

Computing Effect-to-Cause / Cause-to-Effect Diagnoses Within *NdL**

BÉCHIR EL AYEB AND SHENGRUI WANG

Université de Sherbrooke

Faculté des Sciences / DMI

Sherbrooke

CANADA J1K 2R1

Abstract

The computation of effect-to-cause (called also deductive) diagnoses is rather straightforward whatever the considered framework—logical, neural, etc. This paper focuses on the computation of cause-to-effect (called also abductive) diagnoses. Ultimately, we show how mechanize the abduction within a neural framework and illustrate our mechanization by illustrative examples. Finally, we present instructive lessons learnt from our experimentation with *NdL* (a Neural Diagnosis Laboratory) and point out the innovative features of this work as well as the important issues deserving further investigations.

1 Prologue

Many research activities are devoted to diagnosis where different methods have been explored. Among these methods, let us mention default trees [Breuer and Friedman, 1976], model-based diagnosis [de Kleer and Williams, 1987; Reiter, 1987; Raiman and de Kleer, 1992; Ayeb *et al.*, 1993], empirical-based diagnosis [Peng and Reggia, 1987; Ayeb, 1994] associative neural networks [Appolloni *et al.*, 1990; Jakubowicz and Ramannujam, 1990; Gaudill, 1990]. From an abstract point of view, diagnostic methods can be characterized by the *framework* and the *nature of knowledge* used. The framework delimits the adopted approach and the starting background such as logical, neural, data analysis, etc. The nature of knowledge specifies whether we make use of fault trees, statistical data, behavior models, empirical associations, etc.

In this paper we consider two kinds of framework, logical and neural, and make use of one nature of knowledge i.e., empirical associations. Therefore, like in [Peng and Reggia, 1987] we define a **diagnostic problem** (DP for short) as a three-tuple $DP = \langle \mathcal{C}, \mathcal{E}, \mathcal{R} \rangle$, where $\mathcal{C} = \{c_1, \dots, c_n\}$, is a finite set representing the possible causes; $\mathcal{E} = \{e_1, \dots, e_m\}$, is a finite set representing

the possible effects; $\mathcal{R} = \{(c_i, e_j) / c_i \in \mathcal{C} \wedge e_j \in \mathcal{E}\}$, is a subset of the Cartesian product $\mathcal{C} \times \mathcal{E}$ representing the relationships between causes and effects.

Let $DP = \langle \mathcal{C}, \mathcal{E}, \mathcal{R} \rangle$ be a given diagnostic problem, then an **observation** for DP is defined as a subset $E = \{e_{j_1}, \dots, e_{j_k}\}$ of $\mathcal{E}$ representing observed effects. Now considering a couple (DP, E) , we define the concept of a **solution** (or a **diagnosis**). Intuitively, a diagnosis is a subset of causes *explaining* the observed effects. The meaning of *explain* depends on our intuitive interpretation of the diagnostic knowledge embedded in $\mathcal{R}$ and its representation types. In fact, there are two ways to interpret the given diagnostic knowledge and ultimately two definitions of what is a diagnosis.

The first way consists in interpreting this knowledge embedded in $\mathcal{R}$ as a set of empirical associations relating effects to causes. Within a logical framework, $\mathcal{R}$ will be represented as a set of formulas K_d of the following form: $e_{i_1} \wedge \dots \wedge e_{i_k} \Rightarrow c_j$. Within a neural framework $\mathcal{R}$ will be mapped into a neural network having effect cells as its inputs and cause cells as its outputs. Here, we call a diagnosis for a given couple (DP, E) as being an **effect-to-cause diagnosis**. Clearly, the computation of effect-to-cause diagnoses is straightforward both in neural and logical frameworks. Typically, within a logical framework an effect-to-cause diagnosis Δ_d can be obtained by direct deduction from the set $K_d \cup E$ —i.e., Δ_d is a subset of $\mathcal{C}$ such that $K_d \cup E \models \Delta_d$. Within a neural framework, the diagnostic problems are further reduced to mere classification problems. Hence, a Perceptron like neural networks can be very effective to generate the effect-to-cause diagnoses. That is why most of diagnostic neural proposals [Appolloni *et al.*, 1990; Gaudill, 1990; Jakubowicz and Ramannujam, 1990], fall into this category.

The second way consists in interpreting this knowledge embedded in $\mathcal{R}$ as a set of empirical associations relating causes to effects. In this case, we call a diagnosis for a given couple (DP, E) as being a **cause-to-effect diagnosis**. Specifically, within a logical framework, $\mathcal{R}$ will be represented as a set of formulas K_a of the following form: $c_i \Rightarrow e_{i_1} \wedge \dots \wedge e_{i_k}$. A cause-to-effect diagnosis for (DP, E) is a subset Δ_a such that $\Delta_a \cup K_a \models E$, and $\Delta_a \subset \mathcal{C}$ is minimal w.r.t set inclusion or cardinality set. Clearly many difficulties arise both in neural and logical frameworks. In particular, a more complex reasoning

**NdL* is an acronym of *Neural diagnosis Laboratory* which is a component of **dL*. This research is partially supported under FRAI Grants (0238A; 0291A) and NSERC grants (OGP0121468; EQP0121469).

process is required to generate cause-to-effect diagnoses. Typically, within a logical framework, we must require an abductive (called also hypothetical) reasoning since mere deduction is not sufficient. Similarly, within a neural framework, cause-to-effect diagnoses cannot be obtained by a simple association mechanism since diagnostic problems cannot be reduced to a mere classification. That is why very few proposals have been suggested to address cause-to-effect diagnoses computation.

