

4.2 APPROXIMATING ABDUCTIVE DIAGNOSES

In general, a classical abductive diagnosis E can always be arbitrarily expanded to $E \wedge x$ or $E \wedge \neg x$ (this follows directly from lemma's 1 and 2). The following theorem shows that this property (which accounts for the exponential number of abductive diagnoses) continues to hold for ABD_i^S-diagnoses. On the other hand, a classical diagnosis cannot be arbitrarily reduced, since it might get too small to imply O and thereby fail to satisfy formula (1). Surprisingly, ABD_i^S diagnoses *can* always be reduced in a certain way, namely by removing letters not in S. Formally:

Theorem 5 (Changing ABD_i^S diagnoses with fixed S)

$$\begin{array}{rcl} x \in S \wedge \mathrm{ABD}_i^S(E) & \rightarrow & \mathrm{ABD}_i^S(E \wedge x) \quad \vee \\ & & \mathrm{ABD}_i^S(E \wedge \neg x) \\ x \notin S \wedge \mathrm{ABD}_i^S(E) & \rightarrow & \mathrm{ABD}_i^S(E \setminus \{x, \neg x\}) \end{array}$$

Proof First part of the theorem: The condition $x \in S$ is only required for ABD_1^S. For ABD_2 and ABD_3^S, the first part of the theorem already holds without this condition. This is because requirement (1) continues to hold when expanding E (through monotonicity of $\vdash_2$ and $\vdash_3^S$), and requirement (2) continues to hold because at least one of $BM \cup (E \wedge x)$ and $BM \cup (E \wedge \neg x)$ must be consistent if $BM \cup E$ was consistent (extendibility). For $\vdash_1^S$ this extendibility property only holds if $x \in S$.

Second part of the theorem: The second part of the theorem is trivial for ABD_1^S, since, according to theorem 4, $x \notin S$ implies $x \notin E$ in which case $E \setminus \{x, \neg x\}$ simply equals E. The proof for the case ABD_3^S follows from monotonicity (to show consistency of $BM \cup E \setminus \{x, \neg x\}$) and from lemma 5 (to show that $BM \cup E \setminus \{x, \neg x\}$ still implies O). That lemma applies since $x \notin S$. □

Example 6 [Illustrating theorem 5] As an example, we can use the one from example 5, where $\{H_0, H_1, H_3\}$ is an ABD_3^S diagnosis if we take $S = \{H_0, H_1, O_1, O_2\}$, and consequently $\{H_0, H_1\}$ is also an ABD_3^S since $H_3 \notin S$.

Theorem 5 holds for abductive diagnoses under a constant value of S. A related result can be obtained when expanding S with some letter x: when increasing S to $S + x$, for any ABD_1^S-diagnosis E, we can always find a larger ABD_1^{S+x}-diagnosis E' which is a superset of E. Similarly for every ABD_3^S-diagnosis E' there is a smaller ABD_3^{S+x}-diagnosis E with $E \subseteq E'$. As above, such expansion and contraction properties do not in general hold for classical diagnoses.

Theorem 6 (Changing ABD_i^S diagnoses with changing S)

$$\begin{array}{rcl} \mathrm{ABD}_1^S(E) & \rightarrow & \mathrm{ABD}_1^{S+x}(E \wedge x) \quad \vee \\ & & \mathrm{ABD}_1^{S+x}(E \wedge \neg x) \\ x \notin S \wedge \mathrm{ABD}_3^S(E) & \rightarrow & \mathrm{ABD}_3^{S+x}(E \setminus \{x, \neg x\}) \end{array}$$

