Safe Partial Diagnosis from Normal Observations

Roni Stern
Software and Information Systems Engineering
Ben Gurion University
Beer Sheba, Israel
sternron@post.bgu.ac.il

Brendan Juba
Dept. of Computer Science & Engineering
Washington University in St. Louis
1 Brookings Dr., St. Louis, MO, 63130 USA
bjuba@wustl.edu

Abstract

Model-based diagnosis (MBD) is difficult to use in practice because it requires a model of the diagnosed system, which is often very hard to obtain. We explore theoretically how observing the system when it is in a normal state can provide information about the system that is sufficient to learn a partial system model that allows automated diagnosis. We analyze the number of observations needed to learn a model capable of finding faulty components in most cases. Then, we explore how knowing the system topology can help us to learn a useful model from the normal observations for settings in which many of the internal system variables cannot be observed. Unlike other data-driven methods, our learned model is safe, in the sense that subsystems identified as faulty are guaranteed to truly be faulty.

Introduction

Model-based diagnosis (MBD) is an approach for automated diagnosis in which a *model* that describes the diagnosed system's expected behavior is assumed to be given as input. This model is used to infer possible explanations – *diagnoses* – for an observed abnormal system behavior. MBD is a well-established, principled, approach for automated diagnosis that has been studied in the Artificial Intelligence literature for decades, and has been applied in practice (Williams and Nayak 1996; Struss and Price 2003; Feldman et al. 2013). However, a key inhibitor to its widespread use is that obtaining a model of the diagnosed system is often impossible or prohibitively expensive.

Spectrum-based fault localization (SFL) (Abreu, Zoeteweij, and Van Gemund 2009; Gupta et al. 2014) provides an alternative approach for diagnosis that does not require modeling the system behavior. SFL considers the observations collected at the time when the system has failed, and finds diagnoses by "discovering statistical coincidences between system failures and the activity of the different parts of a system" (Abreu et al. 2009). While SFL is lightweight, its diagnostic accuracy is inherently limited compared to model-based approaches. Moreover, SFL approaches do not improve over time, and the number of observations required to obtain accurate diagnoses is currently not fully understood.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Learning approaches for diagnosis (Qin 2012; Keren, Kalech, and Rokach 2011; Hafez, Ross, and Gadd 1997) learn from observations of past failures how to map various observed features to the correct diagnosis. These methods usually rely on having a set of abnormal observations, each associated with its correct diagnosis. Since most systems function properly most of the time, obtaining a large set of abnormal observations and their diagnosis is often very difficult. Moreover, given that such observations are inherently abnormal, the kinds of assumptions underlying data-driven techniques, such as stationarity of the data, are questionable for diagnosis. By contrast, normal observations, i.e., observations in which the system is behaving as planned, are easily obtained in practice. Indeed, systems often collect an abundance of data while operating, to be used for monitoring and optimization purposes. Such data contains valuable information about the (mostly) normal behavior of the monitored system. In this work we explore theoretically how such data can be used for diagnostic purposes.

Our first contribution is a method for learning a partial model from normal observations that can be used for diagnosis. In particular, this partial model enables identifying some components as *surely faulty* and other components as *probably normal*, where surely faulty components must be faulty and probably normal components cannot be in any (minimal) diagnosis. Our learning method is *safe*, in the sense that "surely faulty" components are indeed faulty, in contrast to standard data-driven methods that cannot provide such a guarantee. As can be expected, the effectiveness of this learned model strongly depends on how visible the internal variables of the systems are, and on the number of available normal observations. **Our second contribution** is a statistical analysis of the number of normal observations needed to learn an effective model. This novel form of statistical analysis draws from the well-known probably approximately correct (PAC) framework for analyzing learning algorithms (Vapnik and Chervonenkis 1971; Valiant 1984; Haussler 1992). Our main theoretical result is that under several assumptions, the number of observations needed to learn, with high probability, a model that will identify faults in most cases is only linear in the number of components. **Our third contribution** is a method for using knowledge about the system topology, in addition to the set of normal observations, to learn a useful and more informed par-

tial model of the system. Here too, we provide theoretical bounds relating the number of observations and properties of the system topology to the probability of learning a model that can identify faulty components in abnormal observations with high probability.

We demonstrate our learning approach on a Boolean circuit from the standard ISCAS '85 benchmark (Brglez, Bryan, and Kozminski 1989), showing that with a single probe and only 256 observations, we are able to find a faulty component in 28% of the cases, and this percentage increases rapidly. Importantly, our approach is not specific to Boolean circuits, and it applies to any discrete-valued system, and thus applies to the wide range of systems that can be qualitatively modeled.

Background: Model-Based Diagnosis

There are three key entities in MBD: the set of system components that comprise the diagnosed system (COMPS), a model of the system (SD), and the observed system behavior (OBS). At a given point in time, each component $c \in$ COMPS is either *healthy* or *faulty*, represented by the predicate $h(c)$, which is true iff c is currently healthy. We focus on systems in which the components perform some function, i.e., each component accepts a set of values as input and produces a single output (possibly vector-valued). The relation between the input values and the output values is referred to as the *behavior* of the component. The normal behavior of c, i.e., the behavior when c is healthy, is denoted by $\varphi_h(c)$ and assumed to be deterministic, i.e., for the same inputs we expect the same outputs. We briefly discuss how to relax this assumption in the last section.