In fact, the intuition of effect-to-cause and cause-to-effect diagnoses can be outlined as follows. The set of causes in an effect-to-cause diagnosis can be viewed as an *evidence explaining* the presence of the observed effects, while the set of causes in a cause-to-effect diagnosis can be viewed as a (minimal) *conjecture explaining* the presence of the observed effects. Of course, the interpretation of cause-to-effect diagnoses is more natural since the effect-to-cause diagnoses tend to loose the "structure" of knowledge [Poole, 1988]. Despite the different meanings conveyed by effect-to-cause and cause-to-effect diagnoses, both have been used in many diagnostic applications. Within a logical framework, let us mention the diagnostic expert system MYCIN (effect-to-cause diagnoses) and the abduction based diagnoses [Cox and Pietrzykowski, 1987; Ayeb *et al.*, 1993] (cause-to-effect diagnoses). Within a neural framework, most of proposals such as [Appolloni *et al.*, 1990; Jakubowicz and Ramannujam, 1990; Jakubowicz and Ramannujam, 1990] generate effect-to-cause diagnoses. Nowadays there are few proposals, namely [Peng. and Reggia, 1989; Goel *et al.*, 1988], aiming at the generation of cause-to-effect diagnoses within a connectionist/neural framework. There is no accident since the mechanization of abduction raises many difficulties and challenges regardless the adopted framework, neural or logical.

The purpose of this paper is three-fold. First, bridge the gap between the neural framework and the conventional one. Second, propose an effective neural model mechanizing the abductive reasoning. Third, present instructive lessons learnt from our experimentation with *NdL*¹. The rest of this paper is organized as follows. Section 2 points out the background and principles of this work. Section 3 presents our neural model to mechanize abductive reasoning. Section 4 shows the computation of abductive diagnoses of some illustrative examples within *NdL*. Section 5 briefly analyses our model and points out the important issues deserving further investigation.

2 *NdL*: A Brief Background

According to its definition, a cause-to-effect diagnosis (called also **hypothetical** or **abductive** [Cox and Pietrzykowski, 1987; Ayeb *et al.*, 1993] diagnosis) is a conjecture of causes explaining all the observed effects. Clearly, a pair (*DP*, *E*) may admit several abductive diagnoses, hereafter called *diagnostic candidates*. An im-

portant issue is then to define some criteria pointing out the best diagnosis among the diagnostic candidates. Two criteria are usually adopted in literature: *parsimony covering* and *maximal plausibility*. Parsimony covering states that the best diagnosis must include a minimal set of causes. Such a minimality can be considered w.r.t set inclusion or cardinality set. Maximal plausibility states that the best diagnosis must have the most likelihood w.r.t a given belief function. Depending on each case, the belief function can be attached individually to causes or to causal relationships. In fact, there is a potential conflict between these criteria since the parsimony covering (resp. maximal plausibility) represents a global (resp. local) optimization. Fortunately, such a conflict can be resolved by adopting a precedence relation between them. Unfortunately, the soundness of these criteria is case-dependent and is based on several, however conventional, assumptions—see [Console *et al.*, 1989; Goel *et al.*, 1988; Peng. and Reggia, 1987] for more details. In this work, we adopt these conventional assumptions and characterize the best diagnosis in terms of minimal cardinality and maximal plausibility.

In order to model a diagnostic process generating such best diagnoses, the neural network should support complex patterns of interaction between effects and causes, among causes and possibly among effects. This process should allow those cells representing most plausible causes to be activated and those which are not to be inhibited. The role of competition has long been endorsed [Peng. and Reggia, 1989]. The main issue is how to model it. A simple form of competition is to sum up inputs to each cell and take the most "excited" cell or cells as the results of the competition without any transient adaptation. The precision of this form depends heavily on the connection weights (prewired knowledge) and does not allow the network to sufficiently "relax". Essential causes may be kicked out if their initial activation strengths are not high. Equilibrium based competition provides better chance for network relaxation. However, as being pointed out in [Peng. and Reggia, 1989], it is very difficult, if not impossible, to model, through simple inhibitory connections with static weights, complex relationships such as competition between two causes when a common effect shows up and cooperation between them otherwise. It is also very difficult to model the "essentialness and/or plausibility" of a cause. Thus, the network should involve a dynamical activation mechanism that takes into account varieties of factors to model a desired phenomena of winners-take-others.

