Preliminary Considerations on the Modelling of Belief Change Operators by Metric Spaces

James P. Delgrande
School of Computing Science
Simon Fraser University
Burnaby, B.C.,
Canada V5A 1S6
jim@cs.sfu.ca

Abstract

In this paper, a possible worlds framework for representing general belief change operators is presented. In common with many approaches, an agent's set of beliefs are specified by a subset of the set of possible worlds. The central intuition is that there is a distance given between every pair of possible worlds, giving the similarity of one world to another; the set of worlds together with this distance form a metric space. An operator such as revision is defined as expected: in revising by proposition A, the revised belief state is characterised by those worlds closest to the worlds characterising the agent's original beliefs in which A is true. We propose that a suitable fundamental change operator is one where an agent's beliefs are decreased as a result of the agent becoming more skeptical to a specified degree. This approach is compared and contrasted with the central quantitative framework for belief change, that due to Spohn.

Introduction

The notion of *similarity* between possible worlds (or, alternatively, interpretations) has played a prominent role in the modelling of belief change operators. For example, Grove's modelling of theory change (Grove 1988) is based on the system of spheres semantics of (Lewis 1973), in which a revision function is determined based on the possible worlds most similar to a given set of worlds representing a belief state. Other authors have expanded on this theme, or proposed other approaches to belief change based on a notion of relative similarity or closeness between possible worlds.

We address belief revision based on a notion of similarity between possible worlds in the following way. To begin with, we assume a notion of "conceivable states of affairs", or ways that the world could be imagined to be, together with a subset of these worlds which correspond to the agent's actual set of beliefs. A "conceivable state of affairs" is modelled by the notion of a possible world, an object to which we associate a truth assignment, an assignment of $true$ or $false$ to every atomic sentence. A given subset of the possible worlds characterises the agent's beliefs, stipulating the ways the (actual) world could be assuming that the agent's beliefs are in fact true.

We also assume that there is a *distance* given between every pair of possible worlds, giving the degree to which the two worlds are similar. The intuition is that one has some general, fixed, overarching theory (Newtonian or relativistic mechanics for example) given a priori, by which the "distance" between two worlds would be measured. Thus, given such a theory, the relative similarity between worlds (containing contingent information) would be fixed. That is, since belief change is a result of the acquisition of new contingent information (by revision, update, or whatever), there is no need to change the background distance metric. For example, a world that is similar to the present world might be, at first pass, a world that differs in few details from our own, where the notion of "similarity" is in part bound by physical laws. Hence a world like the present, except that my cat is moved 1 cm. horizontally is presumably quite similar to the world at hand, whereas one in which my cat can levitate 1 cm. above the ground is not.

We make the relatively strong assumption (at least with respect to the literature) that the set of possible worlds and distance function in fact define a *metric space*, where the distance function, among other properties, satisfies the triangle inequality. This distance between worlds allows for the specification of an absolute notion of similarity between worlds; as well, it allows for the specification of relative similarity, for example that world w is more similar to w_1 than to w_2. From this it is straightforward to define a revision function that satisfies the AGM postulates: every subset of possible worlds in this model (given a couple more plausible assumptions) induces a system of spheres, and then via Grove's result we can specify a revision function, where the revision of the set of worlds K by proposition A is characterised by the closest A-worlds to K.

Work along these lines is not new. For example (Dalal 1988) and others propose approaches in which the distance between two worlds is given by the number of atomic sentences on which the worlds disagree with respect to their truth values. (Spohn 1988) uses *ordinal conditional functions* to characterise quantitative extensions to belief change operators such as revision and contraction; see also (Williams 1994). (Lehmann *et al.*

2001) develop a general approach to distance-based belief revision based on a qualitative notion of distance that they call a *pseudo-distance*; see also (Delgrande *et al.* 2003).

Interestingly, (Lehmann *et al.* 2001) argues that revision functions are too weak, or coarse-grained, to adequately characterise pseudo-distance. If one accepts their result, then this would seems to indicate that an investigation of the stronger, quantitative, notion of a metric distance would be of limited use with respect to belief revision. While this may be the case with respect to belief revision, it is not necessarily so with respect to other belief change operators. In particular, we show that an operator of *skepticism*, whereby an agent becomes more skeptical to a given degree, is sufficient to fully capture a given metric space (under a couple of standard assumptions). As well, other belief change operators are easily defined in terms of this notion of skepticism along with standard (set theoretic) operators for working with sets of possible worlds.

The approach would appear to be of interest for several other reasons. First, as with other distance-based approaches, it straightforwardly allows iterated belief revision. Unlike approaches founded on that of (Spohn 1988) it does not rely on the notion of an evolving *epistemic state* (beyond the trivial fact that the agent's contingent knowledge will evolve). As well, unlike (Spohn 1988) and related work, in which the similarity relation between worlds is modified, the proposed approach adopts a static similarity relation. Further, we suggest that the approach at hand may lead to the development of interesting elaborations of the standard belief change framework.

