Bayesian Belief Polarization: Supporting Material

Alan Jern
Department of Psychology
Carnegie Mellon University
ajern@cmu.edu

Kai-min K. Chang
Language Technologies Institute
Carnegie Mellon University
kkchang@cs.cmu.edu

Charles Kemp
Department of Psychology
Carnegie Mellon University
ckemp@cmu.edu

We derive an expression (Equation 3 in the paper) for conditions under which contrary updating occurs; we show why convergence and divergence should be equally frequent in the simulations of Section 5; and we prove that convergence and divergence are not possible for the Bayes net in Figure 1h when V is binary.

1 Derivation of Equation 3 in the paper

In the paper, we point out that contrary updating occurs if

$$[P(h_1|d,b_1) - P(h_1|b_1)]\,[P(h_1|d,b_2) - P(h_1|b_2)] < 0\,. \tag{1}$$

If H is a binary variable, we claim that the condition in Equation 1 is equivalent to

$$[P(d|h_1,b_1) - P(d|h_0,b_1)]\,[P(d|h_1,b_2) - P(d|h_0,b_2)] < 0\,. \tag{2}$$

This claim is true if

$$\mathrm{sgn}(P(h_1|d,B) - P(h_1|B)) = \mathrm{sgn}(P(d|h_1,B) - P(d|h_0,B)) \tag{3}$$

where $\mathrm{sgn}(\cdot)$ is the sign function. We now demonstrate that Equation 3 holds.

Using Bayes' rule,

$$\begin{aligned}\mathrm{sgn}\left(P(h_1|d,B) - P(h_1|B)\right) &= \mathrm{sgn}\left(\frac{P(d|h_1,B)P(h_1|B)}{P(d|B)} - P(h_1|B)\right)\\ &= \mathrm{sgn}\left(\frac{P(h_1|B)}{P(d|B)}\left[P(d|h_1,B) - P(d|B)\right]\right)\end{aligned}$$

We assume that all probabilities are positive and can therefore drop the initial term, giving us

$$\begin{aligned}\mathrm{sgn}(P(d|h_1,B) - P(d|B)) &= \mathrm{sgn}\left(P(d|h_1,B) - \sum_i P(d,h_i|,B)\right)\\ &= \mathrm{sgn}\left(P(d|h_1,B) - \sum_i P(d|h_i,B)P(h_i|B)\right)\\ &= \mathrm{sgn}\left(P(d|h_1,B) - P(d|h_0,B)P(h_0|B) - P(d|h_1,B)P(h_1|B)\right)\\ &= \mathrm{sgn}\left(P(d|h_1,B)P(h_0|B) - P(d|h_0,B)P(h_0|B)\right)\end{aligned}$$

where we use the fact that $P(h_0|B) = 1 - P(h_1|B)$ in our binary hypothesis space. Factoring out $P(h_0|B)$ and dropping it because it is positive leaves us with Equation 3.

Since the paper only considers cases in which both people change their beliefs, it follows that both factors in Equation 2 are non-zero. As mentioned in the paper, Equation 2 does not hold and contrary updating cannot occur if $D \perp B|H$, in which case the expression on the left simplifies to $(P(d|h_1) - P(d|h_0))^2$, which is always positive.

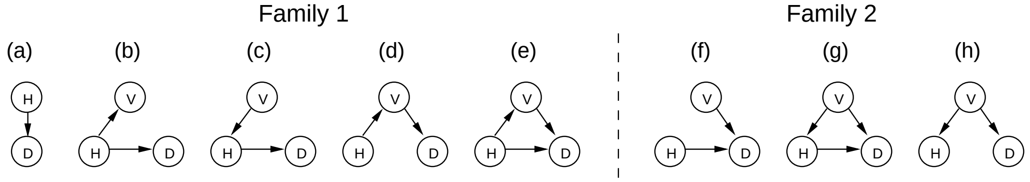
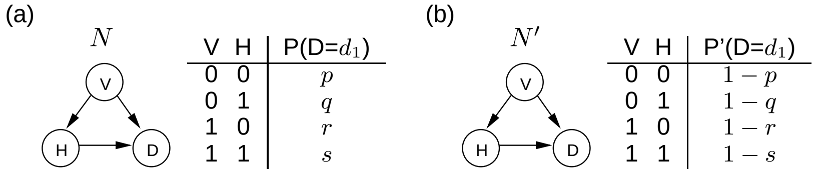


Figure 1: The set of Bayes nets considered in the paper.



V	H	P(D=d_1)
0	0	p
0	1	q
1	0	r
1	1	s

V	H	P'(D=d_1)
0	0	$1-p$
0	1	$1-q$
1	0	$1-r$
1	1	$1-s$

Figure 2: An illustration of how network N in (a) can be transformed into network N' in (b) by altering the CPD for node D.

