Physical Impossibility Instead of Fault Models

Gerhard Friedrich, Georg Gottlob, Wolfgang Nejdl
Christian Doppler Laboratory for Expert Systems
Technical University of Vienna, Paniglgasse 16,
A-1040 Vienna, Austria — friedrich@vexpert.at

Abstract

In this paper we describe the concept of *physical impossibility* as an alternative to the specification of fault models. These axioms can be used to exclude impossible diagnoses similar to fault models. We show for Horn clause theories while the complexity of finding a first diagnosis is worst-case exponential for fault models, it is polynomial for physical impossibility axioms. Even for the case of finding all diagnoses using physical impossibility axioms instead of fault models is more efficient, although both are exponential in the worst case. These results are used for a polynomial diagnosis and measurement strategy which finds a final sufficient diagnosis.

1 Introduction

Model-based diagnosis has traditionally been based on the use of a correct behavior model. Faulty components were assumed to show arbitrary behavior modeled by an *unknown fault* mode.

An interesting extension to this approach is the inclusion of specific fault models, which have been introduced in [3] and [6]. [3] retains an *unknown fault* mode and uses fault models to assign different probabilities to different behavior modes. [6] shows how to exclude impossible diagnoses ("the light of a bulb is on although no voltage is present") by deleting the *unknown fault* mode. However, in this case the fault models have to be complete to find the correct diagnoses. While the correct model behavior can often be expressed as a Horn clause theory (with polynomial consistency checking[1]) the introduction of fault models leads to a non Horn clause theory in any case and thus to a computationally more complex algorithm for finding diagnoses.

[1] In this paper we assume a system model guaranteeing a restricted term depth of all arguments and a restricted number of argument positions. Otherwise the problem would of course be undecidable or exponential.

In this paper we investigate a third approach, which excludes impossible diagnoses by specifying physical impossibility axioms in the form of negative clauses. This approach does not enlarge the diagnosis complexity compared to a correct behavior based system, but usually excludes the same diagnoses as fault models. Starting from a Horn clause description of the correct behavior, the introduction of physical impossibility axioms retains the Horn property. On the other hand the introduction of fault models leads to a non Horn theory resulting in an exponential algorithm for finding even a first diagnosis. Our approach is therefore advantageous in cases where the additional information which can be expressed by specific fault models (like probabilities of different behavior modes) is not needed or not available.

In Section 2 we describe the concept of physical impossibility and discuss the relationship between physical impossibility and fault models. Section 3 shows the computational advantages of our approach and discusses the worst-case complexity of finding diagnoses. A polynomial algorithm for finding a final sufficient diagnosis is given which is not possible if we use fault models.

Because of the space limitations formal definitions and complete proofs are given in a longer version of this paper.

2 Physical Impossibility

To describe the notion of physical impossibility, let us first analyze the possible behavior of the components of a device. This behavior can be represented by specifying constraints between the state variables describing the component. We assume a finite domain for these variables. Each state variable can have only one value.

The domain of a component can be specified by a finite set of value tuples denoting all possible value combinations which can be assigned to the state variables of the component. The arity of a tuple is equal to the number of variables describing a component.

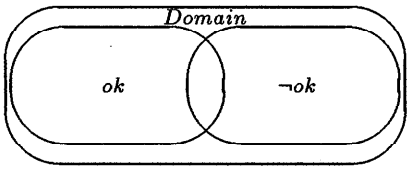


Figure 1: Tuple domain of a component including correct behavior and fault model tuples.

To diagnose a system various subsets of this domain may be specified. The relations of these sets are depicted in Figure 1. We will discuss the following sets (using $S_1 \setminus S_2$ to denote S_1 minus S_2):

- correct behavior set denoted by *ok*
- fault model set $\neg ok$
- physical impossibility set $(Domain \setminus (\neg ok \cup ok))$

In the following paragraphs we will describe how to represent correct, faulty, and impossible behavior. For each definition we show the appropriate rules used for the bulb example which has been introduced in [6]. Note, that the specific formalism used for describing a system model depends on the inference mechanism used. The general concepts defined here do not depend on it.

