STABLE THEORIES IN AUTOEPISTEMIC LOGIC

W.Marek

Department of Computer Science
University of Kentucky
Lexington, KY 40506-0027

ABSTRACT

We investigate the operator producing a stable theory out of its objective part (A stable theory is a set of beliefs of a rational agent). We characterize the objective parts of stable theories. Finally, we discuss the predicate calculus case.

Section 1: Introduction

Recent developments in the artificial intelligence and, in particular, strong interest in the formalizations of the common sense reasonings and nonmonotonic logics ([MC], [Re2], [Li], [MDD]) and reasoning about knowledge leads to new interesting developments in the areas of logic previously left almost exclusively to philosophers. These subjects now get attention of computer scientists and mathematicians, raising hopes of applicability and of increased mathematical rigour.

In one of these contributions, Moore [Mo], successfully formalizes the ideas of [MDD]; the resulting system of modal logic, called autoepistemic logic, deals with the notion of beliefs of a fully rational agent. In particular, Moore has shown that a stable autoepistemic theory is determined by its objective part. (The same result is proved in [HM]). He has not, however, provided the explicit construction of T out of its objective part. In this note, we investigate the connection between the objective part of a stable autoepistemic theory and the theory itself. In particular, we show an effective construction of a theory out of its objective part. We prove that every propositional theory closed under propositional consequence (in a language without modality) is the objective part of a stable autoepistemic theory. The construction is related to that of [FHV]. These results are shown in the Sections 2 and 3. In the Section 4, we consider the case of predicate calculus (which seems to be the original task of both [MDD] and [Mo], although they restrict themselves to the propositional logics in their papers). We extend the stability conditions by strong properties (similar to ω - completeness of Godel). This allows us to prove analogons of Barkan's formulas and, consequently, various normal forms for the formulas. This in turn allows to extend Moore's results to the (restricted) predicate calculus case.

Below we introduce some basic concepts of this paper. We presuppose the acquaintance with Moore's paper, even though we repeat some of the basic definitions.

Let L be the language of the propositional calculus and L_M its corresponding modal extension.

A theory T 'i L_M is *stable* if and only if it satisfies Stalnaker's conditions:

1) T is closed under propositional (tautological) provability.

2) *F‘bT*=>**L** F‘bT

3) *F‘BT*=>–**L** F‘bT

(**L** is the dual of the modal operator **M**)

$L_M|n$ is the set of formulas of L_M of **L**-depth at most n (thus $L = L_M|0$).

If T ‘i L_M then Obj(T) = T ∩ L, is the set of objective formulas of T.

Cn is the usual propositional (tautological) consequence operation.

Given T ‘i L_M define:

$Z(0,T) = Cn(T) \cap L$

$Z(n+1,T) = Cn[Z(n,T) \cup \{\mathbf{L}F: F \text{ ‘b } Z(n,T)\} \cup \{L\text{-}F: F \text{ ‘b } L_M|n - Z(n,T)\}] \cap L_M|(n+1)$

Finally, set:

$Z(T) = \bigcup_0^\infty Z(n,T)$

Section 2: Characterization

The following theorem shows the relevance of the operator Z to the notion of stability:

Theorem 2.1. If T is stable then T is equal to Z(Obj(T)).

Proof: We show that for all n, $Z(n,Obj(T)) = T \cap L_M|n$.

Once this is shown Z(Obj(T)) = T since the inclusion ‘i is obvious and if F ‘b T then for n equal to the **L**-depth of F,

$$F \text{ ‘b } T \cap L_M|n = Z(n,Obj(T)) \text{ ‘i } Z(Obj(T))$$

We will prove the equalities:

$$Z(n,Obj(T)) = T \cap L_M|n$$

by induction on n.

Case of n = 0. We need to show that:

$$Cn(Obj(T)) \cap L = Obj(T)$$

The inclusion $\supseteq$ is obvious. To see the other, notice that if F ‘b $Cn(Obj(T)) \cap L$ then F ‘b $Cn(T) \cap L = Obj(T)$.

Inductive step. Assume that $Z(n,Obj(T)) = T \cap L_M|n$.

We prove that:

$$Z(n+1,Obj(T)) = T \cap L_M|n+1$$

First we prove the inclusion ‘i.