The *system's topology* is a graph whose nodes are the components and there is an edge from component c_i to c_j iff at least one output of c_i is an input of c_j. We say that a system model is *behavioral* if it contains information about the behavior of the components in addition to the system topology. A common behavior system model in the MBD literature is the *weak fault model* (WFM), in which we have rules capturing the normal behavior ($\varphi_h(\cdot)$) of all of the components. The WFM system model SD_{WFM} is given by:

$$\text{SD}_{\text{WFM}} = \bigwedge_{c \in \text{COMPS}} (h(c) \rightarrow \varphi_h(c)) \qquad (1)$$

The union of all the components' inputs and outputs represent the set of *system variables*. The values of these variables may change over time. The *system inputs*, denoted SYSINS, are all the components' inputs that are not the output of any component in the system. Thus, the values of the system inputs are set externally by whoever is using the system. The system's *internal variables* are all the components' inputs that are not system inputs, i.e., the inputs whose values are outputs of the system's components. The *system outputs*, denoted SYSOUTS, are the components' outputs that are visible from outside the system. The internal variables and the system outputs are not necessarily disjoint, as there may be internal variables that are also exposed outside of the system. Such internal variables are sometimes called "probes", and are especially useful for diagnosis. An *observation* is an assignment of values to SYSINS and SYSOUTS.

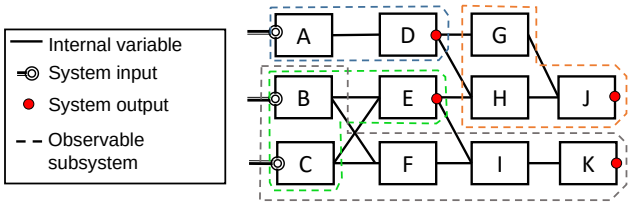

Figure 1: An illustration of a system. Rectangles are components, lines connect outputs to inputs (going from left to right), and circles are the system inputs and outputs. The minimal observable subsystem of each system output is marked with a dashed rectangle.

The vector of the values of SYSINS and SYSOUTS in an observation obs are denoted by $in(obs)$ and $out(obs)$, respectively. The inputs and outputs of a component c in obs are denoted by $in(c, obs)$ and $out(c, obs)$, respectively.

Figure 1 illustrates a system with 11 components, represented by rectangles, 3 system inputs (the inputs of A, B, and C), represented by empty circles, and 4 system outputs (the outputs of D, E, J, and K), represented by red circles. A line connecting the two components represents that an output of the component on the left is an input to the component on the right. An observation obs is thus an assignment of values to $in(A, obs)$, $in(B, obs)$, $in(C, obs)$, $out(D, obs)$, $out(E, obs)$, $out(J, obs)$, and $out(K, obs)$. The outputs of components D and E are *probes*, since they are system outputs and provide inputs to other components.

A diagnosis problem arises when there is an observation obs_{ab} that indicates the system is behaving abnormally. The task in consistency-based diagnosis is to find a hypothesis about which components are faulty that is consistent with the observation and with our knowledge of the system (Reiter 1987). Such a hypothesis is called a *diagnosis*, and is defined as follows:

Definition 1 (Diagnosis)**.** *A set of components ω is a diagnosis iff*

$$\bigwedge_{c \in \text{COMPS} \setminus \omega} h(c) \wedge obs_{ab} \wedge \text{SD} \not\models \bot$$

A diagnosis is minimal iff no proper subset of ω is a diagnosis.

MBD algorithms accept SD, COMPS and obs_{ab} as input and output one or more diagnoses. Minimal diagnoses are often preferred, following Occam's razor and due to the potentially exponential number of possible diagnoses.[1]

Definition 2 (Conflict)**.** *A* conflict *is a set of components γ such that*

$$\bigwedge_{c \in \gamma} h(c) \wedge obs_{ab} \wedge SD \models \bot$$

[1]Definition 1 follows the *consistency-based* approach to diagnosis. In *abductive* diagnosis, the diagnosis is *derived* from (as opposed to only being consistent with) obs_{ab} and SD (Console and Torasso 1991), and is beyond our scope.

Conflicts are useful in diagnosis because every diagnosis is a hitting set of all conflicts (Reiter 1987). Conflict-directed MBD algorithms find diagnoses by searching in the space of possible hitting sets of the found conflicts (Williams and Ragno 2007; de Kleer and Williams 1987; Reiter 1987). Roughly speaking, finding conflicts, and in particular minimal ones, helps to speed up conflict-directed MBD algorithms as it reduces the size of their search space.

Model-based Diagnosis without a Model

In this work, we study a setting in which the system model SD_{WFM} is not available to the diagnostician. Instead, a set of *normal observations* is available, collected when the system was working properly. Thus, they represent observations of the system when all of its components were healthy. We now show how these observations can help us find faulty components, and provide a statistical analysis of how likely this approach is to succeed.

Full Probing

The first setting we analyze is where the output of every component is also a system output, i.e., $\forall c \in$ COMPS we have that $out(c) \in$ SYSOUTS. We refer to this setting as *full probing*, since it corresponds to putting a *probe* after every component. Probing components may be very costly and impractical in many scenarios, but we analyze this setting first as a step towards the partial probing settings analyzed later in this paper.