The next section discussed related work in belief change in more detail, while Section introduces the proposed approach. Section discusses the proposed central belief change operator, constituting increased skepticism on the agent's beliefs. This is followed by a brief concluding section.

Belief Revision

Formal Preliminaries

We assume a logical language $\mathcal{L}$ over a set of atomic sentences $\mathbf{P} = \{p, q, \ldots\}$, closed under the usual connectives $\neg$, $\wedge$, $\vee$, $\supset$, and $\equiv$, and with distinguished element $\bot$ for falsity. The truth relation is denoted by $\models$, and is defined in the usual fashion. A *theory* is a logically closed set of sentences, that is a set of sentences T which satisfies the constraint:

$$\phi \in T \text{ iff } T \models \phi.$$

In the literature, the set of beliefs of an agent have been formally modelled by two principal means. In many approaches to belief revision, including the original papers on the subject (see e.g. (Alchourrón *et al.* 1985)), an agent's beliefs are modelled by sets of sentences, called *belief sets*, corresponding to theories in a logic that includes classical propositional logic. Alternatively, other approaches have employed a modal framework, based on a set of possible worlds W. A possible world $w \in W$ is an object to which a (complete and consistent) truth assignment to atomic sentences is associated. We assume that distinct worlds are assigned distinct truth assignments. The truth of a sentence at a possible world is given by the expected definition. We also write $w \models \alpha$ to assert that α is true at w. Capital letters such as K, A denote subsets of W or *propositions*. For $A \subseteq W$, $\bar{A}$ will stand for $W \setminus A$. We define

$$\|\alpha\| = \{w \in W \mid w \models \alpha\}$$

as the proposition expressed by α. We will generally use K, perhaps subscripted, to denote an agent's knowledge base, expressed as a set of possible worlds. Thus for $K \subseteq W$, an agent believes α just if for every $w \in K$ we have $w \models \alpha$, or, what comes out to the same thing, $K \subseteq \|\alpha\|$. Conversely, given a set of possible worlds, $K \subseteq W$ the corresponding belief set is given by $|K| = \{\alpha \mid$ for every $w \in K$ we have $w \models \alpha\}$ or, what comes out to the same thing, $|K| = \{\alpha \mid \|\alpha\| \subseteq K\}$.

Belief Revision and Similarity

Belief revision is the process whereby an agent changes its beliefs in order to incorporate new information. The seminal work in this area is the *AGM approach* (Alchourrón *et al.* 1985), in which standards for revision functions are given by *rationality postulates*. The intent is to describe belief change at the *knowledge level*, that is on an abstract level, independent of how beliefs are represented and manipulated. As described above, belief states are modelled by sets of sentences, called *belief sets*, corresponding to logical theories. For belief set T and formula ϕ, $T+\phi$ is the deductive closure of $T \cup \{\phi\}$, the *expansion* of T by ϕ. $T_\bot$ is the inconsistent belief set (i.e. $T_\bot = \mathcal{L}$). *Revision* represents the situation in which the new information may be inconsistent with the reasoner's beliefs and needs to be incorporated in a consistent manner where possible. A revision function $*$ is a function from $2^{\mathcal{L}} \times \mathcal{L}$ to $2^{\mathcal{L}}$; given space constraints and general familiarity with the approach, we omit a listing of the revision postulates.

Two well-known constructions for belief revision operators have been proposed. The first is that of *epistemic entrenchment* (Gärdenfors and Makinson 1988). An epistemic entrenchment ordering related to a belief set T is a total preorder $\leq$ on the formulas in $\mathcal{L}$, reflecting the relative degree of acceptance of sentences. Various conditions are given for an entrenchment ordering, including the stipulations such as sentences not in T are minimally entrenched while logical truths are maximally entrenched. Given an entrenchment ordering, a corresponding revision function can subsequently be defined. Gärdenfors and Makinson show that the set of revision functions definable via entrenchment orderings corresponds exactly to the class of functions satisfying the preceding postulates.

The second construction builds on work by David Lewis characterising counterfactual assertions (Lewis

1973). Grove, in (Grove 1988) uses Lewis' *system of spheres* semantics to obtain a modelling of the AGM postulates. We have the following additional notation: A theory U is *complete* just if for every $\alpha \in \mathcal{L}$ we have $\alpha \in U$ or $\neg\alpha \in U$. The set of all complete, consistent theories is denoted $M_{\mathcal{L}}$; hence these theories are analogous to interpretations or possible worlds. The letters U, V, X, ... denote subsets of $M_{\mathcal{L}}$, while T denotes an arbitrary theory. For $\alpha \in \mathcal{L}$ we define $[\![\alpha]\!] = \{I \in M_{\mathcal{L}} \mid I \models \alpha\}$.[1]

Definition 1 ((Grove 1988)) *A set of subsets $\mathcal{S}$ of $M_{\mathcal{L}}$ is a* system of spheres centred on X *where $X \subseteq M_{\mathcal{L}}$, if it satisfies the conditions:*

S1 $\mathcal{S}$ *is* totally ordered *by* $\subseteq$.

S2 X *is the minimum of* $\mathcal{S}$ *under* $\subseteq$.

