

Belief Revision as a Truth-tracking Process

[Extended Abstract]^{* †}

Alexandru Baltag
Institute for Logic, Language
and Computation
P.O. Box 94242
1090 GE Amsterdam
TheAlexandruBaltag@gmail.com

Nina Gierasimczuk
Department of Artificial
Intelligence
Postbus 407
9700AK Groningen
Nina.Gierasimczuk@gmail.com

Sonja Smets
Department of Artificial
Intelligence
Postbus 407
9700AK Groningen
S.J.L.Smets@rug.nl

ABSTRACT

We analyze the learning power of iterated belief revision methods, and in particular their universality: whether or not they can learn everything that can be learnt. We look in particular at three popular methods: conditioning, lexicographic revision and minimal revision. Our main result is that conditioning and lexicographic revision are universal on arbitrary epistemic states, provided that the observational setting is sound and complete (only true data are observed, and all true data are eventually observed) and provided that a non-standard (non-well-founded) prior plausibility relation is allowed. We show that a standard (well-founded) belief-revision setting is in general too narrow for this. We also show that minimal revision is not universal. Finally, we consider situations in which observational errors (false observations) may occur. Given a fairness condition (saying that only finitely many errors occur, and that every error is eventually corrected), we show that lexicographic revision is still universal in this setting, while the other two methods are not.

Categories and Subject Descriptors

I.2.6 [Computing Methodologies]: Artificial Intelligence—*Learning*; I.2.4 [Computing Methodologies]: Artificial Intelligence—*Knowledge Representation Formalisms and Methods*; I.2.3 [Computing Methodologies]: Deduction and Theorem Proving —*Nonmonotonic reasoning and belief revision*

*For a full journal version of this work, which includes the relevant definitions, proofs and further explanations of all our results, see [4].

†The work of Nina Gierasimczuk and Sonja Smets is funded by the VIDI research programme with number 639.072.904, which is financed by the Netherlands Organisation for Scientific Research and is hereby gratefully acknowledged.

ACM COPYRIGHT NOTICE. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Publications Dept., ACM, Inc., fax +1 (212) 869-0481, or permissions@acm.org. TARK 2011, July 12-14, 2011, Groningen, The Netherlands. Copyright ©2011 ACM. ISBN 978-1-4503-0707-9, \$10.00.

General Terms

Theory

Keywords

belief revision, learning theory, universality, identification in the limit, conditioning

1. INTRODUCTION

At the basis of intelligent decision making lies the idea that agents adopt an effective method to integrate new information into a set of prior beliefs and knowledge. Intelligent agents (in Game Theory, AI, Epistemic Logic and Belief Revision theory) are assumed to be endowed with some *belief-changing method*, allowing them to revise their beliefs on the basis of their new observations. But how good are these methods, in the long run? What is their *learning power*: how effective and how reliable are they for eventually finding the truth? Are they *universal*: can they learn everything that is learnable?

The standard setting for revising beliefs in Logic and AI is the classical AGM framework [1], as extended to iterated belief revision and given a firm semantic basis by the adoption of Grove’s “sphere” models [20], and its equivalent relational presentations in [8, 11, 21, 34, etc.]. These frameworks are based on the popular “possible world” semantics, coupled with a *plausibility relation* between possible worlds. Belief is defined as truth in all “plausible enough” worlds. In the most commonly used such setting (which we refer to as the “standard” one), the plausibility relation is assumed to be *well-founded*, in which case belief can be easier defined as truth in all the “most plausible” worlds. In this paper, we do not restrict ourselves to standard (well-founded) relations, since we want to give the agent as many chances for learning as possible. Indeed, non-well-foundedness will turn out to be essential for universality.

The philosophy underlying the AGM postulates is a “conservative” one: keep as much as possible of the old beliefs, and of the old belief structures, while incorporating the new information. A large number of belief-revision methods satisfying these postulates have been proposed (see, e.g., [33]). The most well-known are the three that we study in this paper: *conditioning* (the natural qualitative analogue of probabilistic conditioning), *lexicographic revision* and *minimal revision*. The “most conservative” choice is Boutilier’s “minimal revision”, which seems to ideally capture the AGM ethos of minimal change.