In Figure 2 a simple circuit is shown, consisting of a power supply and three bulbs. Wires connect these components in parallel. The domain theory specifies the correct behavior of the circuit as usual, as described below. The literal $ok(X)$ denotes that the component X works correctly.

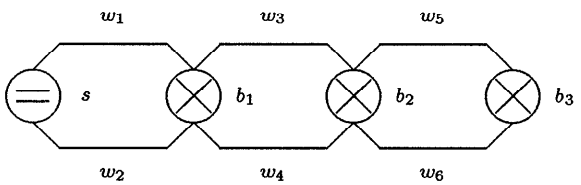


Figure 2: Three bulbs and one voltage supply in parallel

The following axioms describe the *correct behavior* of our components. Variables are stated by capital letters and are universally quantified. In order to achieve a clear and simple presentation, we assume without losing generality that wires always behave correctly.

(1) bulb(X) ∧ ok(X) ∧ val(port(X),+)
→ val(light(X),on).
(2) bulb(X) ∧ ok(X) ∧ val(port(X),0)
→ val(light(X),off).
(3) bulb(X) ∧ ok(X) ∧ val(light(X),off)
→ val(port(X),0).
(4) bulb(X) ∧ ok(X) ∧ val(light(X),on)
→ val(port(X),+).
(5) supply(X) ∧ ok(X)
→ val(port(X),+).

Values are propagated along connections, each state variable can have only one value.

val(Port1,Val) ∧ conn(Port1,Port2)
→ val(Port2,Val).
val(Port1,Val) ∧ conn(Port2,Port1)
→ val(Port2,Val).
val(Port,Val1) ∧ val(Port,Val2)
→ Val1 = Val2.

supply(s). *conn(port(s),port(b_1)).*
bulb(b_1). *conn(port(s),port(b_2)).*
bulb(b_2). *conn(port(s),port(b_3)).*
bulb(b_3).

Additionally the following observations are made:

val(light(b_1),off). *val(light(b_3),on).*
val(light(b_2),off).

The construction of conflict sets leads to four minimal conflict sets $\langle s, b_1 \rangle$, $\langle s, b_2 \rangle$, $\langle b_1, b_3 \rangle$ and $\langle b_2, b_3 \rangle$ which determine two diagnoses $[s, b_3]$ and $[b_1, b_2]$.

Now usually one would not consider b_1 and b_2 to be correct while b_3 and s are faulty producing light when there is no power supply. The additional information a diagnosis expert uses in this case is the knowledge about what is physically possible. If this knowledge is omitted (because only the correct behavior is modeled), "miracles" are possible.

Using the principle of *physical impossibility*, we simply exclude all tuples which are impossible. This can always be done by completely negative clauses and thus without adding a non Horn clause to the description of the correct behavior. In our case:

¬ (bulb(X) ∧ val(light(X),on) ∧ val(port(X),0)).

Using domain closure axioms stating that a light can be *on* or *off* and that the voltage can be 0 or +, we get the following rules, which subsume rule 2 and 4 of our correct behavior rules. They can be used instead of the physical impossibility axiom.

bulb(X) ∧ val(light(X),on) → *val(port(X),+).*
bulb(X) ∧ val(port(X),0) → *val(light(X),off).*

On the other hand, a *fault model* consists of the following axiom to eliminate the undesirable diagnosis $[s, b_3]$:

$$bulb(X) \wedge \neg\ ok(X) \rightarrow (val(port(X),0) \wedge val(light(X),off)) \vee (val(port(X),+) \wedge val(light(X),off)).$$

This axiom can be simplified to

$$bulb(X) \wedge \neg\ ok(X) \quad \rightarrow \quad val(light(X),off).$$

The introduction of fault models into a Horn theory describing the correct behavior always leads to a non Horn theory. (If we use the literal $ab(X)$ denoting abnormal behavior, we have to include the additional axiom $ab(X) \leftrightarrow \neg ok(X)$).