S3 $M_{\mathcal{L}} \in \mathcal{S}$.

S4 *If* $[\![\alpha]\!] \neq \emptyset$ *then there is a least (wrt* $\subseteq$*) sphere* $c(\alpha)$ *such that* $c(\alpha) \cap [\![\alpha]\!] \neq \emptyset$ *and* $U \cap [\![\alpha]\!] \neq \emptyset$ *implies* $c(\alpha) \subseteq U$ *for every* $U \in \mathcal{S}$.

$f_S(\alpha)$ is defined to pick out the least (if such there be) interpretations containing α, that is

$$f_S(\alpha) = [\![\alpha]\!] \cap c(\alpha).$$

From this, a revision function can be defined with respect to a given system of spheres: If $\mathcal{S}$ is a system of spheres in $M_{\mathcal{L}}$ centred on $[\![T]\!]$ and $\alpha \in \mathcal{L}$ where $\not\models \neg\alpha$ then define

$$T * \alpha = \cap\{x \in f_S(\alpha)\}.$$

For $\models \neg\alpha$, the revision is taken to be $\mathcal{L}$.

Grove shows that the set of functions generated by systems of spheres is exactly the set of functions given by the AGM revision postulates. In a sense then, this work can be seen as specifying a static, 3-place similarity relation on interpretations: For a system of spheres centred on X, for $I_1 \in X$ we can say that I_1 is not less similar to I_2 than to I_3 just if there are spheres U, V where $I_2 \in U$, $I_3 \in V$ and $U \subseteq V$.

A limitation of these constructions is that they do not address the issue of iterated belief revision; nor of course do the AGM postulates deal with iterated revision in any substantive way. To address this, it has been argued (perhaps implicitly in some cases) that one needs to consider not just a belief set and recipe for a single revision, but rather one needs to consider, along with the agent's current beliefs, an encoding of the strategy that the agent uses to revise its beliefs. The belief set and revision strategy can be called an *epistemic state*, and the challenge now is to revise not just an agent's belief set, but also its epistemic state.

The seminal work here is that of Spohn (Spohn 1988), which develops *ordinal conditional functions* as a way of characterising belief revision. Subsequent work in this area includes (Boutilier 1993; Williams 1994; Darwiche and Pearl 1997). In this framework, epistemic states are represented by rankings on possible worlds. An ordinal conditional function (OCF) or ranking, denoted κ, is a function from the set of possible worlds into the class of ordinals, representing the degree of plausibility of a possible world, such that some possible world(s) are assigned to 0. The ranking is extended to propositions, represented by sets of possible worlds, so that the rank of a proposition is the least rank assigned to a possible world that satisfies the proposition. That is, for $A \subseteq W$ where $A \neq \emptyset$ define

$$\kappa(A) = \min\{\kappa(w) \mid w \in A\}.$$

A ranking *accepts* a proposition $A \subseteq W$ if the negation of the proposition is implausible, i.e. $\kappa(\bar{A}) > 0$.

A change in belief is represented by a pair (A, m) where $\emptyset \neq A \subseteq W$ is a proposition and m is the post-revision degree of plausibility of A, called the (A, m)-*conditionalisation of* κ:

$$\kappa(A, m)(w) = \begin{cases} \kappa(w) - \kappa(A) & \text{if } w \in A \\ m + \kappa(w) - \kappa(\bar{A}) & \text{if } w \notin A. \end{cases}$$

Thus the A part of the OCF is effectively shifted uniformly[2] so that the least A worlds have value 0, while the $\bar{A}$ part is shifted uniformly so that the least $\bar{A}$ worlds have value m. As Spohn describes, this approach generalized both revision and contraction to quantitative versions of these operators. Thus one can regard A being held with firmness m following the (A, m)-conditionalisation of κ.

As with Grove's construction involving a system of spheres, an OCF can be seen as specifying (or, equivalent to) a three-place similarity relation on (here) possible worlds. OCFs extend Grove's construction in two main ways. First, an OCF provides a *quantitative* ranking on worlds, rather than the relative, *qualitative* ranking on interpretations in a system of spheres. Thus in an OCF one may have no worlds assigned a particular index, although there may be worlds assigned a greater index; this degree of precision in inexpressible in a system of spheres. Second, the similarity relation corresponding to an OCF is *dynamic*, in that following a conditionalisation operation, one obtains a different similarity relation.

Various researchers (Borgida 1985; Dalal 1988; Forbus 1989; Satoh 1988; Weber 1986; Winslett 1988) have proposed specific revision (and update) operators by defining specific distance functions between interpretations. For the revision of formula ψ by μ, the result corresponds to the set of models of μ closest (in the given specific distance) to models of ψ. For example, the revision operator in (Dalal 1988) uses the Hamming distance between interpretations as metric, where the Hamming distance $d(w_1, w_2)$ between two interpretations w_1 and w_2 is the number of propositional variables on which the interpretations differ. The distance between an interpretation w and the models of ψ is given by:

$$d(Mod(\psi), w) = min_{w_i \models \psi} d(w_i, w).$$

[1] Grove uses the notation $|\alpha|$ for $[\![\alpha]\!]$.