Proof First part of the theorem: We must prove that the new diagnosis is 1-$S + x$-consistent and 1-$S + x$-implies the observations. Consistency is guaranteed because $BM \cup E$ is 1-S-consistent (since $\mathrm{ABD}_1^S(E)$), and because we extend the existing diagnosis with either x or $\neg x$, lemma 3 says that both extensions will be 1-$S + x$-consistent. Implication of the observations is guaranteed in the following way: First observe that if $x \in S$, implication is guaranteed trivially by lemma 1, so in the following we can assume $x \notin S$. By extending S with x we might loose consequences (observations) that were explained "for free" under $\vdash_1^S$, but these can be recovered by adding x or $\neg x$ to the diagnosis. Since we only have positive literals as causes, it will only ever be needed to add x and never $\neg x$, so it will never happen that we need to add x for one observation and $\neg x$ for another. More formally: Since $\mathrm{ABD}_1^S(E)$, we know that $BM \cup E \vdash_1^S O$, but by extending S with x, $BM \cup E \vdash_1^{S+x} O$ is no longer guaranteed (theorem 1). If $BM \cup E \vdash_1^{S+x} O$ does hold, then so do $BM \cup (E \wedge x) \vdash_1^{S+x} O$ and $BM \cup (E \wedge \neg x) \vdash_1^{S+x} O$ (by monotonicity), and we are done. So lets assume that $BM \cup E \not\vdash_1^{S+x} O$. We know that $BM \cup E \vdash_1^S O$, so the only reason for $BM \cup E \not\vdash_1^{S+x} O$ must be the extra 1-$S + x$-models introduced by allowing x or $\neg x$ to be true (instead of forcing them to be both false because $x \notin S$). Since we only have positive causes in our network, only x can appear as a premise of a causal link and since no node is both a cause ad and effect (since we are in a two-layer network), only $\neg x$ can appear in the clauses of BM. We can therefore remove the additional 1-$S + x$-models for $BM \cup E$ by forcing $\neg x$ to be false. Since $x \in S + x$, this amounts to forcing x to be true, which amounts to adding x to $BM \cup E$. In other words: $BM \cup E \wedge x \vdash_1^{S+x} O$. This proves the first part of the theorem.

Second part of the theorem: We must show that the new diagnosis is 3-$S + x$-consistent and 3-$S + x$-implies the observations. The proof of the latter is easy: The removal of x and $\neg x$ does not lead to loss of implication of any observations, since these literals were not used in the original diagnosis either (because $x \notin S$). More formally: $\mathrm{ABD}_3^S(E)$ means $BM \cup E \vdash_3^S O$, and since $x \notin S$ this is equivalent to $BM \cup E \setminus \{x, \neg x\} \vdash_3^S O$ (by lemma 5), and this in turn implies $BM \cup E \setminus \{x, \neg\} \vdash_3^{S+x} O$ (by theorem 1), as desired. Showing that the new diagnosis is still consistent is somewhat more subtle: by extending S with x we gain additional causal links, namely those links containing either x or $\neg x$ as premise or conclusion. When enabled these additional causal links might cause inconsistency in the following ways: (1) x or $\neg x$ are premises, and the conclusion of the link is inconsistent with others; (2) there are both links with x and links with $\neg x$ as their consequences. However, in case (1) the causal link would never be enabled, since x and $\neg x$ are removed as axioms, and they can also not be implied by other causal links because they are already premises of a causal link, and we are restricted to two-layer causal networks. Case (2) can never happen because such links would already have caused 3-S-inconsistency, which contradicts the condition of the theorem. □

Example 7 [Illustrating theorem 6] To understand this theorem, it is important to realise that ABD_i^{S+x} can be thought of as "more classical than ABD_i^S", in other words, ABD_i^{S+x} is a little bit more like ABD_2 than ABD_i^S was. We should also recall the intuition from the beginning of this section, which stated that ABD_1^S makes it easier to satisfy diagnostic requirement (1) and ABD_3^S makes it easier to satisfy (2).