Let OKS denote the given set of normal observations. The partial model we can learn in the full probing setting from OKS is a conjunction of statements about how each component is expected to behave when healthy. To formalize this, we introduce the predicate $f(c, i, o)$, which represents the following statement: if the input vector to component c is i then its output vector must be o. The partial model we can learn in the full probing setting, denoted SD_{FP}, is given by

$$\bigwedge_{\substack{obs_{OK} \in \text{OKS}, \\ c \in \text{COMPS}}} \Big(h(c) \rightarrow f\big(c, in(c, obs_{OK}), out(c, obs_{OK})\big) \Big) \tag{2}$$

SD_{FP} states that if a component is healthy and it receives the same input values as it had in one of the normal observations, then we expect it to output the same value as it had in that observation. SD_{FP} can be viewed as an *incomplete* version of SD_{WFM}, as follows.

Proposition 1. $SD_{\text{WFM}} \models SD_{FP}$

Proof. SD_{FP} describes the normal behavior of all components for a subset of their possible input values, while SD_{WFM} describes the normal behavior of all components for all possible input values. Thus, for every observation obs and normal observation $obs_{OK} \in$ OKS it holds that $\varphi_h(c) \rightarrow f(c, in(c, obs_{OK}), out(c, obs_{OK}))$. Thus, (1) entails (2) as required. □

One can compute diagnoses and conflicts with respect to this partial model (see Definitions 1 and 2), but will these diagnoses and conflicts be useful? The following general observation, which follows from the definitions of diagnoses and conflicts (Def. 1 and 2), describes the relationship between conflicts and diagnosis of SD_{WFM} and SD_{FP}.

Proposition 2. *For every pair of system models SD and SD' and an abnormal observation obs_{ab}, if $SD \models SD'$ then (1) a conflict w.r.t. SD' is a conflict w.r.t. SD, and (2) a diagnosis w.r.t. SD is a diagnosis w.r.t. SD'.*

A diagnosis w.r.t. SD_{FP} may not be a diagnosis for SD_{WFM}. For example, if we do not have any normal observations, then SD_{FP} is empty, and consequently assuming all components are healthy is a diagnosis w.r.t. SD_{FP} in this case. Clearly, this is not desirable. In general, in this work we consider SD_{WFM} as a reference, and unless stated otherwise use the term diagnosis to refer to diagnoses w.r.t. SD_{WFM}.

Definition 3 (Surely Faulty and Normally Behaving)**.** *A component c has a* surely normal input *in obs_{ab} if $\exists obs \in$ OKS such that $in(c, obs) = in(c, obs_{ab})$. A component c with a surely normal input in obs_{ab} is called* surely faulty *if $out(c, obs_{ab}) \neq out(c, obs)$ and is called* normally behaving *otherwise.*

That is, a surely faulty component is one that has the same inputs in the abnormal observation obs_{ab} as in some other normal observation $obs \in$ OKS, but has a different output. Since we assume components are deterministic when they are healthy, a surely faulty component c cannot be healthy. Thus, c represents a conflict w.r.t. SD_{FP} and, due to Proposition 2, also w.r.t. SD_{WFM}.

Normally behaving components serve a different purpose. All we know about these components is that they behaved in obs_{ab} just like they did in a normal observation. Faulty components can behave intermittently, i.e., behave as healthy components in some observations (de Kleer 2009). Thus, a normally behaving component may actually be faulty. However, since it followed its normal behavior in obs_{ab}, assuming it is faulty does not help explaining the abnormal system behavior. The following observation summarizes the above discussion.

Proposition 3. *Every diagnosis must contain all of the surely faulty components, and every diagnosis that contains a normally behaving component is not minimal.*

Proposition 4 (At least one surely faulty)**.** *If there is $obs \in$ OKS such that $in(obs) = in(obs_{ab})$ then we will find at least one component that is faulty.*

Proof. Since obs and obs_{ab} have exactly the same input, obs_{ab} is abnormal, and we observe every input and output of every component, then there must be at least one component that has a surely normal input and its output is different from its normal output for that input. Thus, this component is surely faulty. □

Example. Consider the simple Boolean circuit depicted in Figure 2. Now, assume we are given one normal observation $obs = \{i_1 = 1, i_2 = 1, z_1 = 0, z_2 = 0, o = 0\}$ and an abnormal observation $obs_{ab} = \{i_1 = 1, i_2 = 1, z_1 = 1, z_2 = 0, o = 1\}$. The only diagnosis w.r.t. SD_{WFM} is $\{A, C\}$. With the learned model, we can infer that A is surely faulty and B is normally behaving, since both have surely normal inputs. C does not have a surely normal input, so we cannot classify

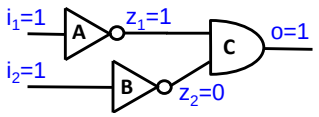

Figure 2: A simple Boolean circuit demonstrating Prop. 3.

it as either surely faulty or normally behaving. This highlights that while the set of surely faulty components must be faulty, it is not necessarily a diagnosis. In fact, there may be diagnoses that cannot be found using only normal observations.

Nonetheless, a direct implication of Proposition 4 is that given a sufficient number of normal observations, we will be able to find at least one faulty component for *every* diagnosis problem.

However, in non-trivial systems, the set of all possible combinations of system input values is extremely large. A key question is how many normal observations are needed to be able to find a faulty component in *most* diagnosis problems. We analyze this in the next section.