[2] That is, so that the relative positions of A-worlds remains unchanged.

A pre-order on interpretations is given by:

$$w_1 \leq_\psi w_2 \text{ iff } d(Mod(\psi), w_1) \leq d(Mod(\psi), w_2).$$

Revision is defined by:

$$Mod(\psi *_D \mu) = Min(Mod(\mu), \leq_\psi).$$

The resulting operator, $*_D$, satisfies the AGM postulates. An iterated version of Dalal's operator follows straightforwardly, since the distance between interpretations (viz. the Hamming distance) is constant over revisions. In (Peppas *et al.* 2000) the notion of similarity is explicitly employed in capturing Winslett's PMA approach (Winslett 1988), in terms of conditions on systems of spheres.

Recently, Lehmann, Magidor, and Schlechta have explored general distance-based approaches to belief revision (Lehmann *et al.* 2001); see also (Delgrande *et al.* 2003). In their approach, a distance function d is defined on all pairs of interpretations. The distance function d is called a *pseudo-distance*, that is, a binary function whose range is a total order. The authors allow that d may not be symmetric. As well d may not respect identity, where d respects identity iff: $d(a, b) = 0$ iff $a = b$. (For this latter property, the authors consider only violating one direction, although it seems that either direction may plausibly not hold.) Belief revision is a function on two arguments, each consisting of sets of formulas, whose result is a belief set, the theory resulting from the revision. Revision of a theory by a theory representing a formula for revision is characterised by those interpretations of the formula that are closest to those of the original theory.

The authors consider conditions corresponding to the AGM postulates together with the following condition:[3]

$*$**S1** If: $T_1 * (T_0 \vee T_2) \not\vdash T_0$, $T_2 * (T_1 \vee T_3) \not\vdash T_1$
$\ldots T_k * (T_{k-1} \vee T_0) \not\vdash T_{k-1}$
then: $T_0 * (T_k \vee T_1) \not\vdash T_1$.

The central result (slightly paraphrased) is the following:

Theorem 1 (Lehmann *et al.* 2001): *A revision operator $*$ is representable by a symmetric, consistency-preserving, identity-preserving pseudo-distance iff it satisfies the AGM postulates and $*$**S1**.*

The authors note that revision operators are relatively crude means for representing distances, and they provide an example in which distances cannot be compared by looking at the results of revisions. Their example is somewhat intricate. Figure 1 is simpler but illustrates the problem. Essentially the difficulty is that (obvious approaches to) using revision functions to represent the relative distances in Figure 1 fail, specifically in the assertion that $d(x_1, x_2) < d(x_3, x_4)$. The problem is that these distances cannot be compared via revision functions. Now, we can reflect, for example, that $d(x_1, x_3) < d(x_1, x_2)$ via $x_1 * (x_3 \vee x_2) = x_1 * x_3$.

[3]This is for the symmetric case; the authors also consider the not-necessarily symmetric case.

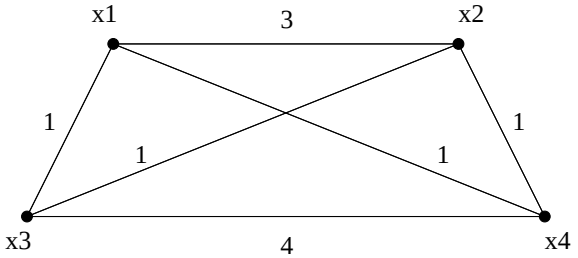

Figure 1

However, it seems that all attempts to extend this to the more general case fail. The obvious extension (viz. involving $(x_1 \vee x_3) * (x_2 \vee x_4)$) fails, as do all similar formulations; in each case, the other distances with value 1 interfere with a comparison between $d(x_1, x_2)$ and $d(x_3, x_4)$.[4]

Regarding underlying similarity relations in belief revision then, it can be seen that there are two distinct approaches: In (Grove 1988) and (Spohn 1988), similarity is treated as a three-place relation. Originally, Lewis (Lewis 1973) used $w_1 \leq_w w_2$ to mean that world w_1 is at least as similar to the world w as the world w_2 is. The relation $\leq_w$ is then asserted to be a total preorder with minimum element w, yielding a system of spheres centred on w. In a Grove-style system of spheres, or a Spohnian OCF centred on a set K, this can be written $w_1 \leq_K w_2$ to mean that world w_1 is at least as similar to the knowledge base K as the world w_2 is. The (A, m)-conditionalisation of an OCF then gives a means of determining the new $\leq$-relation following a belief change operation.

The notion of similarity can be generalised to that of relative distance between pairs of worlds. Hence one can write

$d(w_1, w_2) \leq d(w_3, w_4)$ for
"w_1 is at least as similar to w_2 as w_3 is to w_4".

This reading is common to the specific approaches exemplified by (Dalal 1988), as well as (Lehmann *et al.* 2001).

We note lastly that (Williamson 1988) has studied these general notions of similarity. The three-place relation is given by the predicate:

$S(X, Y, Z)$ for
"X is at least as similar to Y as Z is to Y".