As an example, we take $S = \mathrm{LET}(BM) \setminus \{H_2\}$, and $O = \{O_1, O_2\}$. Then $\{H_3\}$ is an ABD_1^S diagnosis. When increasing S with H_2, $\{H_3\}$ is by itself no longer an explanation for $\{O_1, O_2\}$, and must be expanded to $\{H_2, H_3\}$, in other words: $\{H_2, H_3\}$ is an $\mathrm{ABD}_1^{S+H_2}$ diagnosis (as predicted by the theorem). Notice that $\{H_3\}$ itself is not an $\mathrm{ABD}_1^{S+H_2}$ diagnosis, since it violates requirement (1): excluding H_2 from S gave us the explanation of O_2 "for free". Conversely, $\{H_2, H_3\}$ is not an ABD_1^S diagnosis.

Similarly, for ABD_3^S if we take $S = \mathrm{LET}(BM) \setminus \{H_4\}$ (again with $O = \{O_1, O_2\}$). Then $\{H_2, H_3, H_4\}$ is an ABD_3^S diagnosis. When adding H_4 to S, $\{H_2, H_3, H_4\}$ is no longer an $\mathrm{ABD}_3^{S+H_4}$ diagnosis, since it violates (2). As predicted by the theorem, $\{H_2, H_3\}$ is an $\mathrm{ABD}_3^{S+H_4}$ diagnosis.

Because the superset and subset relations between diagnoses under different values of S, as in theorem 6, will turn out to be very important, we introduce the following notation:

Definition 2 *For any set P and P', $P \subseteq^{*} P'$ and $P \subseteq^{+} P'$ are defined by:*

$$
\begin{array}{lll}
P \subseteq^{*} P' & \equiv & \forall p \in P \quad \exists p' \in P' : \quad p \subseteq p' \\
P \subseteq^{+} P' & \equiv & \forall p' \in P' \quad \exists p \in P : \quad p \subseteq p'
\end{array}
$$

Example 8 [Examples of $\subseteq^{*}$ and $\subseteq^{+}$]

$$
\begin{array}{ccccccc}
 & P \subseteq^{*} P' & & \qquad & & P \subseteq^{+} P' & \\
\{\{a\}, & \rightarrow & \{\{a,b\}, & & \{\{a\}, & \leftarrow & \{\{a,b\}, \\
 & \searrow & & & & \nwarrow & \\
 & & \{a,c\}, & & & & \{a,c\}, \\
\{b,c\}\} & \rightarrow & \{b,c,d\}, & & \{b,c\}, & \leftarrow & \{b,c,d\}\} \\
 & & \{b,e\}\} & & \{e\}\} & & \\
 & \mathit{injective} & & & & \mathit{surjective} &
\end{array}
$$

Notice that $P \subseteq P' \rightarrow P \subseteq^{*} P'$ and $P \subseteq P' \rightarrow P' \subseteq^{+} P$. We can use this notation to summarise our results so far. If we write $\mathcal{ABD}_i^S$ for the set of all diagnoses E which satisfy $\mathrm{ABD}_i^S(E)$, we have:

Theorem 7 (Relations between $\mathcal{ABD}_i^S$)

$$
\begin{array}{ll}
\emptyset = & \mathcal{ABD}_1^{\emptyset} \subseteq \mathcal{ABD}_1^{S} \subseteq^{*} \mathcal{ABD}_1^{S'} \subseteq^{*} \mathcal{ABD}_2 \\
 & \mathcal{ABD}_2 \subseteq^{+} \mathcal{ABD}_3^{S'} \subseteq^{+} \mathcal{ABD}_3^{S} \subseteq^{+} \mathcal{ABD}_3^{\emptyset} = \emptyset
\end{array}
$$

This states that ABD_1^S diagnoses consist of parts of ABD_2 diagnoses, and that ABD_3^S diagnoses contain ABD_2 diagnoses. **Proof** The theorem is a straightforward reformulation of theorem 6 using the notation of definition 2. □

