Learning-theoretic take on: Belief Revision and Possible Histories

Nina Gierasimczuk

Institute for Logic, Language and Computation
University of Amsterdam


NASSLLI
Austin, Texas
June 21st, 2012

A conservative methodologist seeks to minimize the damage done to his current beliefs by new information. A reliabilist, on the other hand, seeks to find the truth whatever the truth might be. [...]

The inductive leap from a hundred black ravens to the universal generalization that all ravens are black is not conservative. Nor was Copernicus revolutionary rejection of conservative tinkering within the Ptolemaic system. [...]

Conservatism is the motivation behind the theory of belief revision proposed by [AGM]. Reliabilism is the principal concern of formal learning theory.


Kelly, K. (1998). Iterated Belief Revision, Reliability, and Inductive Amnesia, Erkenntnis, Vol. 50, pp. 11-58.


Kelly, K. (1998). The Learning Power of Iterated Belief Revision, in: Proceedings of the Seventh TARK Conference, Itzhak Gilboa (ed.), pp. 111-125.


Kelly, K., Schulte, O., and Hendricks, V. (1997). Reliable Belief Revision, in: Logic and Scientfic Methods, M. L. Dalla Chiara, et al. (eds.), Dordrecht: Kluwer.

Belief revision:

Propositional belief state B and new information φ give a new belief state B':

$$B * \varphi = B'.$$

The revision cannot depend on the history, except for what is recorded in B.

Instead:

Revision $*$ can operate on an doxastic state $\mathcal{B}$, that determines the belief:

$$\mathcal{B} * \varphi = \mathcal{B}'.$$

The doxastic state may record some or all of the agents updating history.

- W — the set of possible worlds.
- A proposition is a set of possible worlds.
- Belief state is all worlds that satisfy each proposition in the belief set.
- Hence, a belief state may be represented as a proposition.

The doxastic state $\mathcal{B} = (|\mathcal{B}|, \leq_{\mathcal{B}})$ determines:

- a belief state $B(\mathcal{B})$,
- but also a total pre-ordering $\leq_{\mathcal{B}}$ on the domain of $\mathcal{B}$.

The ordering $\leq_{\mathcal{B}}$ is the implausibility ordering induced by $\mathcal{B}$.

Definition

- Let $min_{\mathcal{B}}(\varphi)$ be the set of all $\leq_{\mathcal{B}}$-minimal elements of $[\varphi] \cap |\mathcal{B}|$.
- Also: $B(\mathcal{B}) = min_{\mathcal{B}}(W)$.

There is an agreement that $B(\mathcal{B} * \varphi)$ should be $min_{\mathcal{B}}(\varphi)$.

But: How should the rest of the doxastic state be revised?

Boutiliers natural method $*_M$ induces minimal modification of $\mathcal{B}$:

worlds in $min_{\mathcal{B}}(\varphi)$ go to the bottom, leaving the rest unaffected.

Definition

Formally, if φ is consistent with $|\mathcal{B}|$ and $\mathcal{B}' = \mathcal{B} *_M \varphi$ then:

$$w \leq_{\mathcal{B}'} w' \text{ iff } w \in min_{\mathcal{B}}(\varphi) \ \vee \ w \leq_{\mathcal{B}} w'.$$

Spohn's (generalized by Nayak) method $*_L$ induces lexicographic revision:

slides all φ-worlds below all $\neg\varphi$-worlds.

Definition

Formally, if φ is consistent with $|\mathcal{B}|$ and $\mathcal{B}' = \mathcal{B} *_L \varphi$, then:

$$w \leq_{\mathcal{B}'} w' \text{ iff } (w \in \varphi \wedge w' \notin \varphi) \vee ((w \in \varphi \Leftrightarrow w' \in \varphi) \wedge (w \leq_{\mathcal{B}} w')).$$

The doxastic state as a (partial) mapping r from possible worlds to ordinal-valued degrees of implausibility.

Definition

If $X \subseteq W$ is consistent with the domain of r, define

$$r_{min}(X) = min\{r(w') \in dom(r) \cap X\}$$

and the conditional implausibility of w given X is:

$$r(w \mid X) = -r_{min}(X) + r(w).$$

Intuition:
$r_{min}(|\mathcal{B}|)$ is the height of the most plausible member of $\mathcal{B}$.
$r(w \mid |\mathcal{B}|)$ is the height of w above the most plausible member of $\mathcal{B}$.

Spohns qualitative generalization of Jeffrey method $*_{J,\alpha}$:

it lowers all φ-worlds until the most plausible of them are at implausibility 0; and then lifts all $\neg\varphi$-worlds till the most plausible of them is at implausibility α.

Definition

$$(r *_{J,\alpha} \varphi)(w) = \begin{cases} r(w|[\varphi]) & \text{if } w \in dom(r) \cap [\varphi] \\ r(w|W - [\varphi]) + \alpha & \text{if } w \in dom(r) - [\varphi] \\ \uparrow & otherwise. \end{cases}$$

Darwiche and Pearl proposed a ratchet method $*_{R,\alpha}$:

it rigidly boosts the $\neg\varphi$ worlds up are by a fixed ordinal α.