Both approaches reduce the conflict sets $\langle s, b_1\rangle$, $\langle s, b_2\rangle$ to $\langle b_1\rangle$ and $\langle b_2\rangle$. This results in the elimination of diagnosis $[s, b_3]$. The reason for the conflict set reduction using the fault model approach is that $ok(b_3)$ can be deduced without assuming it, since the light is *on*. Therefore each single assumption $ok(b_1)$ and $ok(b_2)$ is inconsistent with the system description and the observations. By using physical impossibility, we simply exclude the possibility, that a light is *on* with no voltage present. Transforming the physical impossibility axiom using the domain axioms even lets us directly deduce the presence of voltage. (Note, that we use the domain axioms only during transformation, not for the final generation of the model.)

This example suggests an equivalence between fault models and physical impossibility axioms. This equivalence can be formally described by the following theorem:

Theorem 1 If the domain and the model of correct behavior is represented and $\neg ok(X)$ only appears in the clauses representing the correct behavior (e.g., $ok(c_i) \rightarrow \ldots$) then the additional specification of a fault model is equivalent to the additional specification of the physical impossibility axioms for the task of diagnosis.

We use the usual component oriented description and the assumption that faults are independent from each other. Rules like $\neg ok(c_i) \rightarrow \neg ok(c_j)$ are excluded.

Proof (informal): Using domain axioms and the axioms describing correct and faulty behavior we can deduce the physical impossibility axioms. No additional conflict will result if we add the physical impossibility axioms to the system model.

On the other hand using domain axioms, correct behavior and physical impossibility axioms (which are specified by negative clauses), we can deduce the possible behavior. Additionally for each component c_i in a diagnosis we can deduce $\neg ok(c_i)$ otherwise the diagnosis would not be minimal. This can be deduced only by using the correct behavior clauses, as $\neg ok(c_i)$ does not appear in any other clauses. Therefore every correct behavior tuple leads to a contradiction. Using the possible behavior, we can now derive at least a subset of the faulty behavior subsuming the fault model.

Let us denote the correct behavior axioms as $\mathcal{B}_C$, the faulty behavior axioms as $\mathcal{B}_F$, the physical impossibility axioms as $\mathcal{B}_I$ and the domain axioms as $\mathcal{D}$.

Then we can write the equivalence of fault models and physical impossibility axioms for the purpose of finding all diagnoses somewhat informally as

$$\mathcal{B}_C \cup \mathcal{D} \cup \mathcal{B}_I \cong \mathcal{B}_C \cup \mathcal{D} \cup \mathcal{B}_F$$

In most systems (especially those based on value propagation) only Horn clauses are used for describing correct and faulty behavior modes. Explicit domain axioms are not included in the system model. Notwithstanding the potential incompleteness caused by this omission, we usually use such a simplified theory to avoid combinatorial explosion. Its incompleteness with respect to diagnosis decreases with an increasing set of measurements.

What we would therefore like to prove is the following equivalence of physical impossibility and fault models (without domain axioms):

$$\mathcal{B}_C \cup \mathcal{B}_I \cong \mathcal{B}_C \cup \mathcal{B}_F$$

Although this is indeed valid in many cases it is possible to construct situations where the addition of Horn clause fault models yields a more complete theory than the addition of physical impossibility. If domain axioms are omitted physical impossibility axioms can therefore only be a reasonable approximation. However, although Horn clause fault models yield better results in some cases, they are themselves an approximation (except if completely unrestricted clauses are used).

In Section 3 we will show that physical impossibility axioms do not degrade the efficiency of the diagnosis algorithm. We can still construct a polynomial algorithm finding a final sufficient diagnosis for such a theory. On the other hand we show that fault models are intractable, something we wanted to avoid when we excluded the domain axioms initially. So we are faced once again with the well-known completeness/efficiency tradeoff often encountered in AI.