A more general, four-place, notion of similarity is given as follows:

$T(W, X, Y, Z)$ for
"W is at least as similar to X as Y is to Z".

[4]This is of course an informal argument (as is the one in (Lehmann *et al.* 2001)). While it seems clear that such distances cannot be captured by a revision function, it is not so obvious that a suitably imaginative definition of distance might not prove adequate. Similarly, it is not clear what a proof of impossibility would look like.

From a four-place relation, a corresponding three-place relation is obtained via:

$$S(X, Y, Z) \equiv T(X, Y, Z, Y).$$

As well, (Hansson 1992) addresses belief revision based on this notion of similarity, where the similarity relation is taken as holding among sets of formulas. Specific constructions are proposed founded on intuitions concerning the degree to which sets differ and, alternatively, the degree to which they concur.

Modelling Belief Change Operators by Metric Spaces

A Model for Belief Change Operators

We begin with a set of possible worlds W, corresponding to different ways that the world could conceivably (factually or counterfactually) be. An agent's beliefs K is modelled by a subset of these possible worlds, corresponding to the different ways that the actual world might be, assuming that the agent's beliefs are in fact correct. Thus $w \in K$ just if, for all the agent knows, w could in fact be the actual world. The agent's belief set can then be defined by $\{\alpha \in \mathcal{L} \mid w \models \alpha \text{ for every } w \in W\}$. Further, for every pair of possible worlds, the agent will have an opinion as to how similar one world is to another. This is expressed, at least initially, as a non-negative real number. We assume that no world is more similar to another than it is to itself; as well we assume that similarity is symmetric.

Hence, we assume that we have a distance function d, or *metric*, where $d : W \times W \rightarrow \mathbf{R}^{\geq 0}$ and:

1. $d(w_1, w_2) = 0$ iff $w_1 = w_2$
2. $d(w_1, w_2) = d(w_2, w_1)$
3. $d(w_1, w_2) + d(w_2, w_3) \geq d(w_1, w_3)$.

Condition 3 is the *triangle inequality* and (W, d) is a *metric space*. Before we can define belief revision, we require a further assumption:

4. For $A, B \subseteq W$, $\exists w_1 \in A, \exists w_2 \in B$, such that $\forall w_3 \in A, \forall w_4 \in B$ we have $d(w_1, w_2) \leq d(w_3, w_4)$.

Condition 4 is a *limit assumption* (Lewis 1973). In Lewis's approach (and consequently Spohn's) this assumption restricts the range of d to the set of ordinal numbers. Given our condition of symmetry, this assumption restricts the range of d here to the (non-negative) integers.

The distance function d is extended to sets of possible worlds by, for $\emptyset \neq A, B \subseteq W$:

$$d(A, B) = d(w_1, w_2) \quad \text{where} \quad w_1 \in A, w_2 \in B,$$

$$\text{and} \quad \forall w_3 \in A, \forall w_4 \in B, d(w_1, w_2) \leq d(w_3, w_4).$$

The limit assumption guarantees that this extension is well-defined.

We will ultimately want to relate belief change operators, expressed in terms of operations on sets of possible worlds, to knowledge bases (or belief sets) expressed by formulas or collections of formulas. This in turn necessitates another assumption concerning our models, since we may have $W_1, W_2 \subseteq W$ where $W_1 \neq W_2$ yet $\{\alpha \mid W_1 \models \alpha\} = \{\alpha \mid W_2 \models \alpha\}$ (i.e. different belief sets may be verified by distinct sets of possible worlds.) Consequently we assume that the distance function d is the same for such sets of possible worlds.

Definition 2 *A function f on $2^W \times 2^W$ is* syntax preserving *iff $f(A_1, B) = f(A_2, B)$ whenever $\{\alpha \mid A_1 \models \alpha\} = \{\alpha \mid A_2 \models \alpha\}$.*

Thus we have our last assumption:

5. d is syntax-preserving.

The distance-based approach of (Lehmann *et al.* 2001) addresses this issue by restricting the operation for belief change to that of *definability-preserving* operations. Here, in contrast we restrict the notion of distance rather than any defined operator. For a discussion of underlying issues, see (Lakemeyer and Levesque 2000), who also show how sets of worlds that verify the same sets of formulas may be represented by a (maximal) canonical element.

Belief change operators are defined with respect to a model, as given next.

Definition 3 *An* epistemic metric space *is a tuple $M = \langle W, K, d, P \rangle$ where*

1. *W is a set (of possible worlds);*
2. *$\emptyset \neq K \subseteq W$;*
3. *$\langle W, d \rangle$ is a metric space satisfying assumptions 4 and 5 above;*
4. *$P : \mathbf{P} \mapsto 2^W$.*

Belief Revision

Given an epistemic metric space (henceforth, "model") M, the revision of K by A is easily defined, as being comprised of the closest set of A-worlds to the worlds in K according to d.