Distribution of Normal System Inputs In the following analysis we focus on faults that are caused by malfunctions in the components themselves, and not due to abnormal system inputs. That is, we assume that the system inputs of obs_{ab} are standard system inputs. Another assumption we make is that the system inputs in each of the given observations – normal and abnormal – are drawn independently from the same distribution, denoted by D.

In the normal observation, this distribution over the system inputs induces a distribution over the input values for every component c, denoted D_c. Let $P(c, i)$ be the probability of component c getting input i in a normal observation according to D_c. We denote the domain of possible input values for component c by $dom(c, in)$ and the largest domain size by V, i.e., $V = \max_{c \in \text{COMPS}} |dom(c, in)|$.

Lemma 1 (Common Inputs)**.** *For every system with n components and every $\epsilon, \delta \in (0, 1)$, after observing*

$$m \geq \frac{1}{\epsilon} \ln \frac{n \cdot V}{\delta}$$

normal observations, then with probability at least $1 - \delta$ we have observed every input value i for every component c such that $P(c, i) \geq \epsilon$.

Proof. Consider an input value i for a component c in a normal observation such that $P(c, i) \geq \epsilon$. Since each observation is independent, the probability that we do not observe i in any of the m observations is at most

$$(1 - \epsilon)^m \leq (1 - \epsilon)^{\frac{1}{\epsilon} \ln \frac{n \cdot V}{\delta}} \leq e^{-\ln \frac{n \cdot V}{\delta}} = \frac{\delta}{n \cdot V} \qquad (3)$$

By a union bound over the at most $n \cdot V$ pairs of components and input values, we have that with probability $1 - \delta$, every component c and input value i for which $P(c, i) \geq \epsilon$ will be observed in one of the m normal observations □

Theorem 1. *Given an abnormal observation obs_{ab} and m normal observations, where $m \geq \frac{1}{\epsilon} \ln \frac{n \cdot V}{\delta}$, then with probability at least $1 - \delta$, the learned model SD_{FP} is sufficient to identify at least one surely faulty component with probability at least $1 - \epsilon$.*

Proof. The *scope* of a component c, denoted scope(c), is the set of components that affect its input, i.e., have a path to c in the system's topology. Since obs_{ab} is an abnormal observation and we assumed the system inputs are normal, then there is at least one faulty component c_f such that its scope is either empty or contains only healthy components ($h(c')$ for every $c' \in scope(c_f)$), and (2) its output is different from its normal output, i.e., $\varphi_h(c_f) \wedge in(c_f, obs_{ab}) \wedge out(c_f, obs_{ab}) \models \bot$.

If c_f has a surely normal input in obs_{ab}, then it must be surely faulty (Definition 3), and thus using SD_{FP} we can immediately identify that c_f is faulty. If c_f does *not* have a surely normal input, then there is a component $c \in scope(c_f)$ for which there is no normal observation obs where $in(c, obs) = in(c, obs_{ab})$. Following Lemma 1, this means $P(c, in(c, obs_{ab})) < \epsilon$, i.e., this event occurs with probability smaller than ϵ. □

Note that the number of normal observations required grows only logarithmically with the number of system components (and linearly with the fan-in). For example, consider a system that consists of 100 Boolean gates, each having a fan-in of 2. Assume we wish to learn, with probability at least 0.9, sufficient information to allow us to identify a faulty component in at least 0.9 of the observations. This corresponds to setting $V = 2^2 = 4$, $n = 100$, $\delta = 0.1$, and $\epsilon = 0.1$. Thus, the number of normal observations required is $\frac{1}{0.1} \cdot \ln \frac{400}{0.1} \approx 83$ normal observations. Increasing the system size from 100 components to 1,000 components requires 106 normal observations, which is slightly more than 20% more observations than the result for 100 components.

Partial and No Probing

Next, we analyze the setting where we do not observe the outputs of all of the components. Consider first the extreme case where we have no internal probes, i.e., there is no system output that is also an input to some other component. We refer to this as the *no probing* setting. To determine whether a component is surely faulty or surely normal for a given observation, we must be able to observe both its inputs and some of its outputs. We refer to such a component as a *observed* component. There are no observed components in a no probing setting.[2] Between full probing and no probing is the *partial probing* setting, in which only a subset of the components are observed. The system description SD_{FP} in the partial probing setting is defined as in Eq. 2, but the conjunction is only over the observed components. Thus, in the no probing setting SD_{FP} is empty, and we cannot use it to infer diagnoses.

[2]Strictly speaking, a system can contain components that are not connected to any other component. Such components, are, in fact, observed components, and thus can be identified as surely faulty even without probes.

In the partial probing setting and given enough normal observations, we may be able to identify the observed components as surely faulty or normally behaving in a given abnormal observation. However, it is not clear how likely is it for such identification to occur. The statistical analysis given for the full probing setting does not transfer nicely to the partial probing setting. Lemma 1 by itself is still true in partial probing: given a sufficient number of examples ($m \geq \frac{1}{\epsilon} \ln \frac{n \cdot V}{\delta}$), with high probability ($\geq$1-δ) we observe every common ($P(c, i) \geq \epsilon$) input value of each component. However, since we do not observe the inputs of all of the components, we cannot estimate how likely we are to be able to identify a surely faulty component in an abnormal observation (Theorem 1). The key difference between the full and partial probing settings is that in partial probing we may not have a component that outputs an abnormal value and has a surely normal input, since some of the components in a partial probing setting are not observed.