Our proposal is based on fundamental neural activation principles, discussed in [Grossberg, 1988], and a generalized synaptic adaptation rule of Hebb. Cells in our neural model are based on a shunting equation. Interactions among cause cells and between cause cells and effect cells are emulated. All the cause cells which have one of the observed effects are activated at the beginning by the active effect cells. Activities of cause cells reflect the network's belief on the plausibilities of the causes. Temporal adaptation and normalization of the weights between cause and effect cells are modeled to reflect the changes in the distribution of cells' activities,

¹ *NdL* (Neural Diagnosis Laboratory) is a component of **dL* which is written in C++ with an interface to Rochester system tool [N. H. Goddard and K. J. Lynne and T. Mintz, 1988]. Since the computation of effect-to-cause diagnoses is straightforward, in this paper *NdL* focuses only on the computation of cause-to-effect or abductive diagnoses.

while adaptation of the weights between cause cells regulates inhibition strengths. Both together act in favor of those cells having the strongest support from observed effects or being indispensable to explain certain effects and allow to eliminate those causes that are not essential.

In the proposed neural architecture, cells in the network are grouped into two main layers $\mathcal{E}$ representing all the possible effects, and $\mathcal{C}$ representing all the possible causes. Ultimately, we use the following conventions: the activity value x_i will denote the degree to which a cause or an effect holds. We shall use w_{ij} to represent the weight of the (excitatory) connection between effect cell e_j and cause cell c_i —i.e., w_{ij} stand for the fuzzy or uncertain (including probabilistic) relationship between e_j / c_i . Cause cells c_i and c_k will be connected if they have at least one common effect—i.e. iff there exists j such that $w_{ij} w_{kj} \neq 0$. The inhibitory strength between cause c_i and c_k explaining their common effects will be quantified by I_{ik} .

3 NdL: The Underlying Mechanics

The activity x_j of a given effect cell e_j is set to a given constant ψ_j if the effect e_j is in the given observation E —Otherwise x_j is set to zero. Moreover, their activities will stay unchanged through out the reasoning process. In its turn x_i the activity of the cell c_i is governed by the shunting competitive equation [Grossberg, 1988]: $\frac{dx_i}{dt} = -Ax_i - x_i \sum_k I_{ik} f(x_k) + \sum_j w_{ij} x_j$, where A is a small positive real number (i.e., 0.1) representing the decay rate of the cell c_i . Let us rewrite the above equation as follows: $\frac{dx_i}{dt} = -Ax_i - x_i IH_i + EX_i$, where $IH_i = \sum_k I_{ik} f(x_k)$ and $EX_i = \sum_j w_{ij} x_j$.

In fact, IH_i is the total inhibitory input to the cell c_i from all the other cells in the layer $\mathcal{C}$. Its action on the activity of the cell c_i is modulated by the activation level of the cell c_i itself. This guarantees smooth evolution of cell activity and long transient period during which the network is devoted to the reasoning process. Conversely, EX_i is the total excitatory input to the cell c_i from layer $\mathcal{E}$. Observe that EX_i is not modulated by x_i since EX_i is a supporting evidence of the cause c_i . Note also that in absence of the excitatory inputs (i.e., $EX_i = 0$), the activity x_i will converge to zero due to the decay parameter A .

The reasoning process is modeled by iterative updating of cells' activity. It can be summarized in two major stages. Steps 1 and 2 form the first stage for initializing all possible causes. Steps 3 to 6 form the second stage engaging the network into a competitive reasoning process. Here the different steps:

•Step 1.0• $x_j \leftarrow \psi_j, \forall e_j \in E$

This first step consists of initializing the activity of effect cells e_j by the input data. We should suppose that values ψ_j be positive and normalized (i.e. bounded above).

•Step 2.0• $x_i \leftarrow \frac{EX_i}{A}, \forall x_i$

This second step gives a formula to initialize cause cells c_i . It is obtained assuming that when the effects turn on, the network converges immediately to a stable

state before the competition process begins. This means $\frac{dx_i}{dt} = 0$ while $IH_i = 0$ for all i .

•Step 3.0• $x_i \leftarrow \frac{EX_i}{A+IH_i}, \forall x_i$

This third step consists of resolving the system of non-linear differential equations in its stable state ($\frac{dx_i}{dt} = 0$) assuming that the states of the cells vary smoothly from one iteration to another and the output of each cell $f(x_k)$ is approximately constant within a short time interval. An important property is that x_i increases with EX_i (depending on the connection of all weights w_{ij}) and decreases with IH_i (depending on a great extent on the connection weights I_{ik}). Further more, when $EX_i > 0$, we have $x_i > 0$.

•Step 4.1• $w_{ij} \leftarrow w_{ij} + \alpha x_j x_i, \forall w_{ij} > 0$

•Step 4.2• $w_{ij} \leftarrow \frac{w_{ij}}{\|w_{\cdot,j}\|}, \forall w_{ij} > 0;$

•Step 4.3• $I_{ik} \leftarrow I_{ik} - \beta \delta_{x_i}^2 \delta_{x_k}^2, \forall I_{ik}/w_{ij_0} w_{kj_0} \neq 0$

This fourth step updates the connection weights to reflect the current variations of the network's belief on the plausibilities of the causes. The formula $w_{ij} \leftarrow w_{ij} + \alpha x_j x_i$ together with the normalization $w_{ij} \leftarrow w_{ij}/\|w_{\cdot,j}\|$ consists of a redistribution of belief strengths w_{ij} that e_j is caused by c_i . Clearly, causes gaining greater belief are normally those that resist better the competition. On the other hand, the normalization formula ensures that an exclusive cause is always gaining the maximal belief, which is 1, no matter how strong the other competitive cells are.