The goal of investigating the learning power of iterated belief revision has appeared before within at least two lines of research. On the one hand, in the research line of [31, 32] the connection between Learning Theory and Belief Revision has been rooted in the more syntactic framework of the classical AGM approach [1], and embodied into a first-order framework. While lying within the same learning paradigm as our work (the so-called “set learning” paradigm), this framework is actually very far from our semantical, modal framework. On the other hand, the work of Kelly, Schulte and Hendricks [24, 25, 26, 27, 28] belongs to the “function learning” paradigm: learning (the future of) an infinite stream of incoming data, based on its past (the data already observed). Like us, these authors adopt a semantic, relational framework, based on “possible worlds” and preference relations, and study identifiability in the limit by various belief-revision methods. But their “possible worlds” are identified with the sequence of all future data, and the order (in which the data are observed) matters. Their learning problem is one of prediction: using belief revision to predict the pattern of all future observations. In contrast, we focus on another type of learning problem altogether, belonging to the “set learning paradigm”. This is the problem of learning the set of all data that are true in the “real world”, based on the finite sequence of data that were already observed. Here, it is assumed that the data are observed in more or less random manner, so predicting the future sequence is not feasible, or even relevant. The two problems (function learning versus set learning) are very different from each other. Learning paradigm which requires prediction of the future sequence of events (function learning) makes the corresponding universality results weaker (and easier to prove) than ours. This explains the positive results in [24] according to which all AGM methods are in a sense good (in contrast to our results). However, the setting based on function learning can be recovered as a special case of ours (that of “tree-like” epistemic states).

As in [16, 17, 18], in this paper we link the set learning paradigm with the epistemic and doxastic logics of belief revision [3, 5, 6, 7, 8, 9, 15]. More concretely, we consider sequences of specific learning events, i.e., situations in which finite sets of data (some observable properties) are observed sequentially. In this setting we are concerned with the problem of belief convergence to “full truth” in finitely many steps. This is the doxastic counterpart of one of the central notions in Learning Theory, namely *identifiability in the limit* [19]. Hence, we focus on achieving a *stable true (justified) belief* (rather than $S5$ “knowledge”, in the strong, absolute sense of partition-based epistemic models). This is an advantage, since this absolute sense of “irrevocable knowledge” given by partition models is generally criticized in epistemology as representing a unrealistic concept, not fit to represent the knowledge we possess in day-to-day life or in natural sciences. Indeed, at least one philosophical school [35] considers something like our “stable, true justified belief” to be a philosophically more appropriate definition of knowledge. We relax our notion of knowledge in order to be generous to the learning agent, not requiring him to achieve an unrealistic standard of certainty: it is enough if his beliefs reliably (and justifiably) converge to full truth.¹

¹Achieving absolute, irrevocable knowledge can be linked to a more restrictive kind of learning—finite identification [see 13, 14].

In this setting we show that the most conservative method, Boutilier’s minimal revision [12], is actually *the least favorable one for learning*. In contrast, we show that *condition-alization and lexicographic revision are “universal”* learning methods: i.e. they can reliably learn the real world, whenever the initial epistemic state is such that the real world is reliably “learnable” (via *any* learning method). We show this in a much more general context than the one of Kelly et alia, that of arbitrary epistemic states (corresponding to what in Learning Theory is called “learning from positive data”). This means that we don’t assume closure of observable data under negation: the fact that an observable property does *not* hold is not necessarily observable.² There is a price to pay for this: our proofs are harder. *Not all prior beliefs* (prior plausibility relations) lead to convergence to the truth by conditioning. The choice of prior is important, and very specific to the initial epistemic state. Most of our work in the proof goes into constructing an appropriate prior. Moreover, we show that sometimes only a non-well-founded prior allows our revision methods (conditioning and lexicographic revision) to realize their full learning potential. So our conclusion is that *a non-standard (non-well-founded) setting for Belief Revision is absolutely necessary for its universality*.