This theorem only claims that for increasing values of S, we will find superset diagnoses for $\mathrm{ABD}_1^{S'}$ and subset diagnoses for $\mathrm{ABD}_3^{S'}$. The next theorem considerably strengthens this result, by stating that, in the case of ABD_1^S, not only do we find superset diagnoses, but that whenever we find a subset-diagnosis (when increasing S to S'), this $\mathrm{ABD}_1^{S'}$ subset-diagnosis was already an ABD_1^S-diagnosis. Stated differently: when moving from ABD_1^S to $\mathrm{ABD}_1^{S'}$ we will find superset diagnoses, but we will not find any *new* subset diagnoses. The converse also holds: when, while *reducing* S' to S, we find an ABD_1^S diagnoses E' which is a superset of an $\mathrm{ABD}_1^{S'}$ diagnosis E, then this E' was already a diagnosis under $\mathrm{ABD}_1^{S'}$. Roughly speaking, theorem 7 promises that ABD_1^S diagnoses will become larger with increasing S, and the following theorem promises that we will not find any *new* smaller diagnoses when increasing S. A similar result is obtained for ABD_3^S. Formally:

Theorem 8 (No new subset (superset) diagnoses)

$$
\begin{array}{lll}
\mathrm{ABD}_1^S(E') \wedge \mathrm{ABD}_1^{S'}(E) \rightarrow & \mathrm{ABD}_1^S(E) & \wedge \\
 & \mathrm{ABD}_1^{S'}(E') & \\
\mathrm{ABD}_3^S(E) \wedge \mathrm{ABD}_3^{S'}(E') \rightarrow & \mathrm{ABD}_3^S(E') & \wedge \\
 & \mathrm{ABD}_3^{S'}(E) &
\end{array}
$$

Proof When we write out the definitions for ABD_1^S and $\mathrm{ABD}_1^{S'}$ in the first part of the theorem, we get:

$$
\begin{array}{lll}
BM \cup E' \vdash_1^S O & \wedge & BM \cup E' \not\vdash_1^S \perp \\
\wedge & & \\
BM \cup E \vdash_1^{S'} O & \wedge & BM \cup E \not\vdash_1^{S'} \perp \\
\rightarrow & & \\
BM \cup E \vdash_1^S O & \wedge & BM \cup E \not\vdash_1^S \perp \\
\wedge & & \\
BM \cup E' \vdash_1^{S'} O & \wedge & BM \cup E' \not\vdash_1^{S'} \perp
\end{array}
$$

$BM \cup E \vdash_1^S O$ and $BM \cup E' \vdash_1^{S'} O$ follow from $BM \cup E \vdash_1^{S'} O$ by theorem 1 and lemma 1 respectively. Similarly, $BM \cup E' \not\vdash_1^{S'} \perp$ and $BM \cup E \not\vdash_1^S \perp$ follow from $BM \cup E' \not\vdash_1^S \perp$ by theorem 1 and lemma 1 respectively.

The proof of the second part of the theorem has the same structure, again using theorem 1 and lemma 1. □

One final result on the relations between approximate abductive diagnoses concerns the sizes of $\mathcal{ABD}_i^S$:

Theorem 9 (Sizes of $\mathcal{ABD}_i^S$)

$$
\begin{array}{ll}
0 = & |\mathcal{ABD}_1^{\emptyset}| \leq |\mathcal{ABD}_1^{S}| \leq |\mathcal{ABD}_1^{S'}| \leq |\mathcal{ABD}_2| \\
 & |\mathcal{ABD}_2| \geq |\mathcal{ABD}_3^{S'}| \geq |\mathcal{ABD}_3^{S}| \geq |\mathcal{ABD}_3^{\emptyset}| = 0
\end{array}
$$

Proof Theorem 6 implies an injective map from $\mathcal{ABD}_1^S$ to $\mathcal{ABD}_1^{S'}$, and from $\mathcal{ABD}_3^S$ to $\mathcal{ABD}_3^{S'}$, as indicated in example 8. This immediately yields the theorem. □

The results of the main theorems in this section are summarised in table 1.