The formula $I_{ik} \leftarrow I_{ik} - \beta \delta_{x_i}^2 \delta_{x_k}^2$ aims at alternating inhibitory connections. Keeping in mind that a connection I_{ik} exists if and only if both c_i and c_k have a common effect. The formula is applied only to a I_{ik} connecting two causes that have a common effect appeared in the observation. If the activity x_i and x_k of such two causes vary in the opposite direction, then one of the causes seems to get more evidence and the inhibition weight is increased to locally accommodate the changes. On the other hand, if x_i and x_k both decrease or increase, which is likely to happen at the beginning of the reasoning process, the network is still in the search of a right direction—especially when they both increase. This step is crucial since it implements a form of competition between cells without using the simplist winner-take-all approach. It allows to eliminate causes which don't receive enough support and which are not essential. In our experiments, α and β are chosen to 0.1 and stay unchanged during the reasoning process.

•Step 5.0• $w_{ij} \leftarrow 0, \forall w_{ij} < \epsilon,$

This fifth step is simply for pruning and eliminates those causal relations no more significant for the subsequent reasoning.

•Step 6.1• when ($x_i > 0 \wedge exclusive(c_i), \forall c_i$) return ($\{c_i/x_i > 0\}$)

•Step 6.2• go to •Step 3.0•

²Note that δ_{x_i} (resp. δ_{x_k}) is the variation of x_i (resp. x_k) between two succeeding iterations.

The sixth step includes the termination test. The reasoning process is terminated when each of the remaining causes is exclusive. It is worth noting that a cause c_i is temporarily exclusive if there is an effect e_j such that c_i is the only remained cause that has e_j as an effect. Exclusiveness is thus related to the temporal belief of the network. The goal of the network updating is to eliminate those causes which are not considered to be essential and make every remaining cause exclusive w.r.t. at least one observed effect. Formally, $exclusive(c_i)$ holds iff $\exists e_{j_0} \in E / w_{ij_0} > 0 \wedge w_{kj_0} = 0, \forall k \neq i$.

4 NdL: A Few Experiments

For the sake of illustration, let us consider the following mechanical diagnostic problem adapted from [Console et al., 1989] where we suppose that the mechanical knowledge is expressed as a set of relationships between causes and effects instead of a causal network as in [Console et al., 1989].

K^3	oil_engine_cause	MustCause	ab_smoke_engine
	oil_engine_cause	MustCause	ab_acceleration_engine
	oil_engine_cause	CanCause	ab_vapour_engine
	oil_engine_cause	MustCause	ab_high_temperature
	oil_engine_cause	CanCause	ab_dirty_oiled_engine
	spark_plugs_cause	CanCause	ab_smoke_engine
	spark_plugs_cause	MustCause	ab_starting_engine
	spark_plugs_cause	MustCause	ab_acceleration_engine
	spark_plugs_cause	CanCause	ab_ignition_engine
	spark_plugs_cause	CanCause	ab_noise_engine
	carburettor_cause	CanCause	ab_gasoline_smell
	carburettor_cause	MustCause	ab_noise_engine
	carburettor_cause	MustCause	ab_acceleration_engine
	H2O_level_cause	MustCause	ab_lights_status
	H2O_level_cause	MustCause	ab_weak_battery
	H2O_level_cause	MustCause	ab_alternator
	H2O_level_cause	CanCause	ab_ignition_engine

To build the neural network (see Fig. 4.1 (b)) for our illustrative example, NdL requires a formal specification given in Fig. 4.1 (a)– note that a complete specification within NdL consists of six sections devoted to the syntactic level (e.g., network topology) as well as the semantic level (e.g., relationships interpretation, incompatibility of pairs of causes). However, only a snapshot of the specification is given to save space in this paper.

³We have distinguished two types of relationships (i.e., *MustCause* and *CanCause*) to qualify the causal relationships. For instance, the 1st. (resp. 3rd.) sentence of K tells us that an abnormal smoke (resp. abnormal vapour) from engine must (resp. should) be observed if the initial cause consists in an oil engine. Clearly, this interpretation is for the sake of illustration since one could introduce several qualifications (e.g., *Leads_To* or *MayManifest*) expressing a variety of relationships. Such a discussion on the modeling issue is left for a subsequent long paper on NdL/*dL.

BINDING SECTION	LINKING SECTION
<u>bind</u> oil_engine C ₁ ;;	<u>link</u> C ₁ E ₁ ;;
...	...
<u>bind</u> carburettor C ₄ ;;	<u>link</u> C ₄ E ₈ ;;
...	...
<u>bind</u> ab_smoke_engine E ₁ ;;	...
<u>bind</u> ab_ignition_engine E ₁₂ ;;	CAUSALITY SECTION
...	...

Fig. 4.1 (a): A Partial Network Specification Within NdL.