Definition 4 (Revision) *Let M be a model. The function $* : 2^W \times 2^W \mapsto 2^W$ is given by:*

$$\text{If } K = \emptyset \text{ then } K * A = A.$$

Otherwise:

$K * A = \{w \in A \mid \exists w_1 \in K$ *such that* $\forall w_2 \in A, \forall w_3 \in K,$ *we have* $d(w, w_1) \leq d(w_2, w_3)\}$.

This is different from the standard conception of belief revision, which is a function from a belief set and formula to a belief set. The phrasing above, in terms of possible worlds is conceptually simpler, so we stick with this formulation. It is obvious that any $K \subseteq W$ induces a system of spheres on W, and so revision (once phrased in terms of possible worlds) satisfies the AGM postulates. Alternatively, standard AGM revision, in which the arguments to a revision operator are a belief set and formula, can be expressed by $|(\|K\| * \|\alpha\|)|$.

Since the function $*$ is total, Definition 4 trivially supports iterated revision. As described in the previous section, this approach has been investigated in (Lehmann *et al.* 2001) and (Delgrande *et al.* 2003) for a weaker, qualitative, notion of distance. As was pointed out in these papers, the standard notion of belief revision is too weak, or coarse-grained, to fully capture a distance semantics.[5] Moreover, a straightforward argument shows that the triangle inequality has nothing of substance to say concerning distance-based revision, in that the triangle inequality places no constraints on a revision function defined in terms of a metric space model.

To see this, we need to assume that revisions based on distance are independent of the uniform addition of a constant. As well, we assume here that we have a finite language. So, call two distance-based models M and M' over the same (finite) language *similar* just when M and M' are identical, except that their respective distance functions d and d' are such that for all $\emptyset \neq P, Q \subseteq W$ where $P \cap Q = \emptyset$, we have

$$d(P, Q) = d'(P, Q) + c$$

for some fixed integer constant c. (For $P \cap Q \neq \emptyset$, we have $d(P, Q) = 0$ in any revision function.) Arguably, the revision functions captured by similar models should be considered to be identical in all respects.

Consider a distance based model M with associated distance function d; without loss of generality, assume that d ranges over nonnegative integers. Let m be the maximum (distance) over all values of d, i.e.

$$m = \max\{d(w_1, w_2) \mid w_1, w_2 \in W\}.$$

Define a new model M' that is identical to M, except that the distance function d' is defined by:

1. If $d(P, Q) = 0$ then $d'(P, Q) = 0$.
2. If $d(P, Q) \neq 0$ then $d'(P, Q) = d(P, Q) + m$.

The models M and M' are *similar* according to the above definition. Consider arbitrary distinct possible worlds w_1, w_2, w_3. We have that

$$d(w_1, w_2), d(w_2, w_3), d(w_1, w_3) \in [1,\ m].$$

Thus $d'(w_1, w_2), d'(w_2, w_3), d'(w_1, w_3) \in [m + 1,\ 2m]$. Hence $d'(w_1, w_2) + d'(w_2, w_3) \in [2m + 2,\ 4m]$. But this means that $d'(w_1, w_2) + d'(w_2, w_3) > d'(w_1, w_3)$. Hence the triangle inequality is satisfied for w_1, w_2, w_3.

Since w_1, w_2, w_3 are arbitrary, this means that the triangle inequality is satisfied by every triple of possible worlds in M'. (The case of zero distances is straightforward and doesn't alter the argument.) Thus, for every model there is a *similar* model in which the triangle inequality trivially holds. Consequently, since the same revision function is determined by similar distance based models, the triangle inequality has nothing of substance to say concerning distance based revision.

These results would seem to limit the usefulness of metric spaces as a means of modelling revision functions. However, as we argue in the full paper, the additional granularity of metric spaces makes them appropriate for modelling other belief change operators, in particular belief set merging. This also raises the question as to whether there are other operators that may adequately capture a metric space. In the next subsection we examine one such operator.

Skeptical Belief Change

Given the notion of a model as defined above, we propose a primitive, characterising, belief change operator that, for lack of a better term, we will call *(quantitative) skepticism.*[6] In a model, an agent has a set of beliefs $K \subseteq W$ modelled by a set of possible worlds, and a notion of the distance between every pair of worlds. Thus it makes sense to ask what an agent would believe, should it become more skeptical, or cautious, about what it believes. Assuming integral distances, if an agent were to become more skeptical by degree i, then the agent's beliefs would be given by the set of worlds of distance no greater than i from the worlds in K. This gives rise to the following:

Definition 5 *Let M be a model. The function $Sk : 2^W \times \mathbf{I} \mapsto 2^W$ is given by:*

$$Sk(K, i) \ = \ \{w \in W \ \mid \ d(w, K) \leq i\}$$

It is straightforward to define belief revision in terms of Sk:

Definition 6 *Let M be a model. The function $* : 2^W \times 2^W \mapsto 2^W$ is given by:*

$$\textit{If } \ K = \emptyset \ \textit{ then } \ K * A \ = \ A.$$

Otherwise: $K * A \ = \ Sk(K, d(K, A)) \ \cap \ A.$

That is, revision by A corresponds to just those worlds in A that are obtained by the agent becoming minimally (more) skeptical so that it believes A is possible. Clearly contraction can be similarly defined, either directly, or via the Harper Identity. As well, Sk is closely related to (a quantitative version of) *severe withdrawal* (Rott and Pagnucco 1999): the severe withdrawal of A would require the agent becoming skeptical to a minimal degree sufficient to include a $\bar{A}$ world.