3 Efficiency and Complexity

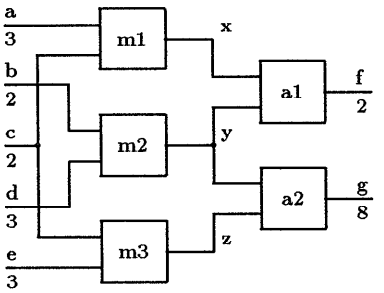
3.1 Efficiency Considerations

To allow efficient consistency checking and diagnosis generation, we use Horn clauses for our system model as much as possible. This corresponds to the use of value propagation as inference engine. Usually only the

subset of the correct behavior which can be expressed by functional dependencies is used in the system model.

It is clear that by extending such a Horn clause theory by physical impossibility axioms (which are negative or definite clauses) we do not increase the complexity of the diagnosis process, while even the inclusion of Horn clause fault models automatically makes the theory non Horn, leading to the well-known combinatorial explosion.

Example 1 We use the standard d74 circuit depicted in Figure 3.1 with six different behavior modes (as used in [4], see Figure 3.1).



1. output is correct
2. output is zero
3. output is left input
4. output is right input
5. output is one
6. output is shifted left one bit

Figure 3: D74-circuit

We do not use an *unknown fault* mode, as such a mode would allow any possible behavior. Such a fault mode is therefore only interesting if we rely on probability ranking.

Initial measurements are $a = d = e = 3$, $b = c = 2$, $f = 2$ and $g = 8$. The two double fault diagnoses for these measurements are $[(a1, right), (m2, left)]$ and $[(m1, zero), (m2, left)]$. Using physical impossibility we get the same diagnoses as using fault models, without the additional fault mode information. However, if we just want to change the faulty components, the exact fault mode is irrelevant. (In this example, the physical impossibility axioms define that any behavior not covered by the behavior modes is inconsistent.)

Using the MOMO system described in [5] we got the following normalized model generation times (for finding all diagnoses):

- physical impossibility (4 or 6 modes): 1
- fault models (first 4 modes): 7.6
- fault models (all 6 modes): 22.9

In this example we achieved a runtime improvement factor of 22.9 by using physical impossibility axioms instead of fault models. Note, that this does not depend on our algorithm, but simply mirrors the combinatorial explosion caused by the non Horn theory (see also [4]). Each fault model introduces alternative rules used for value propagation and we have exponentially many combinations of fault models.

On the other hand physical impossibility is barely affected by the introduction of additional fault models, as only the checks to exclude impossible diagnoses get slightly more complicated. No new values are deduced because of the physical impossibility axioms. Consistency has to be checked only for a Horn clause theory.

3.2 Complexity

In the following we will concentrate on Horn clause theories for the correct and faulty behavior and the physical impossibility axioms. This is sufficient for most cases and usually used by value propagation systems. It also allows us to capture all functional dependencies.

For consistency based model-based diagnosis we can state the following complexity theorems. They are independent of the inference strategy used.

Theorem 2 Assume a description of the correct behavior by a (propositional) Horn clause theory, a set of observations and a set of (already found) diagnoses $\mathcal{D}$. The complexity of deciding whether *a next diagnosis* exists which is not in $\mathcal{D}$ is $\mathcal{NP}$-complete.

Proof (informal): The problem is obviously in $\mathcal{NP}$. By reduction to SAT we can show that it is also $\mathcal{NP}$-complete.

Let C be a set of propositional clauses in SAT form. Let U be the set of variables used in C. Assume further for each x ($\neg x$) that there exists at least one clause $c \in C$ such that x ($\neg x$) does not occur in c. We use the following instance of the next diagnosis problem $ND = \langle COMP, SD, OBS, D \rangle$ consisting of a set of components, a system description, a set of observations and of diagnoses.

$$
\begin{aligned}
COMP &= \{x, \bar{x} \mid x \in U\} \\
SD &= G_1 \cup G_2 \cup G_3 \cup G_4 \\
OBS &= \{\neg z\} \\
D &= \{\{x, \bar{x}\} \mid x \in U\} \\
G_1 &= \left\{\left[\textstyle\bigwedge_{x \in c} ok_x \wedge \bigwedge_{\neg x \in c} ok_{\bar{x}}\right] \rightarrow f \mid c \in C\right\} \\
G_2 &= \{ok_x \rightarrow \hat{x}, ok_{\bar{x}} \rightarrow \hat{x} \mid x \in U\} \\
G_3 &= \left\{\textstyle\bigwedge_{x \in U} \hat{x} \rightarrow a\right\} \\
G_4 &= \{a \wedge f \rightarrow z\}
\end{aligned}
$$