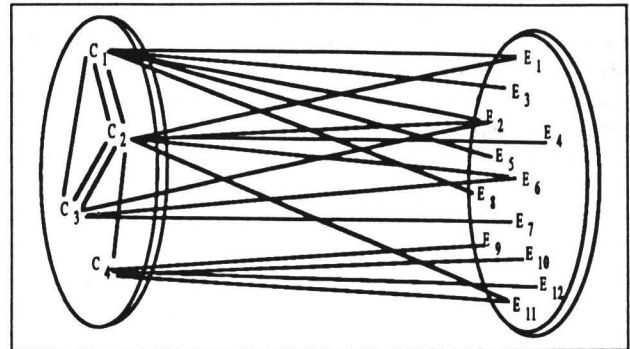


Fig. 4.1 (b): The Build Network.

From a theoretical viewpoint, excitatory and inhibitory connection weights (i.e., w_{ij} and I_{ik}) can be initialized to any values. However, NdL does not initialize these weights neither randomly nor in a brute force way but by first asking the user to select a initialization directive. NdL provides a variety of directives which help the network to quickly converge. Succinctly, a directive is a mapping that quantifies the causal relation between causes and effects (excitatory connections) and those between causes (inhibitory connections). For our illustrative example, we select the following simple directive assigning 1 (resp. $\frac{1}{2}$) for an excitatory connection w_{ij} between a cause c_i and an effect e_j depending on whether e_j must (resp. might) be observed. The rationale behind this directive stems from the following fact: the excitatory connections must reflect our belief of the causal relation between causes and effects.

Considering the inhibitory connections I_{ik} between two causes (say c_i and c_k), one common directive suggested by NdL is to “normalize” the number of common effects between both causes. This directive is motivated by the following observation: since the connection weight between two causes reflect mutual inhibition strength it should be proportional to the number of the common effects. Nonetheless, NdL normalizes these numbers to yield small connection weights– otherwise the effect of inhibition will be so strong that the cause will not be sufficiently activated to sustain the competition process. Of course, if two causes do not share common effects then I_{ik} does not exist– see Fig. 4.1b. Consider our illustrative example, $I_{oil_engine_cause, H2O_level_cause}$ and $I_{carburettor_cause, H2O_level_cause}$ do not exist, whereas the other inhibitory connections are: $I_{oil_engine_cause, spark_plugs_cause} = I_{spark_plugs_cause, oil_engine_cause} = 2/5 = 0.4$; $I_{oil_engine_cause, carburettor_cause} = I_{carburettor_cause, oil_engine_cause} = 1/5 = 0.2$;

$I_{spark_plugs_cause, carburettor_cause} =$
 $I_{carburettor_cause, spark_plugs_cause} = 2/5 = 0.4;$
 $I_{spark_plugs_cause, H2O_level_cause} =$
 $I_{H2O_level_cause, spark_plugs_cause} = 1/5 = 0.2;.$

Assume that we observe an abnormal status of lights and an abnormal acceleration, and notice a likely abnormal noise of engine. By starting *NdL*, a short dialog⁴ leads to the specification of $E_0 = \{ab_lights_status, ab_acceleration_engine, ab_noise_engine\}$ as the set of observed effects and $\Psi_0 = \{\psi_{ab_lights_status} = \psi_{ab_acceleration_engine} = 1, \psi_{ab_noise_engine} = 0.5\}$ as the initial activity of the cells corresponding to observed effects—step 1 of the reasoning.

Now, *NdL* enters the network in the competition process summarized in Fig. 4.2.

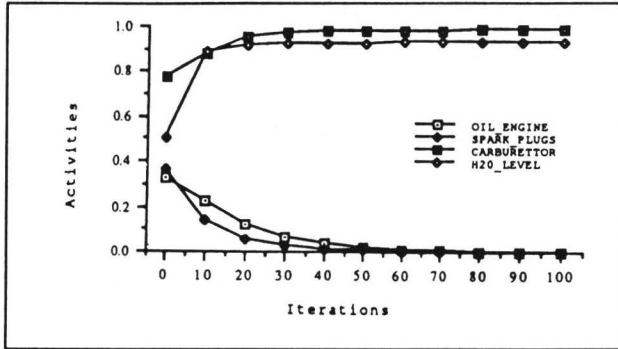


Fig. 4.2 (a): Activities Evolution.

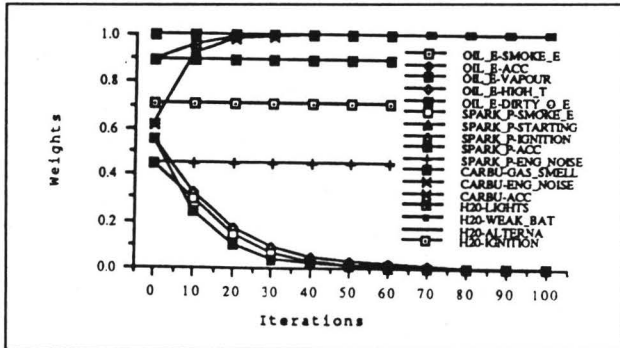


Fig. 4.2 (b): Weights Evolution.