Table 1: Summarising the results of section 4.2

ABD_i^S	S	new superset diagnosis	new subset diagnosis	number
$i=1$	bigger	yes	no	more
$i=1$	smaller	no	only	less
$i=3$	bigger	no	yes	more
$i=3$	smaller	only	no	less

5 HOW TO USE APPROXIMATE DIAGNOSES

We will now briefly discuss some ways in which these results can be exploited for diagnosis. The follow-

Table 2: Behaviour of an anytime algorithm for ABD_1^S for $O = \{O_1, O_2\}$

S	nr	E comment
$\{O_1, O_2\}$	0	*no* S violates theorem 3
$+\{O_3, H_1\}$	0	*no* $H_4 \notin S$, therefore $BM \vdash_1^S \neg O_2$ therefore violation of requirement (2)
$+H_4$	6	$\{\}, \{H_1\}, \{\neg H_1\}, \{\neg H_4\}, \{H_1, \neg H_4\}, \{\neg H_1, \neg H_4\}$ H_2 and H_3 are not in S, therefore $BM \vdash_1^S O_1 \wedge O_2$ (i.e. O_1 and O_2 are explained "for free"), therefore all consistent literal-sets over H_1 and H_4 are diagnoses.
$+H_2$	6	all previous solutions extended with H_2 H_2 is now in S, so O_2 is no longer explained "for free", and therefore H_2 must be included in the solutions.
	4	$\{H_1\}, \{H_1, \neg H_4\}, \{H_1, \neg H_2\}, \{H_1, \neg H_2, \neg H_4\}$ The other explanation for O_2 is H_1 (classically this would be $H_0 \wedge H_1$, but this becomes H_1 since $H_0 \notin S$). We also obtain all consistent extensions of H_1 with letters from S
total	10	
$+H_0$	18	all previous solutions already containing H_2 (6 of them), possibly extended with H_0 or $\neg H_0$ since H_0 is now in S, it is now allowed to include H_0 or $\neg H_0$, but not required, since, $H_3 \notin S$, which still explains O_1 "for free".
	4	$\{H_1, H_0\}, \{H_1, \neg H_4, H_0\}, \{H_1, \neg H_2, H_0\}, \{H_1, \neg H_4, \neg H_2, H_0\}$ these solutions are obtained from those 4 in the previous step where H_1 was required as the explanation for O_2, (i.e. those solutions where H_2 was missing). Since H_0 is now in S, H_1 alone can no longer explain O_2, and we must add H_0.
total	22	
$+H_3$	22	all previous solutions extended with H_3 since H_3 is now in S, it is now allowed to include H_3
	6	$\{H_1, H_2, H_0, \neg H_3\}, \{H_1, \neg H_4, H_2, H_0, \neg H_3\}, \{H_1, H_0, \neg H_3\}$, $\{H_1, H_0, \neg H_4, \neg H_3\}, \{H_1, \neg H_2, H_0, \neg H_3\}$, $\{H_1, \neg H_2, H_0, \neg H_4, \neg H_3\}$
	6	$\{H_1, H_2, H_0\}, \{H_1, \neg H_4, H_2, H_0\}, \{H_1, H_0\}, \{H_1, H_0, \neg H_4\}$, $\{H_1, \neg H_2, H_0\}, \{H_1, \neg H_2, H_0, \neg H_4\}$ In the previous 22 cases, if H_3 is not required for explaining O_1 (because H_0 and H_1 are present), then we are allowed to remove H_3 or to add $\neg H_3$.
total	34	

ing gives us an anytime algorithm (Russel and Zilberstein 1991) for computing diagnoses: start by computing $\mathcal{ABD}_1^S$ for some small value of S, and iteratively increase the value of S. This will include ever more causes in the set of diagnoses $\mathcal{ABD}_1^S$. This set will be an ever better approximation of the set of classical diagnoses $\mathcal{ABD}_2$. The behaviour of this algorithm is described in table 2. This table shows which ABD_1^S solutions are computed when starting with $O = S = \{O_1, O_2\}$, and gradually adding letters to S as indicated. In the final row of the table, $S = \text{LET}(BM)$, and the diagnoses correspond to exactly the set of all 34 classical ABD_2 diagnoses. This algorithm can be interrupted at anytime, and will give monotonically improving results as run-time increases. If for example, starting by including in S only the most urgent causes, we obtain only the urgent parts of ABD_2 diagnoses. When we gradually increase S by adding less urgent causes, we obtain ever larger segments of ABD_2 diagnoses.