While first restricting ourselves (as it is standard in Learning Theory) to truthful observations (sound and complete data streams), later on we extend the setting to allow for *errors in observations*. Provided that errors occur only finitely often and are always eventually corrected, we show there still exist belief-revision methods that are universal (as successful as it is possible) in such fallible observational settings. But now *only one of the three investigated methods can do this: lexicographic revision*.

2. CONVERGENCE TO TRUTH

An *epistemic space* (S, Φ) consists of a set S of epistemic possibilities (“possible worlds”), together with a family of relevant observable properties $\Phi \subseteq \mathcal{P}(S)$. An infinite stream $\epsilon = (\epsilon_1, \epsilon_2, \dots) \in \Phi^\omega$ of successive observations is *sound and complete* with respect to a given world $s \in S$ iff the set $\{\epsilon_n : n \in N\}$ of all properties that are observed in the (infinite history of the) stream coincides with the set $\{P \in \Phi : s \in P\}$ of all observable properties that are true in the given world. A *learning method* is a map L that associates to any epistemic space (S, Φ) and any finite sequence of observations $\sigma = (\sigma_0, \dots, \sigma_n)$ (of *any finite length* n), some *hypothesis*, i.e. a subset $L(S, \Phi; \sigma) \subseteq S$ of the possible worlds. A world $s \in S$ is *learnable by a method* L if, for every observational stream ϵ that is sound and complete for s , there exists a finite stage N such that $L(S, \Phi; \epsilon_0, \dots, \epsilon_n) = \{s\}$ for all $n \geq N$. The epistemic space (S, Φ) is itself said to be *learnable* by L if all its worlds are learnable by L .

A *plausibility space* (S, Φ, \leq) is an epistemic space (S, Φ) together with a total preorder \leq on S , called plausibility order. A *belief-revision method* is a function R that associates

²This generality is important for applications: there are experimental studies suggesting that, in many situations, such as in language acquisition, negative examples are irrelevant, and the agent learns almost only from observing positive examples of language use.

to any plausibility space (S, Φ, \leq) and any observational sequence $\sigma = (\sigma_0, \dots, \sigma_n)$, some new plausibility space

$$R(S, \Phi, \leq; \sigma) := (S^\sigma, \Phi^\sigma; \leq^\sigma),$$

with $S^\sigma \subseteq S$ and $\Phi^\sigma = \{P \cap S^\sigma : P \in \Phi\}$. A belief-revision method R , together with a prior-plausibility assignment $(S, \Phi) \mapsto \leq_S$, generate in a canonical way a learning method L , and given by: $L(S, \Phi, \sigma) := \text{Min } R(S, \Phi, \leq_S, \sigma)$, where $\text{Min}(S', \leq')$ is the set of all the least elements of S' with respect to \leq' (if such least elements exist) or $= \emptyset$ (otherwise). An epistemic space (S, Φ) is learnable by a belief-revision method R if there exists some prior plausibility assignment $(S, \Phi) \mapsto \leq_S$ such that (S, Φ) is learnable by the associated learning method $L(S, \Phi, \leq_S)$.

Learning methods differ in their learning power. We are interested in the most powerful among them, those that are *universal*: they can learn any epistemic state that is learnable (by any other method). We are particularly interested in finding whether any AGM-like belief revision methods are universal. Our main result in this sense is:

THEOREM 1. *There exist universal AGM-like iterated belief revision methods. Namely, conditioning and lexicographic revision are universal.*

The main technical difficulty of the proof is the construction of an appropriate prior plausibility order. For this, we use some classical learning-theoretic concepts and results (locking sequences introduced in [10], finite tell-tale sets proposed in [2], as well the simple non-computable version of Angluin’s theorem [2]). But in order to construct a suitable prior plausibility we need to refine these concepts and improve on the above-mentioned results.