The initial activities of cause cells are 0.79 for carburettor, 0.52 for H2O.level, 0.37 for spark-plugs and 0.32 for oil.engine. Due to the mutual inhibition, oil.engine and spark-plugs causes are quickly eliminated as depicted in the evolution of the activities (Fig. 4.2 (a)) and the evolution of the excitation weights (Fig. 4.2 (b)). Conversely, carburettor and H2O.level causes are the winners of the competition process. From the Fig. 4.2 (b), one can see that for a surviving cause there is one weight that gradually converges to 1. This means that cause becomes exclusive—here carburettor is exclusive to gasoline.smell. At the end of this competition, *NdL* collects the winners causes $\{carburettor, H2O.level\}$, that is the final diagnosis.

Now let us consider two additional observations to point out an interesting behavior of *NdL*. Let E_1 and E_2

be two observations together with their initial Ψ_1 and Ψ_2 specified as follows: $E_1 = E_2 = \{ab_acceleration_engine, ab_ignition_engine, ab_noise_engine, ab_gasoline_smell\}$, whereas $\Psi_1 = \{\psi_{ab_ignition_engine} = 0.5, \psi_{ab_acceleration_engine} = \psi_{ab_noise_engine} = \psi_{ab_gasoline_smell} = 1\}$, and $\Psi_2 = \{\psi_{ab_ignition_engine} = 1, \psi_{ab_acceleration_engine} = \psi_{ab_noise_engine} = \psi_{ab_gasoline_smell} = 0.5\}$. As before, *NdL* engages the network in a competition process summarized in Fig. 4.3.

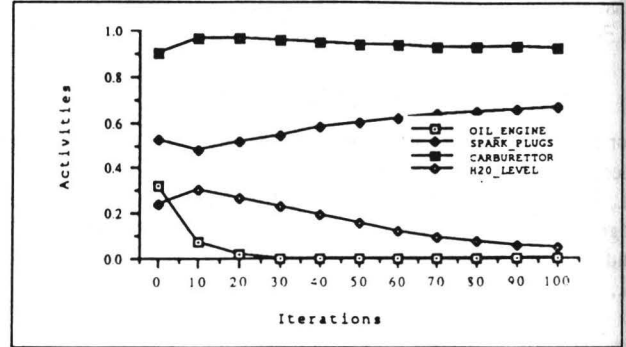


Fig. 4.3 (a): Activities Evolution w.r.t (E_1, Ψ_1).

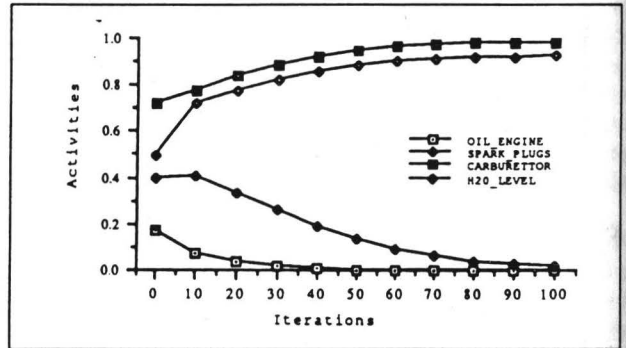


Fig. 4.3 (b): Activities Evolution w.r.t (E_2, Ψ_2).

This time *NdL* proposes $\{carburettor, spark_plugs\}$ (resp. $\{carburettor, H2O.level\}$) as the final diagnosis for E_1 (resp. E_2). From the formal viewpoint, both diagnoses are pretty good. However if we take the interpretation viewpoint we would prefer that both observations yield the same diagnosis $\{carburettor, spark_plugs\}$. The reason leading *NdL* to propose different diagnoses stems from the following: our exclusiveness criterion does not take the initial activities of effect cells (i.e., Ψ) into account. A slight adaptation of our neural model is under development.

5 Epilogue

Many challenges are raised by the mechanization of abductive reasoning regardless of the considered framework, conventional or neural. For example, the computational challenges within a conventional framework have been pointed out by many authors [Baylander et al., 1991; Goel et al., 1988; Console et al., 1989; Ayeb et al., 1993; Ayeb, 1992]. Considering a neural framework such a mechanization is difficult because it involves multiple and bidirectional inferences. Several causes may have a common observed effect, which one of them is the actual cause? Neither Perceptron-like associative neural networks nor more flexible models like ART [Carpenter

⁴Unfortunately, we cannot report here sample snapshots of dialogs which are space consuming.

and Grossberg, 1987], bidirectional associative networks [Kosko, 1988] can easily be used to produce the desired winners-take-others phenomena.

A common approach to model symbolic processing (e.g., abduction-based diagnosis) by neural networks, consists in mapping the original problem into the searching for an ideal stable state of a dynamic neural system. This stable state usually correspond to the minimum energies of the system. This has been the basic principle used in [Hopfield and Tank, 1985] to resolve the Traveler Salesman Problem. The proposals suggested by Peng and Reggia [Peng. and Reggia, 1989] and by Goel et al. in [Goel et al., 1988] are also based on this principle. Our proposal can also be analyzed in terms of minimization of an energy function. In fact, the main advantage of using an energy function is to model straightforwardly the process aiming at the construction of the best diagnoses w.r.t minimal cardinality and maximal plausibility. Within *NdLa* diagnosis is generated from the diagnostic candidates through competition and cooperation between causes. The plausibility of the best diagnosis results from the competition that strongly favors those causes either with a high strength or exclusive. Minimal cardinality also results from this competition process in that as long as there is a non-exclusive cause the competition will continue.