The complexity results in (Schaerf and Cadoli 1995) ensure that this converging computation of $\mathcal{ABD}_2$ will not be more expensive than the direct computation of $\mathcal{ABD}_2$ by classical means. However, the results of (Schaerf and Cadoli 1995) that we have applied to abduction concern the worst-case complexity of the declarative characterisation of deduction. Further study and experimentation is needed to see how well these results apply to specific algorithms for abduction known in the literature.

An alternative anytime algorithm is obtained by using ABD_3^S: Again, we start by including in S only the most urgent causes. The $\subseteq$-*min*-ABD_3^S diagnoses are now exactly those ABD_2 diagnoses which consist of only these urgent causes from S. When increasing S by including less urgent causes, we obtain more and more of $\mathcal{ABD}_2$. Thus, ABD_1^S are the urgent subsets of classical diagnoses, whereas $\subseteq$-min-ABD_3^S are only those classical diagnoses consisting entirely of urgent causes. Both of these anytime algorithms ensure that we only loose less urgent causes from our approximate diagnoses (theorem 4). Another example of the choice for S would be the most frequently occuring causes, etc.

In (ten Teije and van Harmelen 1996), we have explored some of such properties of the behaviour model BM in order to select predicate letters in S. Some examples from (ten Teije and van Harmelen 1996) are:

- *specificity of observations*: observations are more specific if they have fewer possible causes in the model. For example, in a medical context, headache or a mild fever would be aspecific symptoms, while a lump in the breast would be a very specific symptom. Beginning with the most specific symptoms in S, and gradually adding less specific symptoms, we obtain a decreasing series of diagnoses where inconsistency with specific observables is taken more seriously then inconsistency with a-specific observables. (For this example we used an approximation of a consistency-based notion of diagnosis á la (Reiter 1987), instead of the abductive notion used in this paper).
- *strength of causal connections*: following (Console and Torasso 1990), we distinguish causes which *necessarily* imply their effects from causes which only *possibly* do so, but necessarily. Using ABD_3^S, we first include only necessary causal links in S, and only add possible causal links if no diagnosis can be obtained without them. This results in an algorithm which computes the most reliable diagnoses first, before investigating diagnoses that are based on less reliable causal links.
- *structure in the causal model*: in the context of a simple causal model of an automobile in (ten Teije and van Harmelen 1996), we distinguished submodels for electrical or mechanical faults. Different sub-models can be gradually incorporated in the diagnostic process by added their letters to S.

6 CONCLUSION AND FUTURE WORK

In this paper, we have extended a widely accepted definition of diagnosis by using approximate deduction relations instead of the usual classical deduction. This has yielded a number of interesting approximate versions of diagnosis. We have proven theorems which state how these approximations can be used to increase and decrease both the total number and the individual size of the computed diagnoses, while guaranteeing certain properties (e.g no loss of classical diagnoses, or only loss of certain types of classical diagnoses). We have exploited our results in efficient anytime algorithms for computing both approximate and classical diagnoses. Finally, we have sketched how the results from this paper can be put to practical use.

As stated above, the results from (Schaerf and Cadoli 1995) ensure that the problem of iterated computation of approximate abduction is not harder (and often easier) than the problem of computing classical abduction, but these are only worst-case results that apply to the declarative presentation of the problem. Further study and experimentation is needed to see how well these results apply to specific algorithms for abduction known in the literature.