On the other hand, we show that Boutilier’s minimal revision [12] is the least favorable method for learning:

PROPOSITION 1. *Minimal revision is not universal.*

For the above universality results, our non-standard setting (involving non-well-founded plausibility orders) is essential: no AGM-like belief revision method is universal with respect to well-founded prior plausibility orders.

PROPOSITION 2. *No AGM-like belief-revision method is standardly universal.*

While first restricting ourselves to truthful observations, we can extend the setting to allow for observational *errors*. For this, we now give up the soundness of data streams, and replace it by a “fairness” assumption: errors occur only finitely often and are always eventually corrected. Unsurprisingly, conditioning (which assumes that absolute veracity of the new observations) is no longer a good strategy. If erroneous observations are possible, then eliminating worlds that don’t fit these observations is risky and irrational.

PROPOSITION 3. *Conditioning and minimal revision are not universal for fair streams.*

Only one of our three belief-revision methods can successfully cope with errors.

PROPOSITION 4. *Lexicographic revision is standardly universal for fair streams.*

3. CONCLUSIONS

In this paper, we consider iterated belief-revision policies of conditioning, lexicographic and minimal revision. In the full paper, we identify certain features of those methods relevant in the context of iterated revision: data-retention, conservatism, and history-independence. We define learning methods based on those revision policies and we have shown how the aforementioned properties influence the learning process. Throughout this paper we are mainly interested in convergence to the actual world on the basis of infinite data streams. We show that conditioning and lexicographic revision generate universal learning methods for sound and complete data streams. Minimal revision fails to be universal, and the crucial property that makes it weaker is its strong conservatism. Moreover, we show that the full power of learning cannot be achieved when the underlying prior plausibility assignment is assumed to be well-founded. Finally, in the setting of fair streams (containing a finite number of errors that all get corrected later in the stream) lexicographic revision again turns out to be universal, while conditioning and minimal revision fail to be so. The conclusion is that the choice of a specific belief revision method, as well as the choice of prior (conditional) beliefs (i.e. prior plausibility relation), is *not* an indifferent matter, from a Learning theoretic perspective: various belief-revision policies differ with respect to their reliability.³

The epistemic state resulting from a *successful* belief-revision process need not be as strong as irrevocable knowledge, i.e., the S5 type of knowledge known in Logic or Aumann Knowledge in Game Theory and Economics. It rather matches the *defeasible* type of “knowledge” proposed by Lehrer [29, 30] and others, formalized by Stalnaker in [35] and rediscovered by modal logicians under the name of “safe belief”. The strength of safety is in the guarantee it gives: a safe belief is not endangered by the new truthful observations. Hence, if we restrict our considerations to truthful information, or at least assume that mistakes happen rarely, safety can be reformulated in terms of stability. In other words, defeasible “knowledge” emerges when stability is reached. The need for such a notion appeared in many different frameworks: from reaching an agreement in a conversational situation [see, e.g., 29, 30] to the considerations in the domain of philosophy of science [see, e.g., 22]. In this paper we account for the emergence of such stable true belief: we explicitly formulate the conditions under which certain belief states may give rise to this kind of defeasible “knowledge”.

References

- [1] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.
- [2] D. Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45(2):117–135, 1980.
- [3] G. Aucher. A combined system for update logic and belief revision. *Master’s thesis, ILLC, University of Amsterdam*, 2003.

³The analysis of inductive inference in terms of reliability has been for the first time provided in [23].