Let $U_{n+1}(Obj(T)) = Z(n,Obj(T)) \cup \{\mathbf{L}F : F \text{ ‘b } Z(n,Obj(T))\} \cup \{\text{-}\mathbf{L}F : F \text{ ‘b } L_M - Z(n,Obj(T))\}$

We claim that $U_{n+1}(Obj(T))$ ‘i T.

If F ‘b $U_{n+1}(Obj(T))$ then one of the following cases holds:

i) F ‘b Z(n,Obj(T))

ii) F ‘b {**L**G : G ‘b Z(n,Obj(T))}

iii) F ‘b {-**L**G : G ‘B Z(n,Obj(T))}

If i) holds then the formula F belongs to the theory T by the inductive assumption. If ii) holds then F is in

T by the inductive assumption and Stalnaker condition 2. Finally if iii) holds then F is -**L**G for some G of **L**-depth at most n, G ‘B Z(n,Obj(T)). By the inductive assumption G does not belong to T so -**L**G belongs to T (by condition 3) i.e. F is in T.

By condition 1 of stability $Cn(U_{n+1}(Obj(T))) \text{ ‘i } T$ and so $Cn(U_{n+1}(Obj(T))) \cap L_M|n+1$ is included in $T \cap L_M|n+1$. This proves the inclusion ‘i.

For the proof of the inclusion, ⊇, notice the normal form of Moore [Mo], that is the fact that every formula of L_M can be tautologically transformed to one of the form $\text{‘x } \Theta_i$ where each Θ_i is of the form:

$$G \text{ v } \mathbf{L}H_1 \text{ v ... v } \mathbf{L}H_k \text{ v -}\mathbf{L}H_{k+1} \text{ v ... v -}\mathbf{L}H_{k+r}$$

with G objective.

This normal form does not increase the **L**-depth of the resulting formula. Since both T and Z(n+1,Obj(T)) are closed under tautological consequence, it is sufficient to show that every formula of the form:

$$G \text{ v } \mathbf{L}H_1 \text{ v ... v } \mathbf{L}H_k \text{ v -}\mathbf{L}H_{k+1} \text{ v ... v -}\mathbf{L}H_{k+r}$$

of the depth at most n+1 and in T belongs to Z(n+1,Obj(T)).

Hence assume that this formula belongs to T. If any of $\mathbf{L}H_1$,... ,$\mathbf{L}H_k$ belongs to Z(n+1,Obj(T)) then the whole alternative belongs to Z(n+1,Obj(T)) and we are done.

The same argument works if any of $\text{-}\mathbf{L}H_{k+1}$, ... ,$\text{-}\mathbf{L}H_{k+r}$ belongs to Z(n+1,Obj(T)).

Hence we are left with the case when none of the $\mathbf{L}H_1$, ... ,$\mathbf{L}H_k$, $\text{-}\mathbf{L}H_{k+1}$, ... ,$\text{-}\mathbf{L}H_{k+r}$ is in Z(n+1,Obj(T)).

By virtue of the construction either **L**H or -**L**H belongs to $U_{n+1}(Obj(T))$ for every formula H of **L**-depth at most n. Therefore $\text{-}\mathbf{L}H_1$, ... ,$\text{-}\mathbf{L}H_k$,$\mathbf{L}H_{k+1}$, ... ,$\mathbf{L}H_{k+r}$ are all in $U_{n+1}(Obj(T))$. Consequently, :

$$H_1, \text{ ... },H_k \text{ ‘B } Z(n,Obj(T)), H_{k+1}, \text{ ... },H_{k+r} \text{ ‘b } Z(n,Obj(T))$$

By inductive assumption

$$H_1, \text{ ... },H_k \text{ ‘B T}, H_{k+1}, \text{ ... },H_{k+r} \text{ ‘b T}$$

Using Stalnaker’s conditions 2 and 3 we find that:

$$\text{-}\mathbf{L}H_1, \text{ ... },\text{-}\mathbf{L}H_k \text{ ‘b T}, \mathbf{L}H_{k+1}, \text{ ... },\mathbf{L}H_{k+r} \text{ ‘b T}$$

Using the resolution principle k+r times we find that G belongs to T. But G is objective so being in T it must belong to Obj(T). Now Obj(T) ‘i Z(n+1,Obj(T)) so G ‘b Z(n+1,Obj(T)), hence, the alternative

$$G \text{ v } \mathbf{L}H_1 \text{ v ... v } \mathbf{L}H_k \text{ v -}\mathbf{L}H_{k+1} \text{ v ... v -}\mathbf{L}H_{k+r}$$

belongs to Z(n+1,Obj(T)). This completes the proof and shows the complete characterization of stable T in its objective part.