Using Topology and Normal Observations

So far, we have ignored an important and commonly available source of information about the system: its topology. Lamperti and Zanella (2012) explored how system topology can be used to perform consistency-based diagnosis and how it relates to MBD with a behavioral model. Here we extend their discussion to consider diagnosis with normal observations and a topology.

The system topology allows us to extend the notion of a *observed component* to an *observed subsystem*. A *subsystem* is a connected induced subgraph of the system topology G. The inputs of a subsystem S are the set of inputs of the components in S that are either system inputs or are outputs of components that are not in S. The subsystem outputs are defined in a similar way.

Definition 4 (Observed Subsystem)**.** *An observed subsystem is a subsystem in which all its inputs and at least one of its outputs are either system inputs or system outputs.*

Note that an input of an observed subsystem can be a system output. The key property of an observed subsystem is that all the inputs and at least one of its outputs must be *observable* (i.e., given in an observation).

Figure 1 shows several of the observed subsystems that exist in the depicted system. For example, the set of components $\{G, H, J\}$ forms an observed subsystem with inputs from D E and outputs from J.

Let $\mathcal{S}$ be the set of all observed subsystems of the diagnosed system. This set allows us to extend the model we can learn from normal observations described in Eq. 2, to the following model, denoted by $SD_{\mathcal{S}}$:

$$\bigwedge_{\substack{S \in \mathcal{S}, \\ obs_{OK} \in \text{OKs}}} \Big(\big(\bigwedge_{c \in S} h(c) \big) \rightarrow f\big(S, in(S, obs_{OK}), out(S, obs_{OK})\big) \Big) \tag{4}$$

where $in(S, obs_{OK})$ and $out(S, obs_{OK})$, are the input and output values of S, respectively, in obs_{OK}, and f is generalized in the natural way from components to subsystems.

One can adapt SD_{FP} from the full probing setting to partial probing, by including only the clauses in Eq. 2 that consider observed components. However, such a system model will be less powerful than $SD_{\mathcal{S}}$ since every observed component is, in fact, an observed subsystem where the subsystem contains a single component. Since $SD_{\mathcal{S}}$ still only describes the normal behavior of the components, it is still weaker than SD_{WFM}, in the sense that $SD_{\text{WFM}} \rightarrow SD_{\mathcal{S}}$.

The concepts of surely faulty components and normally behaving components can be adapted to the subsystem level: a surely faulty *subsystem* and a normally behaving *subsystem* are subsystems that have a surely normal input in obs_{ab} and whose outputs are different and the same, respectively.

Proposition 5. *Every diagnosis must contain at least one member from every surely faulty subsystem.*

Proof outline: A surely faulty subsystem is a conflict w.r.t. $SD_{\mathcal{S}}$, as it follows by induction: if all subsystems are healthy and all inputs are normal, then the output should also be normal; contrapositively, if the output is abnormal but the inputs are normal, then there must be a faulty component.

Minimal Observed Subsystem

The set of all observed subsystems ($\mathcal{S}$) is never empty, as every system output induces at least one observed subsystem. In fact, in a full probing setting $\mathcal{S}$ can even be exponential in the number of components. Thus, we focus on the set of *minimal* observed subsystems, where an observed subsystem S is *minimal* if no proper subset of S is also an observed subsystem. We denote this set of subsystems by $\mathcal{S}_{min}$. Every system output o has a corresponding minimal observed subsystem, defined by going backwards on the system topology until reaching either a system output or the system inputs. In fact, $\mathcal{S}_{min}$ is exactly the set of all these subsystems. Figure 1 shows, in dashed lines, all the subsystems in $\mathcal{S}_{min}$ for the example system. Note that $\mathcal{S}_{min}$ can be found in polynomial time as described above, and it contains at most $|\text{SysOuts}|$ subsystems. Next, we show that this set is sufficient for diagnostic purposes and there is no need to find all of the observed subsystems.

Let $SD_{\mathcal{S}_{min}}$ be the subset of the system model $SD_{\mathcal{S}}$ that uses only the minimal observed subset in $\mathcal{S}_{min}$.

Proposition 6. $SD_{\mathcal{S}_{min}} \equiv SD_{\mathcal{S}}$

Proof outline: Since $\mathcal{S}_{min} \subseteq \mathcal{S}$, then $SD_{\mathcal{S}} \models SD_{\mathcal{S}_{min}}$. To prove the other direction, we observe that a non-minimal observed subsystem S is equivalent to a union of the set of minimal observed subsystems that correspond to the outputs of S. Thus, the statement in $SD_{\mathcal{S}}$ added due to S can be derived from the set of statements added to $SD_{\mathcal{S}_{min}}$ due to S's constituent minimal observed subsystems.