Definition

$$(r *_{R,\alpha} \varphi)(w) = \begin{cases} r(w \mid [\varphi]) & \text{if } w \in dom(r) \cap [\varphi] \\ r(w) + \alpha & \text{if } w \in dom(r) - [\varphi] \\ \uparrow & otherwise. \end{cases}$$

Goldszmidt and Pearl discuss the procedure $*_{A,\alpha}$:

it boosts all $\neg\varphi$-worlds to the fixed ordinal α.

Definition

$$(r *_{A,\alpha} \varphi)(w) = \begin{cases} r(w \mid \varphi) & \text{if } w \in min_{\leq_r}([\varphi]) \\ \alpha & \text{if } w \in dom(r) - [\varphi] \\ \uparrow & otherwise. \end{cases}$$

The rules differ in how much of a boost they apply to refuted possible worlds:

- $*_M$ boosts them by one step, along with many non-refuted worlds.
- $*_L$ can boost infinitely.
- $*_{J,\alpha}$ may provide a negative boost if $\alpha <$ the most plausible refuted world.
- $*_{R,\alpha}$ sends refuted worlds up by a fixed increment.
- $*_{A,\alpha}$ sends refuted worlds up all to the same fixed level.

Problem: w' is the true state and a refuted w gets beneath w'.

If that happens, the agent forgets the past data refuting w.

Kelly: Sometimes the rules must forget if they want to predict the future.

Conditioning $*_C$ induces drastic modification of $\mathcal{B}$:

$\neg\varphi$-worlds are thrown away.

Definition

Formally, if φ is consistent with $|\mathcal{B}|$ and $\mathcal{B}' = \mathcal{B} *_C \varphi$ then:

$$r *_C \varphi = r(.|[\varphi]).$$

Rough Ball Experiment

Many people believe that a rough ball will have more air drag than a smooth one. Suppose we invite such a subject to consider the results of an experiment. We mount both balls in a wind tunnel and measure the drag on each. We start the tunnel at a small wind velocity and raise the velocity incrementally. After each increment, we report a 0 *to the subject if the drag on the smooth ball is no greater than that on the rough ball and we report a* 1 *otherwise. The experiment is run. The subject is smug as several* 0*s are presented, consistently with her current belief. But to her utter surprise, the sequence reverses and* 1*s continue for some time. Thereafter, the sequence flips back to* 0*s and yields* 0*s thereafter.*

Fred and Jane are engaged in an indefinitely repeated prisoners dilemma. Fred fully believes that Jane is a patsy who will cooperate no matter how often he defects. But maybe Jane simply has a veneer of civility that affords a fixed grace period of unconditional cooperation to new acquaintances to encourage good behavior, and punishes defections for eternity once this grace period is over. Or maybe she punishes the first infraction after the grace period for a fixed time and then returns to being a patsy forever (two tail reversals). Or maybe she punishes the first infraction after the grace period for a fixed time, offers a new grace period, and punishes the next infraction for eternity, etc.

$$e = (0000001111000\ldots)$$

- An outcome stream is an infinite sequence e of 0s and 1.
- U is the set of all such outcome streams.
- An empirical proposition's truth depends only on the outcome stream.

Example

Proposition: 'outcome b will be observed at position n in the outcome stream':

$$[n, b] = \{e' \in U \mid e'(n) = b\}.$$

We will use the following notation:

- data stream generated by e is $[e] = ([0, e(0)], [1, e(1)], [2, e(2)]\ldots)$;
- $e[k] = (e(0), e(1), \ldots, e(k_1))$;
- data sequence $[e|k] = ([0, e(0)], [1, e(1)], \ldots; [k-1, e(k-1)])$.

Given $*$ and $\mathcal{B}$, a revision agent is then a pair $(*, \mathcal{B})$.

- she is starting out in an a priori epistemic state $\mathcal{B}$;
- and then successively $*$-revising on the propositions input thereafter.

An adaptation of Gold's concept of identification in the limit.

DEFINITION

1. $(*, \mathcal{B})$ identifies $P \subseteq U$ in the limit just in case for each $e \in P$ there is a stage k such that for each subsequent $k' \geq k, (\mathcal{B} * [e[k']]) = \{e\}$.
2. $*$ identifies $P \subseteq U$ in the limit if there exists a $\mathcal{B}$ such that $(*, \mathcal{B})$ identifies P in the limit.
3. P is identifiable in the limit just in case some $*$ identifies P.
4. $*$ is universal iff $*$ can identify every identifiable P. Else, $*$ is restrictive.

The principal concern:
determine whether the iterated belief revision methods are universal.

1. Possible worlds
2. Problem
3. Environments
4. Scientist
5. Success

Definition

The tail reversal operation on the outcome streams is defined as follows:

$$(e' \ddagger k)(n) = \begin{cases} e'(n) & \text{if } n \leq k \\ \neg e'(n) & \text{otherwise,} \end{cases}$$

where $\neg$ denotes bit reversal.