Once we have indicated *how* a stable theory depends on its objective part (the fact that a stable theory depends on its objective part has been proved by Moore, see Theorem 2 in [Mo]) we may ask which theories in L are objective parts of stable theories. As may have been expected the answer is the most natural one; these are exactly the theories closed under the tautological consequence.

Theorem 2.2. Let T_0 be a consistent theory in L which is closed under consequence. Then there exists a unique stable theory T in L_M such that Obj(T) = T_0.

Proof: Consider $Z(T_0)$. We claim that this is a consistent and stable theory T such that Obj(T) = T_0.

To show consistency, we need to prove that there exists a valuation **I** under which $Z(T_0)$ is true. Since T_0 is consistent, there exists a valuation **I**' under which T_0 is true. This valuation is extended, inductively, to a valuation **I** under which all $Z(T_0)$ is true. Hence, $Z(T_0)$ is consistent. In fact one checks that every valuation making any of $Z(n,T_0)$ true can be extended to make all $Z(T_0)$ true.

We now prove that $Z(T_0)$ is stable.

1) $Z(T_0)$ is closed under tautological consequence.

If F ‘b $Cn(Z(T_0))$ and is of the depth at most n, choose $m \geq n$ such that all the axioms used to prove F from $Z(T_0)$ are in $Z(m,T_0)$. By construction, F ‘b $Z(m,T_0)$ i.e., F ‘b $Z(T_0)$.

2) If F ‘b $Z(T_0)$ then F ‘b $Z(m,T_0)$ for some m. But then **L**F ‘b $Z(m+1,T_0)$ and consequently: **L**F ‘b $Z(T_0)$.

3) If F ‘B $Z(T_0)$ and m is the **L**-depth of F then F ‘B $Z(m,T_0)$ hence -**L**F ‘b $Z(m+1,T_0)$ ‘i $Z(T_0)$

i.e. -**L**F ‘b $Z(T_0)$.

Now we prove that Obj(T) = T_0. Otherwise, let F ‘b Obj(T) - T_0. (Let us point that T_0 is obviously included in T thus in its objective part). Since $Cn(T_0) = T_0$ there exists, by completeness theorem, a valuation **I** which makes T_0 true and F false. This valuation **I** can be extended all the way up to the valuation making T true. But then F must be true under the extended valuation, hence, contradiction.

This, in fact, proves our theorem but let us notice that we have an even stronger fact:

$$Z(n,T_0) = Cn(T) \cap L_M|n$$

(This uses the extension of valuations property mentioned above).

Let us look at some corollaries to the theorems 1 and 2.

Corollary 2.1. (Theorem 2 of [Mo]). If T_1 and T_2 are stable and $Obj(T_1) = Obj(T_2)$ then $T_1 = T_2$.

Proof: Both T_1 and T_2 are equal Z(T), where $T = Obj(T_1) = Obj(T_2)$.

Let us recall (from [Mo]) that we call a theory T *sound* w.r.t. an initial set of premises A iff every autoepistemic interpretation of T in which all the formulas of A are true is an autoepistemic model of T.

Corollary 2.2. A stable theory T is sound w.r.t A ‘i L if and only if Obj(T) ‘i Cn(A).

Corollary 2.3. ([HM], Proposition 2) No stable set properly includes another stable set.

Proof. If $Obj(T_1) = Obj(T_2)$ then $T_1 = T_2$. Hence, assume that F ‘b $Obj(T_1)$ - $Obj(T_2)$. Then **L**F_1 ‘b T_1 and -**L**F ‘b T_2, contradiction.