Discussion

It is instructive to compare this framework with that of (Spohn 1988). The primary point of similarity between these approaches is their quantitative aspect; further, both allow iterated revision. However, beyond these points the approaches differ significantly. First, similarity as defined in (Spohn 1988) is a three-place relation

[5] That is, in the sense of the Grove representation result, in which there is a correspondence between systems of spheres and revision operators.

[6] The term "skeptical" isn't terrific, given its use as a type of default reasoning; however it isn't clear what would make a better alternative.

(see Section), whereas here similarity can be regarded as a 4-place relation. In (Spohn 1988), an ordinal conditional function (OCF) represents an epistemic state, and a conditionalisation of an ordinal conditional function yields a new epistemic state. That is to say, the three-place similarity relation is *dynamic*. In contrast, in the approach at hand, we have a *static* four place similarity relation, that nonetheless allows iterated revision. Consequently, in the present approach, we do not require the notion of an epistemic state, or, perhaps more accurately, we need deal only with a single, static epistemic state within which belief revision can be defined for all knowledge bases and sentences for revision.[7]

This gives rise to the question as to whether there are reasons for preferring one approach over another. I feel that a plausible case can be made for the metric space approach proposed here. First, the proposed approach is arguably founded on a plausible intuition, that given a fixed background theory, an agent will have a fixed, contingent-information-independent notion as to how similar two possible worlds are. Revision then concerns new contingent information about the world at hand, applied in the context of this background information. In OCFs there is similarly some sort of background information, reflected in the initial OCF, but where similarity is relative to the contingent state of affairs, and the similarity relation itself is modified as a result of new contingent information. Arguably then an OCF conflates the distinction between contingent knowledge and a background theory. As well, there are various ways in which one may modify an OCF to reflect belief change operators. However, of the well-known proposals (for example (Boutilier 1993; Williams 1995; Darwiche and Pearl 1997)) inappropriate properties are obtained in some plausible scenarios (see (Darwiche and Pearl 1997; Delgrande and Schaub 2003)).

Last, it is not clear that conditionalisation in an OCF appropriately reflects how an underlying 3-place similarity relation should be modified. In a revision in an OCF by a proposition A, the closest A-worlds are moved to rank 0, and no $\bar{A}$-world has rank 0. There are two main strategies by which $\bar{A}$-worlds are dealt with. In (Boutilier 1993; Darwiche and Pearl 1997) the $\bar{A}$-worlds are shifted minimally, and so the least $\bar{A}$-worlds will have rank 1 whenever some $\bar{A}$-world was considered possible by the original knowledge base. In (Papini 2001) the opposite tack is taken, and no $\bar{A}$-world has lower rank than any A-world.

Arguably both strategies may sometimes yield a non-intuitive notion of similarity. Consider the following example: I currently believe that Sherlock Holmes was an actual person; further I believe that the current temperature is 18°. I am informed that Sherlock Holmes was in fact not a real person and that the current temperature is 19°. (That is, I believed something of the form $p \wedge q$ and am informed that $\neg p \wedge \neg q$.) In the case of an OCF, for approaches such as (Boutilier 1993; Darwiche and Pearl 1997), following the revision asserting that Sherlock Holmes was not a real person and the temperature is 19°, at the least set of worlds not at level 1, it would be true that Sherlock Holmes existed and that the temperature is 18°. Given a second revision, with the fact that the temperature is 18°, one would also lose the information that Sherlock Holmes was not a real person, in reverting to the closest set of possible worlds in which it is raining. Hence, this means of updating an OCF seem to employ an at-least-sometimes inappropriate notion of locality in implementing revisions, in that the most recently-discarded information is closest (or: most similar) to the current knowledge base. In (Papini 2001), the opposite approach is taken: given that an agent believes that the temperature is 18° and is informed that it is 19°, *every* world in which the temperature is 18° is ranked higher than any world where the temperature is 19°. Thus, among the 19° worlds there is a world in which the polar icecaps have melted, and this world is ranked more similar to the knowledge base than another world in which the temperature is 18° but the icecaps are intact. Again, this seems counterintuitive with respect to the ranking of contingent information.

Conclusion

We have discussed a general approach to distance-based belief revision in a possible worlds framework, in which the set of possible worlds and distance function form a metric space. In revising a set of beliefs by proposition A, the revised belief state is characterised by the closest A worlds to the worlds representing the agent's original beliefs. Revision functions are however generally too weak to fully capture notions of distance – in fact too weak to capture even weaker notions of distance than employed here. Consequently we propose an alternative belief change operator, in which the agent becomes more skeptical about its beliefs. This operator, together with standard (set-theoretic) operations on possible worlds, is adequate to capture the suite of revision-type belief change functions. As well, by modifying the definition of "closest" between propositions, the approach could easily capture the update-style operators of (Katsuno and Mendelzon 1992).