Exponential explosion is often the main issue of abductive reasoning to find the best diagnosis w.r.t the adopted criteria. In fact, there are two methods to construct a neural (connectionist) system mechanizing the abductive reasoning. The conventional method consists of two steps: (1) propose a target/energy function modeling the adopted criteria; (2) design the architecture of the network to optimize (maximize or minimize) the proposed function. Both proposals discussed earlier have adopted such a conventional method. The second less conventional method also consists of two steps: (1) design the architecture of the network w.r.t the diagnostic problem at hand; (2) analyze the network dynamics and the properties of the generated diagnosis. Our proposal adopts the second method. The main advantage of the conventional method is to make available the target/energy function in a earlier stage. Unfortunately, such an earlier availability imposes several constraints on the network design and optimization mechanisms to be used. On the other hand, the second method offers the advantage of imposing serious constraints on the network architecture, allowing to take into account the characteristics of the diagnostic problems. Nevertheless, such an advantage depends on the availability of means (e.g. an energy function) to analyze the network dynamics. That is why we suggest the following energy function for the proposed network. Consider the following function: $\alpha \sum_i x_i^2 + \beta \sum_j (\sum_i w_{ij} - \sqrt{\sum_i w_{ij}^2})^2$, where α and β are two given constants. Clearly, if the weights issued from each effect are normalized with the Euclidean norm (i.e., $\sum_i w_{ij}^2 = 1$), then we can rewrite the above function as: $\alpha \sum_i x_i^2 + \beta \sum_j (\sum_i w_{ij} - 1)^2$. Observe that both terms in the energy function are related to cardinality criterion since each active cell of a diagnostic candidate con-

tributes to each of them. The second term corresponds to the concept of exclusiveness since $(\sum_i w_{ij} - 1)^2 = 0$ iff $\exists i_0$ such that $w_{i_0j} = 1 \wedge w_{ij} = 0, \forall i \neq i_0$. Consequently, the exclusivity can be guaranteed provided that β is sufficiently large w.r.t α . Yet, one could argue that the energy function does not explicitly take into account the plausibility criterion. However, this criterion is a built-in part in the competition process. Indeed, the second phase of the proposed procedure ensures that only those cells with high sustained inputs, namely highly plausible causes, or cells which are/become exclusive, namely the only plausible causes for some effects, can survive the competition.

Before closing this paper, let us point out the distinctive features of this work w.r.t related work⁵. The first distinctive feature of our proposal w.r.t those of [Peng. and Reggia, 1989] and [Goel et al., 1988] stems from the additive updating rule while still having an energy function taking minimal cardinality and maximal plausibility into account. This results in an effective neural network architecture. Note that the proposals of [Peng. and Reggia, 1989] and [Goel et al., 1988] are rather connectionist network architecture (or high order neural network architecture). In fact, this results from their multiplicative function to take minimal cardinality and maximal plausibility into account. The second feature of our proposal consists of the temporal adaptation of connection weights. The connection weights are adapted to reflect the network's temporary belief on the plausibility of causes. Thus, the competition between causes is modeled through two mechanisms, updating activities of units and updating weights. Clearly, this adaptation speeds up the convergence to generate the appropriate diagnosis. It also determines the termination condition without using arbitrary thresholds.

This work can be extended in several directions related to methodological, theoretical and developmental aspects. For example considering the methodological aspect, it seems important to study the different classes of diagnostic problems and define the appropriate criteria characterizing the best diagnosis of each class. In the same vein, addressing the mechanization issue of diagnosis is worthwhile. Of course, our goal does not aim at decreasing the exponential complexity of the diagnosis procedure in general. Rather, we should design an efficient and adequate mechanization taking into account the requirements of a given class of diagnostic problems. Finally, let us mention another direction aiming at a distributed diagnostic problem solving where a complex problem is broken down into small pieces that can be solved either by a conventional diagnosis or by a neural one [Wang and Ayeb, 1992]. Such a direction is the long-term goal of the ongoing project **dL* where we combine in a skillful way both conventional and neural diagnosis.

Acknowledgment We are grateful to E. Jobidon who implemented the graphic interface of *NdL*.

⁵An experimental analysis w.r.t [Wald. et al., 1989; Tagamets and Reggia, 1989], and an extension of *NdL* to include non-monotonic abduction features [Josephson et al., 1987; Thagard, 1989; Baylander et al., 1991] are under investigation.