Definition

The tail reversal operation on the outcome streams is defined as follows:

$$(e' \ddagger k)(n) = \begin{cases} e'(n) & \text{if } n \leq k \\ \neg e'(n) & \text{otherwise,} \end{cases}$$

where $\neg$ denotes bit reversal.

Example

Let z be a stream consisting only of 0s:

$$z = (0000000000000\ldots)$$

Then:

$$e = (z \ddagger 6) \ddagger 11$$

is the following outcome sequence:

Definition

The tail reversal operation on the outcome streams is defined as follows:

$$(e' \ddagger k)(n) = \begin{cases} e'(n) & \text{if } n \leq k \\ \neg e'(n) & \text{otherwise,} \end{cases}$$

where $\neg$ denotes bit reversal.

Example

Let z be a stream consisting only of 0s:

$$z = (0000000000000\ldots)$$

Then:

$$e = (z \ddagger 6) \ddagger 11$$

is the following outcome sequence:

$$?$$

Definition

The tail reversal operation on the outcome streams is defined as follows:

$$(e' \ddagger k)(n) = \begin{cases} e'(n) & \text{if } n \leq k \\ \neg e'(n) & \text{otherwise,} \end{cases}$$

where $\neg$ denotes bit reversal.

Example

Let z be a stream consisting only of 0s:

$$z = (0000000000000\ldots)$$

Then:

$$e = (z \ddagger 6) \ddagger 11$$

is the following outcome sequence:

$$?$$

We can also write: $e = z \ddagger \{6, 11\}$.

The grue is a sinister, lurking presence in the dark places of the earth. Its favorite diet is adventurers, but its insatiable appetite is tempered by its fear of light. No grue has ever been seen by the light of day, and few have survived its fearsome jaws to tell the tale.

- 'Grue' and 'bleen' are artificial predicates.
- Nelson Goodman, "Fact, Fiction, and Forecast".
- "The new riddle of induction".

- 'Grue' and 'bleen' are artificial predicates.
- Nelson Goodman, "Fact, Fiction, and Forecast".
- "The new riddle of induction".

An object is $\begin{cases} \text{grue} & \text{if it is green and examined before time } t, \text{ or else blue.} \\ \text{bleen} & \text{if it is blue and examined before time } t, \text{ or else green.} \end{cases}$

GOODMAN'S PARADOX

All emeralds examined thus far are green. This leads us to conclude (by induction) that also in the future emeralds will be green, and every next green emerald discovered strengthens this belief. Assuming t has yet to pass it is equally true that every emerald that has been observed is grue. Why, then, do we not conclude that emeralds first observed after t will also be grue, and why is the next grue emerald that comes along not considered further evidence in support of that conclusion? The problem is to explain why induction can be used to confirm that things are "green" but not to confirm that things are "grue".

Definition

Define for each $n < \omega$,

$$G^n(z) = \{z \ddagger S \mid card(S) \leq n\}.$$

Then let:

$$G^\omega(z) = \bigcup_{i<\omega} G^i(z).$$

DEFINITION

Define for each $n < \omega$,

$$G^n(z) = \{z \ddagger S \mid card(S) \leq n\}.$$

Then let:

$$G^\omega(z) = \bigcup_{i<\omega} G^i(z).$$

Suppose we want to identify $G^\beta(z)$, where $\beta < \omega$.
Popperian procedure: believe at each stage that the observed tail reversals are the only ones that will ever be observed.

- it identifies $G^\beta(z)$;
- no other method identifying $G^\beta(z)$ dominates it in convergence time;
- no other method identifying $G^\beta(z)$ has a better retraction rate.

Proposition

1. $*_M$, $*_{J,1}$, $*_{A,1}$ *cannot identify* $G^1(z)$.
2. $*_M$, $*_{A,2}$, $*_{R,1}$ *cannot identify* $G^2(z)$.
3. *for all* $n > 0$, $*_{A,n}$ *cannot identify* $G^n(z)$.

Proposition

1. $*_L$, $*_{A,\omega}$, $*_{J,\omega}$, $*_{R,\omega}$ *are universal, and hence can identify* $G^\omega(z)$.
2. $*_{J,2}$, $*_{R,2}$ *can identify* $G^\omega(z)$.

problem	M	A, α	J,α	R,α	L	C
$G^{\omega}(e_0)$	no	$\alpha = \omega$	$\alpha = 2$	$\alpha = 2$	yes	yes
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$G^{n}(e_0)$	no	$\alpha = n+1$	$\alpha = 2$	$\alpha = 2$	yes	yes
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$G^{2}(e_0)$	no	$\alpha = 3$	$\alpha = 2$	$\alpha = 2$	*yes*	*yes*
$G^{1}(e_0)$	no	$\alpha = 2$	$\alpha = 2$	$\alpha = 1$	yes	yes
$G^{0}(e_0)$	*yes*	$\alpha = 0$	$\alpha = 0$	$\alpha = 0$	yes	yes

END OF DAY 4