The approach can be contrasted with that of ordinal conditional functions. While superficially similar (both employ quantitative distances between possible worlds), there are significant differences. First, an OCF provides a three-place similarity relation, while the approach at hand effectively employs a four-place relation. In an OCF, the similarity relation itself is modified; hence an epistemic state is modified in a belief change operation, rather that simply the agent's beliefs. In contrast, in the approach at hand, a model can be regarded as providing a (static) epistemic state, reflecting in part

[7] Obviously the current approach could be generalised to allow for the distances between worlds to vary following a belief change, but it is not clear what advantages would accrue to such a generalisation.

an agent's background knowledge or theory. Hence belief revision, and other belief change operations, concern changes in an agent's contingent beliefs against this background theory.

There are several ways in which this approach can be extended. First, properties of the basic system can be further developed and explored. As well, we are interested in using the framework to explore an extended version of belief revision, in which an agent's beliefs are held with various degrees of conviction, along with its non-beliefs. Further, the approach is readily extendible to deal with multiple agents, both in the case of merging the belief sets of different agents, and, in an extension to the approach, having a different distance function and knowledge base associated with each agent.

References

C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

A. Borgida. Language features for flexible handling of exceptions in information systems. *ACM Transactions on Database Systems*, 10, 1985.

C. Boutilier. Revision sequences and nested conditionals. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 519–531, 1993.

M. Dalal. Investigations into theory of knowledge base revision. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, pages 449–479, St. Paul, Minnesota, 1988.

A. Darwiche and J. Pearl. On the logic of iterated revision. *Artificial Intelligence*, 89:1–29, 1997.

J. Delgrande and T. Schaub. A consistency-based approach for belief change. *Artificial Intelligence*, 151(1-2):1–41, 2003.

J.P. Delgrande, O. Papini, and O. Schulte. Considerations on distance-based belief revision. In *IJCAI-03 Workshop on Nonmonotonic Reasoning, Action and Change*, pages 80–85, Acapulco, Mexico, August 2003.

K.D. Forbus. Introducing actions into qualitative simulation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1273–1278, 1989.

P. Gärdenfors and D. Makinson. Revisions of knowledge systems using epistemic entrenchment. In *Proc. Second Theoretical Aspects of Reasoning About Knowledge Conference*, pages 83–95, Monterey, Ca., 1988.

A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

S. O. Hansson. Similarity semantics and minimal changes of belief. *Erkenntnis*, 37:401–429, 1992.

H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In P. Gärdenfors, editor, *Belief Revision*, pages 183–203. Cambridge University Press, 1992.

S. Konieczny and R. Pino Pérez. Merging information under constraints: A logical framework. *Journal of Logic and Computation*, 12(5):773–808, 2002.

G. Lakemeyer and H.J. Levesque. *The Logic of Knowledge Bases*. MIT Press, Cambridge, Mass., 2000.

D. Lehmann, M. Magidor, and K. Schlechta. Distance semantics for belief revision. *Journal of Symbolic Logic*, 66(1):295–317, 2001.

D. Lewis. *Counterfactuals*. Harvard University Press, 1973.

P. Liberatore and M. Schaerf. Arbitration (or how to merge knowledge bases). *IEEE Transactions on Knowledge and Data Engineering*, 10(1):76–90, 1998.

T. Meyer. On the semantics of combination operations. *Journal of Applied NonClassical Logics*, 11(1-2):59–84, 2001.

O. Papini. Iterated revision operations stemming from the history of an agent's observations. In M.-A. Williams and H. Rott, editors, *Frontiers in Belief Revision*, volume 22 of *Applied Logic Series*, pages 279–301. Kluwer Academic Publishers, 2001.

P. Peppas, N. Foo, and A. Nayak. Measuring similarity in belief revision. *Journal of Logic and Computation*, 10(4):603–618, 2000.

H. Rott and M. Pagnucco. Severe withdrawal (and recovery). *Journal of Philosophical Logic*, 28(5), 1999.

K. Satoh. Nonmonotonic reasoning by minimal belief revision. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, pages 455–462, Tokyo, 1988.

W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, volume II, pages 105–134. Kluwer Academic Publishers, 1988.

A. Weber. Updating propositional formulas. *Proc. First Conference on Expert Database Systems*, pages 487–500, 1986.

M.-A. Williams. Transmutations of knowledge systems. In J. Doyle, P. Torasso, and E. Sandewall, editors, *Proceedings of the Fourth International Conference on the Principles of Knowledge Representation and Reasoning*, pages 619 – 629, Bonn, Germany, May 1994.

M.-A. Williams. Iterated theory base change: A computational model. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1541–1547, Montréal, 1995.

T. Williamson. First-order logics for comparative similarity. *Notre Dame Journal of Formal Logic*, 29(4):457–481, 1988.

M. Winslett. Reasoning about action using a possible models approach. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, pages 89–93, St. Paul, Minnesota, 1988.