For a diagnosis $\Delta \notin D$ the following truth value assignment satisfies C:

$$\forall u \in U : \varphi(u) = \begin{cases} true & if\ u \in \Delta \\ false & otherwise \end{cases}$$

Using this assignment we can show

$$C\ satisfiable \Longleftrightarrow \exists \Delta \notin D\ for\ ND$$

This complexity theorem is valid if the system description includes just a model of correct behavior consisting of propositional Horn clauses. Extending the model by fault models or physical impossibility axioms can not decrease the complexity.

Sometimes it is sufficient to find just one initial diagnosis, especially if we take various repair or measurement strategies into account. Let us therefore compare the complexity of this problem for physical impossibility and fault models.

While both physical impossibility and fault models exclude impossible diagnoses, the difference between them is, that the use of a fault model also influences the candidate space and the use of physical impossibility does not. This is expressed by the following theorem:

Theorem 3 If we add physical impossibility axioms to the correct behavior model, each superset of a diagnosis is consistent.

Proof (informal): No clause from the description of the correct behavior and the physical impossibility axioms contains the positive literal $ok(c)$. Only negative literals $\neg ok(c)$ appear in the clauses describing the correct behavior. Therefore adding $\neg ok(c)$ for some component c to a diagnosis can not lead to a contradiction as we cannot derive $ok(c)$ from the given theory.

We can define a polynomial algorithm for a system description consisting of correct behavior and physical impossibility axioms to find a diagnosis:

Algorithm 1 (Finding the First Diagnosis)

1. Take the candidate which assumes all components to be faulty. This candidate has to be correct otherwise the system description itself is inconsistent.
2. Now remove an arbitrary component from the candidate, i.e. assume the component to be correct. The component has to be chosen in such a way that the remaining candidate is consistent. Components need only be checked once. In a value propagation system new values may be deduced for each component which is assumed to be correct. If the theory proves to be inconsistent, these values have to be retracted.
3. Do this until no more components can be removed from the candidate (i.e. all components have been tried). The (minimal) candidate found can be output as first diagnosis.

Proof (informal): As the candidate space is contiguous, algorithm 1 always finds a minimal candidate. The inclusion of $ok(C)$ is monotonous so the algorithm performs exactly n consistency checks.

Note, that checking consistency of all single faults by a simple algorithm exhibits also a worst-case complexity of n and a average case complexity of $n/2$, if we set the cost for a consistency check to 1. If we use conflict sets to compute the single faults the complexity is exponential in the worst case. (Consider the case, where we have exponential many conflict sets.)

Finding the first diagnosis using a system description with several incompatible behavior modes is exponential in general. For fault models which do not exclude any diagnosis compared to the correct behavior model alone (e.g. if the *unknown fault* mode is included), we can find the first diagnosis in polynomial time simply by deleting all fault model axioms.

Theorem 4 Let us assume, that we extend the description of the correct behavior by clauses describing the faulty behavior and that these clauses include the positive literal $ok(c_i)$ for the described components c_i, which appears in negative form in the correct behavior clauses. Then deciding whether a first diagnosis exists is $\mathcal{NP}$-complete.

Proof (informal): The proof is very similar to the *next diagnosis problem*. We transform sets of assumptions which are inconsistent (like $\{ok(c_i), \neg ok(c_i)\}$) into already found diagnoses. (By the way, even deciding whether there exists an arbitrary consistent candidate is $\mathcal{NP}$-complete.)

Similar results to theorem 2 and 4 have been shown in an interesting paper of Bylander et al ([1]) in the context of abductive reasoning. However, the transformation from a consistency-based diagnosis problem into an abductive one sketched in their paper using conflict sets is not preferable, as the number of conflict sets can grow exponentially resulting in an exponential algorithm for the transformed problem.