The set of observable subsystems to which a component belongs provides additional, useful information:

Proposition 7. *If every observable subsystem that a component c belongs to is normally behaving, then any diagnosis that contains c is not minimal.*

The proof is omitted due to space constraints.

m\P	1		4		8		16		Full (57)
4	0.01	(6.9)	0.05	(2.3)	0.09	(2.1)	0.18	(1.7)	0.67 (0.0)
16	0.05	(9.6)	0.12	(5.2)	0.20	(4.4)	0.35	(3.1)	0.95 (0.0)
64	0.13	(9.5)	0.21	(6.6)	0.37	(5.6)	0.56	(4.3)	1.00 (0.0)
256	0.30	(12.4)	0.38	(9.6)	0.55	(7.5)	0.72	(5.7)	1.00 (0.0)
1024	0.52	(16.6)	0.57	(13.1)	0.70	(9.5)	0.83	(7.2)	1.00 (0.0)
2048	0.62	(18.9)	0.66	(15.3)	0.76	(10.7)	0.87	(7.8)	1.00 (0.0)
4096	0.75	(21.0)	0.78	(17.3)	0.86	(12.1)	0.93	(8.7)	1.00 (0.0)

Table 1: Experimental results on the 74181 system.

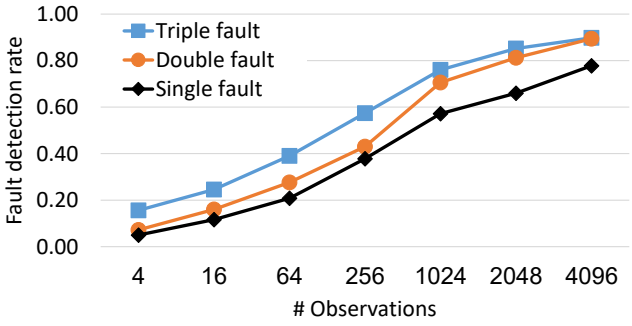

Figure 3: Fault detection rate for the 74181 system with 4 probes and 1, 2, and 3 injected faults.

Statistical Analysis

The system model $SD_{\mathcal{S}_{min}}$ allows us to extend Theorem 1 to consider the system topology, making it also applicable for any probing setting (full, partial, or none).

Theorem 2. *Let $V_{\mathcal{S}_{min}}$ denote the size of the largest domain for the input of a minimal observable subsystem. Given*

$$m \geq \frac{1}{\epsilon} \ln \frac{|\textsc{SysOuts}| \cdot V_{\mathcal{S}_{min}}}{\delta}$$

normal observations, with probability at least $1 - \delta$, the learned model $SD_{\mathcal{S}_{min}}$ is sufficient to identify at least one surely faulty minimal observed subsystem with probability at least $1 - \epsilon$.

The proof of this theorem follows the proof of Theorem 1 and Lemma 1 exactly, where the only change needed is that instead of the number of components n we have the number of system outputs $|\textsc{SysOuts}|$, and instead of the size of the largest domain for a component's input V we have $V_{\mathcal{S}_{min}}$.

Experimental Results

As a proof-of-concept, we implemented our diagnosis algorithm and ran it on system 74181 from the ISCAS '85 Boolean circuit benchmark (Brglez, Bryan, and Kozminski 1989). This system has 65 components, 14 inputs, and 8 outputs. We created training sets by setting uniform-random inputs to the system. In each trial, we injected faults randomly to the system, ran our diagnosis algorithm, and checked if the injected fault was identified as part of a surely faulty subsystem. We varied the size of the training set (4, 16, 64, 256, 1024, 2048, and 4096) and the number of probes (1, 4, 8, 16, and 57, where 57 means full probing for this system). Table 1 shows the proportion of trials out of 100 where the learned model was able to identify the failed component. We show the average number of other components that were in the same observable subsystem as the faulty component in parentheses. The size of the training set varies across rows, and the number of probes varies across columns. Our results are incomparable to that of MBD algorithms that know the system's model. However, as expected, more probes and larger training sets increase the likelihood that our algorithm will detect the fault. Although the system has 14 inputs (2^{14} possible input vectors), with full probing and only 256 training instances we always found exactly the faulty component. Even with only 8 probes, our algorithm found the faulty component in more than half (55%) of the cases. Due to the system topology, in 30% of the problems even one probe is sufficient to identify a set of fewer than 13 components (on average) that contain the faulty component. Observe that for the same number of probes, adding more observations results in identifying more surely faulty subsystems in more instances. Consequently, the number of components in observed subsystems identified as surely faulty also increases.

Figure 3 plots the ratio of cases where a surely faulty minimal observed subsystem was found (on the y axis) as a function of the training size (on the x-axis), for the 74181 system with four probes when injecting one, two, and three faults. As can be seen, our approach is not limited to observations with a single fault and works well for faults of higher cardinality. In fact, having more faults increases the chances that a surely faulty minimal observed subsystem will be found.

We also performed some preliminary experiments on the c3540 system, which is significantly larger than 74181 – it has over 1500 components, 50 system inputs and 22 system outputs. With 4,096 normal observations and only 4 probes our algorithm was able to identify a surely faulty component in over 27% of the cases.

Related Work

Niggemann et al. (2012) proposed a method for learning behavioral models of a system described by a hybrid timed automaton, but the learned model was used for fault detection and not for diagnosis (fault isolation). Feldman et al. (2015) proposed an algorithm for learning how to construct a hybrid model from a set of models of different fidelity. Preliminary work by Sadov et al. (2010) studied how to test the system in order to efficiently learn a model for diagnosis. We do not assume the ability to test the system, and only infer a partial model from a given set of observations. Also, we exploit the system topology, which they did not.