Section 3: An algorithm

Theorems 2.1 and 2.2 show how stable autoepistemic theories depend on their objective parts. It happens that a stable autoepistemic theory is produced out of its objective part in an infinite process of iterated construction which resembles a fixed point construction in the theory of inductive definitions. In fact, we deal with a sequence of operators, each operator $Z(n,\cdot)$ acting on subsets of $L_M|n$ only.

The algorithm for testing if the formula F belongs to $Z(T_0)$, where T_0 is a recursive theory included in L, can be derived from the proof of the Theorem 2.1. We list it below.

Given F, find its modal-objective normal form:

$$F <=> \text{`x } \Theta_i$$

where each Θ_i is of the form:

$$G \vee \mathbf{L}H_1 \vee ... \vee \mathbf{L}H_k \vee \text{-}\mathbf{L}H_{k+1} \vee ... \vee \text{-}\mathbf{L}H_{k+r}$$

Then F belongs to $Z(T_o)$ iff each Θ_i belongs to $Z(T_o)$

This gives the rise of the following algorithm written in a Pidgin Pascal:

Put F into the modal-objective normal form `x Θ_i

Check(Θ) {Θ is represented as the array of its clauses}

for i := 1 **to** s **do**

  if bad(Θ_i) **then return** *false;*

return *true*

bad(Θ_i)

{Θ_i is represented as a tuple <k,r,(F_0,...,F_{k+r})>}

if memb(F_o, T_0) **then return** *false;*

{memb($\cdot$,T_0) is the membership test for T_0}

for i := 1 **to** k **do**

if check(F_i) **then return** *false;*

for i := k+1 **to** k+r **do**

if not check(F_i) **then return** *false;*

return *true*

Correctness of the above algorithm follows from the argument of Theorem 2.2, stability conditions and the fact that F_j's are always of smaller **L**-rank than F.

As a bonus we get the following proposition:

Proposition 3.3: If T_0 is closed under consequence then $Z(T_0)$ is recursive in T_0. Consequently, if T_0 is recursive, so is $Z(T_0)$.

The representation $T = \bigcup_0^\infty Z(n,Obj(T))$ shows that the nonmonotonicity is appearing at every level of the construction.

One may treat the generation of Z(n+1,˙) from Z(n,˙) as an application of a version of the Closed World Assumption (see eg., [GMN] or [Re1]) on the sentences of form **L**F (add -**L**F if F cannot be derived from Z(n,˙)). Our intuition tells us that this similarity is not accidental; some common principle stands behind it.

Section 4: Predicate autoepistemic logic

For the purpose of this section, we shall consider a restricted predicate calculus in which there is a set of constants **A** with some strong properties with regard to all possible sets of beliefs of a rational agent.

In fact, the question of properties of the set of beliefs of a rational agent becomes much more complicated in the case of predicate calculus. In particular, what does it mean that the formula of a form ‘Qxϕ belongs to the set of beliefs of the agent? There are various, more or less constructive, approaches to this problem. The one we adopt is the following: ‘Qx ϕ(x) is believed by the agent if and only if for every object *a* he can name, ϕ(**a**) belongs to his set of beliefs (**a** is the name of the object *a*). Similarly, ‘qx ϕ(x) belongs to his set of beliefs if and only if he is able to pinpoint an object *a* such that ϕ(**a**) is in his set of beliefs.

Formalization of these principles is contained in conditions 4 and 5 below. These conditions are very strong but at the same time one can say that the set of constants **A** has the "Occam razor" property; i.e., there is essentially no entity whose properties are different from those of elements of **A**. One can also say that according to these conditions the model of the objective part of T is an elementary extension of the restriction of that model to **A**.

We formulate the stability conditions in the present context as follows:

1) T is closed under predicate calculus provability (cf [RS]).

2) $\phi\text{‘}bT => L\,\phi\text{‘}bT$

3) $\phi\text{‘}BT == > -L\,\phi\text{‘}bT$

4) ‘Qxϕ ‘b T <==> For all *a* belonging to **A**, ϕ(**a**) ‘b T

5) ‘qxϕ ‘b T <==> There exists *a* in **A** such that ϕ(**a**) ‘b T

A theory satisfying conditions 1 - 4 is called *weakly stable*, and a theory satisfying conditions 1 - 5 *stable*.

Proposition 4.1. A complete theory is weakly stable if and only if it is stable.