- [4] A. Baltag, N. Gierasimczuk, and S. Smets. Belief revision as a truth-tracking process. manuscript, 2011.
- [5] A. Baltag and S. Smets. Conditional doxastic models: a qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science*, 165:5–21, 2006.
- [6] A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *LOFT'06: Proceedings of 7th Conference on Logic and the Foundations of Game and Decision Theory*, pages 11–24. University of Liverpool, 2006.
- [7] A. Baltag and S. Smets. The logic of conditional doxastic actions. In R. van Rooij and K. Apt, editors, *New Perspectives on Games and Interaction. Texts in Logic and Games*. Amsterdam University Press, 2008.
- [8] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *LOFT'08: Proceedings of 8th Conference on Logic and the Foundations of Game and Decision Theory*, number 3 in Texts in Logic and Games, pages 9–58. Amsterdam University Press, 2008.
- [9] J. Van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 2:129–155, 2007.
- [10] L. Blum and M. Blum. Toward a mathematical theory of inductive inference. *Information and Control*, 28:125–155, 1975.
- [11] O. Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49(1):49–80, 2004.
- [12] C. Boutilier. Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic*, 25:263–305, 1996.
- [13] C. Dégremont and N. Gierasimczuk. Can doxastic agents learn? On the temporal structure of learning. In X. He, J. F. Horty, and E. Pacuit, editors, *LORI'09: Proceedings of 2nd International Workshop on Logic, Rationality, and Interaction*, volume 5834 of *Lecture Notes in Computer Science*, pages 90–104. Springer, 2009.
- [14] C. Dégremont and N. Gierasimczuk. Finite identification from the viewpoint of epistemic update. *Information and Computation*, 209(3):383–396, 2011.
- [15] H. Van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Knowledge, Rationality & Action (Synthese)*, 147:229–275 (41–87), 2005.
- [16] N. Gierasimczuk. Bridging learning theory and dynamic epistemic logic. *Synthese*, 169(2):371–384, 2009.
- [17] N. Gierasimczuk. Learning by erasing in dynamic epistemic logic. In *LATA'09: Proceedings of 3rd International Conference on Language and Automata Theory and Applications*, volume 5457 of *Lecture Notes in Computer Science*, pages 362–373. Springer, 2009.
- [18] N. Gierasimczuk. *Knowing One's Limits. Logical Analysis of Inductive Inference*. PhD thesis, Universiteit van Amsterdam, 2010.
- [19] E. M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.
- [20] A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.
- [21] J. Halpern. *Reasoning about uncertainty*. Cambridge, MA: MIT Press, 2003.
- [22] V. Hendriks. *The Convergence of Scientific Knowledge: A View from The Limit*. Dordrecht: Kluwer Academic Publishers, 2001.
- [23] K. Kelly. *The Logic of Reliable Inquiry*. Oxford University Press, Oxford, 1996.
- [24] K. Kelly, O. Schulte, and V. Hendriks. Reliable belief revision. In *Proceedings of the 10th International Congress of Logic, Methodology, and Philosophy of Science*, pages 383–398. Kluwer Academic Publishers, 1995.
- [25] K. T. Kelly. Iterated belief revision, reliability, and inductive amnesia. *Erkenntnis*, 50:11–58, 1998.
- [26] K. T. Kelly. The learning power of belief revision. In *TARK'98: Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 111–124, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.
- [27] K. T. Kelly. Learning theory and epistemology. In I. Niilniuto, M. Sintonen, and J. Smolenski, editors, *Handbook of Epistemology*. Dordrecht: Kluwer, 2004.
- [28] K. T. Kelly. Ockham's razor, truth, and information. In P. Adriaans and J. van Benthem, editors, *Handbook of the Philosophy of Information*. Elsevier, 2008.
- [29] K. Lehrer. Knowledge, truth and evidence. *Analysis*, 25(5):168–175, 1965.
- [30] K. Lehrer. *Theory of Knowledge*. Routledge, London, 1990.
- [31] E. Martin and D. Osherson. Scientific discovery based on belief revision. *The Journal of Symbolic Logic*, 62(4):1352–1370, 1997.
- [32] E. Martin and D. Osherson. *Elements of Scientific Inquiry*. MIT Press, Cambridge, 1998.
- [33] H. Rott. Shifting priorities: Simple representations for twenty-seven iterated theory change operators. In H. W. David Makinson, Jacek Malinowski, editor, *Towards Mathematical Philosophy*, pages 269–296. Springer, 2009.
- [34] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In *Causation in Decision, Belief Change, and Statistics, volume II*, pages 105–134, 1988.
- [35] R. Stalnaker. Iterated belief revision. *Erkenntnis*, 70(2):189–209, 2009.