Approximate MBD has been studied in the Fault Detection and Isolation (FDI) community, based on Fuzzy models (Dexter and Benouarets 1997; Mendonça, Sousa, and da Costa 2009) and other ways to capture how the system description can be inaccurate (e.g., robust FDI) (Chen and Patton 2012; Frank and Ding 1997). We do not assume a-priori knowledge about the inaccuracy of the model, and, unlike these prior works, propose a framework that supports logic-based reasoning over the model. Juba (2016) proposed a method for PAC-learning of abductive reasoning that could in principle be used for diagnosis, extended to partial observations by Juba et al. (2018). But, this approach requires

that the training set contains both normal and abnormal observations. In contrast, we do not need to train on abnormal observations. The notion of a *surely faulty* subsystem is related to the discussion of *partial diagnoses* by Shchekotykin et al. (2016). A partial diagnosis is a diagnosis that is created by a hitting set of a subset of all minimal conflicts. Thus, a partial diagnosis may or may not be correct. By contrast, a surely faulty component must be faulty, but it may not be a diagnosis in the sense that it may not explain all of the abnormal behavior we currently observe. Deriving partial diagnoses from normal observations is a topic for future work.

Grouping components into a set of minimal observable subsystems bears some resemblance to the system abstractions (Stern, Kalech, and Elimelech 2014; Siddiqi and Huang 2011; Sachenbacher and Struss 2005; Torta and Torasso 2003). The set of minimal observed subsystems, however, does not create a disjoint partition over the components, i.e., there are components that are part of more than one minimal observed subsystem (e.g., B in Figure 1).

Casanova et al. (2014) considered how to bound the limits of diagnostic accuracy given partial probing. They assumed a distribution over possible inputs and showed how to infer from this knowledge which components can be isolated if faulty and which components are indistinguishable. This is somewhat similar to our notion of observable subsystem. They did not relate the number of available observations to diagnostic accuracy. Also, their work was focused on diagnosis engines based on spectrum based fault localization (SFL). Diagnosis based on SFL does not use a model of the system components' behavior, and considers instead only which components were used when the abnormal behavior was observed.

Conclusion and Discussion

We provide a theoretical foundation for diagnosis using only normal observations, system topology, and probes if available. We showed how to learn a partial system model from this information, and how to use it to for diagnosis. Our approach is unique in that the learned model is *safe*, in the sense that conflicts extracted from it are conflicts that would have been extracted if we knew the system's model. The number of observations needed to obtain an effective partial model is analyzed, and a small experimental proof of concept is reported, suggesting that our method is applicable.

The analysis in Theorems 1 and 2 only guarantees that we find one surely faulty observed subsystem, even in multiple fault scenarios. However, our approach can sometimes detect more than a single fault. This occurs when multiple observed subsystems are detected as surely faulty, and these observed subsystems have no overlap. Analyzing the number of observations and probing needed to identify a complete multiple fault diagnosis is a topic for future work. Another direction for future work is to incorporate other methods for detecting faulty outputs that can be used to identify more subsystems as surely faulty.

In this paper, we made several significant assumptions. We now discuss approaches for relaxing them.

Deterministic Normal Behavior. If a component may output different values in two different observations even if it received the same inputs and it is healthy, it has a *non-deterministic* normal behavior. To capture such a non-deterministic behavior formally, for every component c we define a relation R_c that is the set of all the input-output pairs that are normal. If we assume that the normal behavior of the system components is *stationary*, i.e., the choice of output for a given input follows some stationary distribution, then we can adapt our theoretical results by replacing the size of the input domain (V in Theorem 1) with the size of the relation R_c. Notably, we can never fully establish a component is surely faulty, but we can establish that it is faulty with high probability.

We stress that this approach cannot diagnose faults with respect to a *probabilistic* specification. As an extreme case, suppose that one of our components is supposed to output an unbiased random bit, but it outputs 0 with probability 1. It is faulty with respect to our specification, but we cannot diagnose this from a single observation of the system, even with full probing. Many observations of this faulty behavior are needed to have any hope of detecting it.

Faults due to Abnormal Inputs. Faults may occur due to abnormal or illegal system inputs, and not faulty components. This means the correct diagnosis may be that a system input was illegal. We can model this setting by considering every system input as a "buffer" component: when healthy, it outputs normal system inputs, and when faulty it does not. Thus, our analysis is fundamentally the same.

Cyclic System Topology. In some systems there is no clear notion of components' inputs and outputs, and components may affect each other in both directions. Thus, a proper analysis of the system's behavior requires a temporal aspect, to represent the propagation of values in the system (Feldman and de Kleer 2017). Learning a model from normal observations over time is an exciting direction for future work.

Components with a Continuous Domain. The analysis in Theorems 1 and 2 implicitly assumes that the components' input domain is discrete, as otherwise it has an infinite number of possible values. However, various methods have been proposed to learn functions over continuous-valued inputs (and with continuous-valued outputs). For example, strong generalization bounds have been proved for the Rademacher complexity (Bartlett and Mendelson 2002), and analyses of the Rademacher complexity have been given by Kakade et al. (2009) for the many common kinds of real-valued functions that can be viewed as linear functions. Thus, it may be possible to derive polynomial sample complexity bounds for our problem from these works given some a-priori knowledge about the range of normal behaviors ($\varphi_h(c)$), e.g., some parametric model for the healthy components.