Proposition 4.2. If a theory T is weakly stable then:

a) $M\text{‘}qx\,\phi\text{‘}bT <=> \text{‘}qxM\,\phi\text{‘}bT$

b) $L\,`Qx\,\phi\,`bT <=> `QxL\,\phi\,`bT$

Proof: a) ==> Assume M ‘qxϕ ‘b T, i.e. for ψ = -ϕ, -‘Qxψ ‘b T (by condition 1). By condition 3, ‘Qxψ ‘B T. Using now condition 4 we find an *a* ‘b **A** such that ψ(**a**) ‘B T and so -Lψ(**a**) ‘b T. Again by condition 1, ‘qx -Lψ ‘b T and finally ‘qx M ϕ ‘b T.

<== Assume that ‘qx M ϕ ‘b T and that T is consistent (otherwise the argument is obvious!). Then -‘Qx L ψ ‘b T, hence ‘Qx L ψ ‘B T and so by condition 4 there is an *a* ‘b **A** such that L ψ(**a**) ‘B T. For this *a*, ψ(**a**) ‘B T. Hence ‘Qxψ ‘B T and so -L‘Qxψ ‘b T. Hence M‘qxϕ ‘b T.

b) Rather straightforward.

Proposition 4.3. If T is stable, then T proves the commutativity of quantifiers and autoepistemic operators in other words, in addition to a) and b) of Proposition 2, we have:

c) M ‘Qxϕ ‘b T <==> ‘Qx M ϕ ‘b T

d) L ‘qxϕ ‘b T <==> ‘qx L ϕ ‘b T

Corollary 4.1. If T is stable, then for every formula θ there is a prenex normal form formula θ’ such that:

$$T \mid\!- \theta <=> \theta'$$

Moreover we can assume that the matrix of θ’ is in the modal normal form of Moore [Mo].

Proposition 4.4. If T is a stable theory, then T is uniquely determined by its quantifier-less part.

Proof: Put θ in the prenex normal form. Then eliminate quantifiers using conditions 4 and 5.

Corollary 4.2. If T is a stable theory, then T is uniquely determined by its objective, quantifier-less part.

Proof: Using Proposition 4 and the theorem 3.2. of Moore [Mo].

Finally let us state a slight extension of one of the results of Section 3.

Proposition 4.5. If T is stable theory, then T is hyperarithmetical in its objective, quantifier-less part.

Actually, a stonger result, locating T in the hyperarithmetical hierarchy with respect to its objective quantifier-less part is possible.

June 1986

References

[FHV] Fagin, R., Halpern, J.Y., Vardi, M.Y., A Model-Theoretic Analysis of Knowledge: Preliminary Report. *Proceedings of the 25th F.O.C.S. Symposium* (1984), pp. 268-278.

[GMN] Gallaire, H., Minker, J., Nicolas, J.-M., Logic and Databases: A Deductive Approach. *ACM Computing Surveys 16* (1984), pp. 153-186.

[HM] Halpern, J.Y., Moses, Y., Towards a Theory of Knowledge and Ignorance: Preliminary Report. *Non-Monotonic Reasoning Workshop* A.A.A.I. 1984, pp. 125-143.

[Li] Lifschitz, V., Some Results on Circumscription, Non-Monotonic Reasoning Workshop, A.A.A.I. 1984.

[MC] McCarthy, J., Circumscription- A Form of Non-Monotonic Reasoning, *Artificial Intelligence 13* (1980) pp. 27-39.

[MDD] McDermott, D., Doyle, J. Non-monotonic Logic I, *Artificial Intelligence 13* (1980) pp. 41-72.

[Mo] Moore, R.C., Semantical Considerations on Nonmonotonic Logic, *Artificial Intelligence 25* (1985), pp. 75-94.

[RS] Rasiowa, H., Sikorski, R., *Mathematics of the Metamathematics*, PWN, Warszawa, 1966.

[Re1] Reiter, R., On closed world databases. In: *Logic and Databases*, H. Gallaire and J. Minker Editors, Plenum Press, New York 1978, pp. 56-76.

[Re2] *Non-Monotonic Reasoning Workshop* A.A.A.I. 1984, R.Reiter, editor.