3.3 Polynomial Diagnosis Strategies

The results described in the previous section indicate the complexity of the consistency based diagnosis problem. However, it is still possible to define a polynomial diagnosis algorithm for finding a sufficient[2] diagnosis

[2]By *sufficient* we mean a correct diagnosis we want to accept as the final one depending on some termination criterion.

by using our first diagnosis algorithm for correct behavior and physical impossibility axioms.

Unfortunately, a measurement selection function derived from entropy (e.g. [2], [3]) tries only to minimize the number of measurements (and therefore measurement costs). What is not included in the minimization process are the inference costs which, however, can get exponential. We have to use measurement selection heuristics, which need to compute only one diagnosis.

Algorithm 2 Polynomial algorithm for finding a sufficient diagnosis (if correct behavior and physical impossibility rules are given):

1. Find the first diagnosis using all available observations (algorithm 1).
2. If the diagnosis found fulfills the termination criterion, then exit. This could be the case if we can prove the components included in the diagnosis to be faulty without assuming the correctness of other components. In other cases, an immediate repair may be more cost efficient than further testing.
3. Take additional actions to get new information such as
   - Take one or more additional measurements.
   - Try to prove a component to be correct or faulty.
   - Replace a component by a good one, etc.

   Which strategy we take and which measurements we choose may depend on the conflicts found so far, the failure probability of the components, cost of testing etc. Trying to prove or disprove the current diagnosis is also a good heuristic. If we can prove a component c_i to be correct for the given exogenous variables (i.e. by measuring its direct inputs and outputs), it can be excluded from a diagnosis. We can assume $ok(c_i)$ for such a component. This might also be done by using an internal test. Replacing a component by a good one has usually the same effect.
4. goto 1

The difference to the algorithm used in [2] and similar algorithms is that only one diagnosis is computed at each iteration. As only polynomially many measurement points exist and the number of consistency checks is polynomial, the algorithm halts in polynomial time.

4 Conclusion

We have described the concept of physical impossibility as an alternative to fault models. Compared to fault models physical impossibility axioms result in a more efficient computation of diagnoses. We also described a polynomial algorithm for finding the first diagnosis using physical impossibility axioms. The inclusion of fault models even into a Horn clause system model was shown to lead to a $\mathcal{NP}$-complete decision procedure to check for a first diagnosis. For both finding the next diagnosis is exponential in general. By relaxing the optimality criterion for measurement selection as defined in [2] we are able to define a simple algorithm for finding a final sufficient diagnosis in polynomial time using correct behavior and physical impossibility axioms.

Acknowledgements

We thank Peter Struss, Oskar Dressler, Hartmut Freitag, Olivier Raiman, and Johan de Kleer for their comments to a previous version of this paper.

References

[1] Tom Bylander, Dean Allemang, Michael C. Tanner, and John R. Josephson. Some results concerning the computational complexity of abduction. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, pages 44–54, Toronto, May 1989. Morgan Kaufmann Publishers, Inc.

[2] Johan de Kleer and Brian C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130, 1987.

[3] Johan de Kleer and Brian C. Williams. Diagnosis with behavioral modes. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1324–1330, Detroit, August 1989. Morgan Kaufmann Publishers, Inc.

[4] Oskar Dressler and Adam Farquhar. Problem solver control over the ATMS. In *Proceedings of the German Workshop on Artificial Intelligence*, pages 17–26, Eringerfeld, September 1989. Springer-Verlag.

[5] Gerhard Friedrich and Wolfgang Nejdl. MOMO — Model-based diagnosis for everybody. In *Proceedings of the IEEE Conference on Artificial Intelligence Applications*, Santa Barbara, March 1990.

[6] Peter Struss and Oskar Dressler. Physical negation — Integrating fault models into the general diagnostic engine. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1318–1323, Detroit, August 1989. Morgan Kaufmann Publishers, Inc.