2 Proof that convergence and divergence are equally frequent

For the simulations in the paper, we claimed that convergence and divergence should be equally common. This section explains this result.

Suppose that we are working with a Bayes net N that belongs to Family 2 where both H and D are binary variables. Suppose also that network N produces belief divergence. Without loss of generality, we assume that person 1 begins with the stronger belief in h_1 (i.e. $P(h_1|b_1) > P(h_1|b_2)$). Since the network produces divergence,

$$P(h_1|d_1, b_1) - P(h_1|b_1) > 0 \tag{4}$$

where d_1 represents the observed data.

We now create a second Bayes net N' that is closely related to N but that produces convergence rather than divergence. N' captures a joint distribution $P'(\cdot)$ over the variables in the network, and we choose the CPDs in N' such that $P'(H|B) = P(H|B)$ but $P'(d_1|H, B) = 1 - P(d_1|H, B)$. If these conditions are satisfied then

$$\begin{aligned} P'(d_1|h_1, b_1) - P'(d_1|h_0, b_1) &= 1 - P(d_1|h_1, b_1) - (1 - P(d_1|h_0, b_1)) \qquad (5) \\ &= -(P(d_1|h_1, b_1) - P(d_1|h_0, b_1)) \qquad (6) \end{aligned}$$

We showed in Section 1 that $P'(h_1|d_1, b_1) - P'(h_1|b_1)$ and $P'(d_1|h_1, b_1) - P'(d_1|h_0, b_1)$ must have the same sign. Equations 4 and 6 therefore imply that person 1's belief increases in the case of network N but decreases in the case of network N'. Similarly, person 2's beliefs are updated in opposite directions in these two cases. Note, however, that the initial beliefs of both people are the same in both cases (e.g. $P'(h_1|b_1) = P(h_1|b_1)$). It follows that the divergence in network N has been transformed into convergence in network N'.

It remains to be described how the CPDs in network N can be altered to produce network N'. Figure 2 illustrates how this can be achieved for a network N that is an instance of the Bayes net in Figure 1g. In effect, the transformation relabels the two possible values of D, converting d_1 into d_0 and vice versa. Since D is always a child node, it does not appear among the parents of H, meaning that the transformation does not alter the distribution $P(H|B)$. As required, however, the transformation ensures that $P'(d_1|H, B) = P(d_0|H, B) = 1 - P(d_1|H, B)$.

In each trial of our simulations, the rows in each CPD were sampled independently from a Beta distribution. It follows that the CPDs in Figures 2a and 2b must be equally likely. In other words, for every set of CPDs that produces divergence, there is an equally likely set of CPDs that produces convergence, and vice versa.

3 Proof that Bayes net 1h with a binary V node cannot produce contrary updating

Consider the network in Figure 1h, and suppose that V is a binary variable. We now compute the prior and posterior probabilities of h_1 for person 1.

$$\begin{aligned} P(h_1|b_1) &= \sum_i P(h_1|v_i)P(v_i|b_1) \\ &= P(h_1|v_0)P(v_0|b_1) + P(h_1|v_1)P(v_1|b_1) \\ &= P(h_1|v_0)(1 - P(v_1|b_1)) + P(h_1|v_1)P(v_1|b_1) \end{aligned}$$

Similarly,

$$P(h_1|d, b_1) = P(h_1|v_0)(1 - P(v_1|d, b_1)) + P(h_1|v_1)P(v_1|d, b_1)$$

It follows that

$$P(h_1|d, b_1) - P(h_1|b_1) = [P(v_1|d, b_1) - P(v_1|b_1)][P(h_1|v_1) - P(h_1|v_0)]$$

Similarly, for person 2,

$$P(h_1|d, b_2) - P(h_1|b_2) = [P(v_1|d, b_2) - P(v_1|b_2)][P(h_1|v_1) - P(h_1|v_0)]$$

From Equation 2 in the paper, we know that if the product of these two differences is less than zero, then contrary updating occurs.

$$\begin{aligned} &[P(h_1|d, b_1) - P(h_1|b_1)][P(h_1|d, b_2) - P(h_1|b_2)] = \\ &\qquad (P(h_1|v_1) - P(h_1|v_0))^2 (P(v_1|d, b_1) - P(v_1|b_1))(P(v_1|d, b_2) - P(v_1|b_2)) \end{aligned}$$

The expression on the right is the product of a term that is always positive and a term that is equivalent to the left side of Equation 2 in the paper, with variable V taking the place of variable H. As we showed in Section 1 of this supporting material, this expression cannot be less than zero if D is conditionally independent of B given H (V in this case). This conditional independence exists for Bayes net 1h. Therefore, it is not capable of producing contrary updating.