Acknowledgments

Brendan Juba was supported by an AFOSR Young Investigator Award and NSF award CCF-1718380. Roni Stern was supported by the Cyber Security Research Center at Ben Gurion University of the Negev and ISF no. 210/17.

References

Abreu, R.; Zoeteweij, P.; Golsteijn, R.; and Van Gemund, A. J. 2009. A practical evaluation of spectrum-based fault localization. *Journal of Systems and Software* 82(11):1780–1792.

Abreu, R.; Zoeteweij, P.; and Van Gemund, A. J. 2009. Spectrum-based multiple fault localization. In *IEEE/ACM International Conference on Automated Software Engineering*, 88–99.

Bartlett, P. L., and Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3(Nov):463–482.

Brglez, F.; Bryan, D.; and Kozminski, K. 1989. Combinatorial profiles of sequential benchmark circuits. In *IEEE International Symposium on Circuits and Systems*, 1929–1934.

Casanova, P.; Garlan, D.; Schmerl, B.; and Abreu, R. 2014. Diagnosing unobserved components in self-adaptive systems. In *International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 75–84.

Chen, J., and Patton, R. J. 2012. *Robust model-based fault diagnosis for dynamic systems*, volume 3. Springer.

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

de Kleer, J., and Williams, B. C. 1987. Diagnosing multiple faults. *AIJ* 32(1):97–130.

de Kleer, J. 2009. Diagnosing multiple persistent and intermittent faults. In *IJCAI*, 733–738.

Dexter, A. L., and Benouarets, M. 1997. Model-based fault diagnosis using fuzzy matching. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 27(5):673–682.

Feldman, A., and de Kleer, J. 2017. Diagnosing multiple faults in sequential logic. In *DX*, 2–8.

Feldman, A.; de Castro, H. V.; van Gemund, A.; and Provan, G. 2013. Model-based diagnostic decision-support system for satellites. In *Aerospace Conference, IEEE*, 1–14.

Feldman, A.; Provan, G.; Abreu, R.; and de Kleer, J. 2015. Learning diagnosis models using variable-fidelity component model libraries. *International Federation of Automatic Control (IFAC)* 48(21):428 – 433.

Frank, P. M., and Ding, X. 1997. Survey of robust residual generation and evaluation methods in observer-based fault detection systems. *Journal of process control* 7(6):403–424.

Gupta, S.; Abreu, R.; de Kleer, J.; and van Gemund, A. J. 2014. Automatic systems diagnosis without behavioral models. In *IEEE Aerospace Conference*, 1–8.

Hafez, W.; Ross, T.; and Gadd, D. 1997. Machine learning for model-based diagnosis. In *American Control Conference*, volume 1, 42–46. IEEE.

Haussler, D. 1992. Overview of the probably approximately correct (PAC) learning framework. *Information and Computation* 100(1):78–150.

Juba, B.; Li, Z.; and Miller, E. 2018. Learning abduction under partial observability. In *AAAI*, 1888–1896.

Juba, B. 2016. Learning abductive reasoning using random examples. In *AAAI*, 999–1007.

Kakade, S. M.; Sridharan, K.; and Tewari, A. 2009. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, 793–800.

Keren, B.; Kalech, M.; and Rokach, L. 2011. Model-based diagnosis with multi-label classification. In *DX*.

Lamperti, G., and Zanella, M. 2012. Consistency-based diagnosis: a topological approach. In *DX*, 207–214.

Mendonça, L. F.; Sousa, J.; and da Costa, J. S. 2009. An architecture for fault detection and isolation based on fuzzy methods. *Expert systems with applications* 36(2):1092–1104.

Niggemann, O.; Stein, B.; Vodencarevic, A.; Maier, A.; and Büning, H. K. 2012. Learning behavior models for hybrid timed systems. In *AAAI*, volume 2, 1083–1090.

Qin, S. J. 2012. Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control* 36(2):220–234.

Reiter, R. 1987. A theory of diagnosis from first principles. *AIJ* 32(1):57–95.

Sachenbacher, M., and Struss, P. 2005. Task-dependent qualitative domain abstraction. *AIJ* 162(1-2):121–143.

Sadov, V.; Khalastchi, E.; Kalech, M.; and Kaminka, G. A. 2010. Towards partial (and useful) model identification for model-based diagnosis. In *DX*.

Shchekotykhin, K. M.; Schmitz, T.; and Jannach, D. 2016. Efficient sequential model-based fault-localization with partial diagnoses. In *IJCAI*, 1251–1257.

Siddiqi, S. A., and Huang, J. 2011. Sequential diagnosis by abstraction. *Journal of Artificial Intelligence Research* 41:329–365.

Stern, R.; Kalech, M.; and Elimelech, O. 2014. Hierarchical diagnosis in strong fault models. In *DX*.

Struss, P., and Price, C. 2003. Model-based systems in the automotive industry. *AI magazine* 24(4):17–34.

Torta, G., and Torasso, P. 2003. Automatic abstraction in component-based diagnosis driven by system observability. In *IJCAI*, 394–402.

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

Vapnik, V., and Chervonenkis, A. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications* 16(2):264–280.

Williams, B. C., and Nayak, P. P. 1996. A model-based approach to reactive self-configuring systems. In *AAAI*, 971–978.

Williams, B. C., and Ragno, R. J. 2007. Conflict-directed A* and its role in model-based embedded systems. *Discrete Applied Mathematics* 155(12):1562–1595.