Furthermore, in this paper we have studied the existence of diagnoses ($\mathrm{ABD}_i^S(E)$, e.g. theorems 5, 6 and 8), and the set of all diagnoses ($\mathcal{ABD}_i^S$, e.g. theorems 7 and 9). Since the set of all S-diagnoses is exponential (up to 2^S), we will in general not want to compute all of these. Other properties which might be fruitfully studied in future work are: the complexity of finding the next diagnosis, testing whether a formula holds in all or some diagnoses, etc.

In section 3, we stated the restriction to causal networks of only two layers, and we argued that this restriction does not affect the expressiveness of the causal networks. Although this is true, this restriction does affect the different possibilities for approximation: since all intermediate nodes have been removed from the network, these letters can no longer be used to characterise approximations by excluding them from the set S. An example of this is the following simple causal theory T:

$$T = \{H_1 \rightarrow N, H_2 \rightarrow N, N \rightarrow O_1, H_1 \wedge H_2 \rightarrow O_2\}$$

This can be transformed to the two layer theory T':

$$T' = \{H_1 \rightarrow O_1, H_2 \rightarrow O_1, H_1 \wedge H_2 \rightarrow O_2\}$$

Although T and T' are classically equivalent, in the sense that for any sentence ϕ made from the letters H_1, H_2, O_1 and O_2 we have

$$T \vdash \phi \text{ iff } T' \vdash \phi,$$

this is not the case for approximate deduction. In particular, if $S = \mathrm{LET}(T) \setminus \{N\}$,

$$T' \cup \{H_1, H_2\} \vdash_3^S O_2 \wedge O_1$$

while

$$\begin{array}{ll} T \cup \{H_1, H_2\} \vdash_3^S O_2 & \text{but} \\ T \cup \{H_1, H_2\} \not\vdash_3^S O_1, & \end{array}$$

This shows that in T', $\{H_1, H_2\}$ is an ABD_3^S-diagnosis for both O_1 and O_2, while in T it is only an ABD_3^S-diagnosis for O_2. In fact, in T' it is impossible to find any choice for S that will yield an ABD_3^S-diagnosis for

O_2 that not also implies O_1, while in T this is possible ($S = \mathrm{LET}(T) \setminus \{N\}$ as above). This shows that two-layer networks really do restrict our options for choosing particular approximations. We are currently investigating under which conditions our results still hold for networks with intermediate layers in networks.

More speculative is the use of our results to model the iterative behaviour of various existing diagnostic systems. Such systems iterate over multiple models (Abu-Hanna 1994), different abstraction levels (Mozetic 1991, Console and Torasso 1992), or request additional observations (Mcilraith and Reiter 1992). Our claim is that such iterative behaviour can be formalised in a uniform way through our results. This would yield insights in the differences and commonalities between such systems, and would make our anytime algorithms available to these existing systems. Further work is required to make these claims more precise.

References

T. Bylander, D. Allemang, M. C. Tanner, and J. R. Josephson. The computational complexity of abduction. *Artificial Intelligence*, 49:25–60, 1991.

L. Console and P. Torasso. Hypothetical reasoning in causal models. *Int. J. of Intelligent Systems*, 5(1):83–124, 1990.

L. Console and P. Torasso. A spectrum of logical definitions of model-based diagnosis. *Computational Intelligence*, 7(3):133–141, 1991.

S. McIlraith and R. Reiter. On tests for hypothetical reasoning. In W. Hamscher, L. Console, and J. de Kleer, editors, *Readings in Model-Based Diagnosis*, pages 89–96. Morgan Kaufman, 1992.

R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–96, 1987.

S. J. Russell and S. Zilberstein. Composing real-time systems. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence, Sydney, Australia*, volume 1, pages 212–217, San Mateo, 1991. Morgan Kaufmann.

M. Schaerf and M. Cadoli. Tractable reasoning via approximation. *Artificial Intelligence*, 74(2):249–310, April 1995.

A. ten Teije and F. van Harmelen. Examples of approximations in diagnosis based on approximate entailment. In *Submitted to DX-96, Seventh International Workshop on Principles of Diagnosis*, October 1996.