References

- [Appolloni *et al.*, 1990] B. Appolloni, G. Avanzini, N. Cesa-Bianchi, and G. Ronchini. Diagnosis of Epilepsy via Backpropagation. In *Proceedings of the International Joint Conference on Neural Networks*, pages II-539-II542, Washington, DC (USA), 1990.
- [Ayeb *et al.*, 1993] B. e. Ayeb, P. Marquis, and M. Rusinowitch. Preferring Diagnoses By Abduction. *IEEE, Transactions on Transactions on Systems, Man and Cybernetics*, 23, May 1993.
- [Ayeb, 1992] B. e. Ayeb. Diagnosis of Complex Industrial Devices. In *ACM Symposium on Applied Computing*, pages 512-520, Kansas (USA), 1992.
- [Ayeb, 1994] B. e. Ayeb. Towards Systematic Construction of Diagnostic Systems for Large Industrial Plants—Two parts. To appear in *IEEE Transactions on KDE*, 1994.
- [Baylander *et al.*, 1991] T. Baylander, D. Allemang, M. C. Tanner, and J. R. Josephon. The Computational Complexity of Abduction. *Artificial Intelligence*, 49:25-60, 1991.
- [Breuer and Friedman, 1976] M. A. Breuer and A. Friedman. *Diagnosis and Reliable Design of Digital Systems*. Computer Science Press, Rockville, MD, 1976.
- [Carpenter and Grossberg, 1987] G. A. Carpenter and S. Grossberg. A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. *Computer Vision, Graphics, and Image Processing*, 37:54-115, 1987.
- [Console *et al.*, 1989] L. Console, D. T. Dupré, and P. Torasso. A Theory of Diagnosis for Incomplete Causal Models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1311-1317, Detroit, Michigan (USA), 1989.
- [Cox and Pietrzykowski, 1987] P. T. Cox and T. Pietrzykowski. General Diagnosis by Abductive Inference. In *Symposium on Logic Programming*, pages 183-189, 1987.
- [de Kleer and Williams, 1987] J. de Kleer and B. C. Williams. Diagnosing Multiple Faults. *Artificial Intelligence*, 32:97-130, 1987.
- [Gaudill, 1990] M. Gaudill. Using Neural Nets: Diagnostic Expert Nets. *AI Expert*, September:43-47, 1990.
- [Goel *et al.*, 1988] A. Goel, J. Ramanujam, and P. Sadayappan. Towards a "neural" Architecture of Abductive Reasoning. In *Proceeding of Second International Conference on Neural Networks*, pages I-681-I-688, 1988.
- [Grossberg, 1988] S. Grossberg. Nonlinear Neural Networks: Principles, Mechanisms, and Architectures. *Neural Networks*, 1(1):17-60, 1988.
- [Hopfield and Tank, 1985] J. J. Hopfield and D. W. Tank. Neural Computation of Decisions in Optimization Problems. *Biological Cybernetics*, 52:141-152, 1985.
- [Jakubowicz and Ramannujam, 1990] O. Jakubowicz and S. Ramannujam. A Neural Network Model for Fault-Diagnosis of Digital Circuits. In *Proceedings of the International Joint Conference on Neural Networks*, pages II-611-II615, Washington, DC (USA), 1990.
- [Josephson *et al.*, 1987] J. R. Josephson, B. Chandrasekaran, J. W. Smith, and M. C. Tanner. A Mechanism for Forming Composite Explanatory Hypotheses. *IEEE Transactions on Transactions on Systems, Man and Cybernetics*, 17:445-454, 1987.
- [Kosko, 1988] B. Kosko. Bidirectional Associative Memories. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-18:49-60, 1988.
- [N. H. Goddard and K. J. Lynne and T. Mintz, 1988] N. H. Goddard and K. J. Lynne and T. Mintz. Rochester Connectionist Simulator. Technical Report TR-233, University of Rochester, New York, 1988.
- [Peng. and Reggia, 1987] Y. Peng. and J. A. Reggia. A Probabilistic Causal Model for Diagnostic Problem Solving (Two Parts). *IEEE Transactions on Transactions on Systems, Man and Cybernetics*, 17:146-162/395-406, 1987.
- [Peng. and Reggia, 1989] Y. Peng. and J. A. Reggia. A Connectionist Model for Diagnostic Problem Solving. *IEEE Transactions on Systems, Man and Cybernetics*, 19:285-289, 1989.
- [Poole, 1988] D. Poole. Representing Knowledge for Logic-Based Diagnosis. In ICOT, editor, *International Conference on Fifth Generation Computer Systems*, pages 1282-1290, 1988.
- [Raiman and de Kleer, 1992] O. Raiman and J. de Kleer. Critical Reasoning. In *Proceedings of the International Workshop On Diagnosis DX 92*, pages 1-5, Orcas, WA (USA), 1992.
- [Reiter, 1987] R. Reiter. A Theory of Diagnosis from First Principles. *Artificial Intelligence*, 32:57-95, 1987.
- [Tagamets and Reggia, 1989] M. A. Tagamets and J. A. Reggia. A data flow implementation of a competition-based connectionist model. *Journal of Parallel and Distributed Computing*, 6:704-714, 1989.
- [Thagard, 1989] P. Thagard. Explanatory coherence. *Behavioral and Brain Sciences*, 12:435-502, 1989.
- [Wald. *et al.*, 1989] J. Wald., M. Farach, M. Tagamets, and J. A. Reggia. Generating Plausible Diagnostic Hypotheses with Self-Processing causal Networks. *Journal of Experimental & Theoretical Artificial Intelligence*, 2:91-112, 1989.
- [Wang and Ayeb, 1992] S. Wang and B. e. Ayeb. Diagnosis: Hypothetical Reasoning With A Competition-Based Neural Architecture. In *Proceedings of the International Joint Conference on Neural Networks*, pages I-7-I-12, Baltimore, MA (USA), 1992.