

# Nielson and Stewart, Bayesian polarization

Kevin Dorst

24.223, Rationality

**TOTAL:** In ideal evidential scenarios (when evidence is clear and shared), ideally rational (Bayesian) agents expected to converge in opinions.

→ N&S claim that **TOTAL** is presupposed in many popular and social-scientific discussions of (e.g.) political disagreements.

→ But N&S claim that **TOTAL** is false: simple examples show that in any case of finite learning, polarization can be ideally rational.

And more subtle reasoning shows that even in cases of *infinite* and *complete* evidence, polarization is still possible.

## I. Local Polarization

Let  $P$  be my (ideally rational) credence function and  $Q$  be yours.

Assume probabilistic, obey ratio formula, and update by conditioning.

$P$  and  $Q$  *locally* polarize on a given proposition  $A$  upon learning  $E$  iff

$$P(A|E) < P(A) \leq Q(A) < Q(A|E)$$

This can (obviously) happen!

**Election.** Abby and Bill are Democrats facing off in a primary; Christa and Dan and Republicans facing off in a primary. We know only one of each pair will win their primaries, and only one of the four will win the general election. I think Bill is the stronger Democrat; you think that Abby is. Precisely:

	$a$	$b$	$c$	$d$
$P:$	1/6	1/4	1/3	1/4
$Q:$	1/2	1/12	1/4	1/6

As a result, learning that Abby and Christa won their primaries ( $\{a, c\}$ ) makes me lower my credence that a Democrat will win ( $\{a, b\}$ ), and you raise your credence that a Democrat will win. Where  $E = \{a, c\}$ :

	$a$	$b$	$c$	$d$
$P(\cdot E):$	1/3	0	2/3	0
$Q(\cdot E):$	2/3	0	1/3	0

$P(\{a, b\}) \approx 0.42$  and  $Q(\{a, b\}) \approx 0.58$ ,  
yet  $P(\{a, b\}|E) \approx 0.33$  and  
 $Q(\{a, b\}|E) \approx 0.66$ .

In general, whether  $E$  polarizes  $P$  and  $Q$  on  $A$  depends on whether  $P$  and  $Q$  disagree on the likelihood ratios:

**Thm.** if  $0 < P(A) \leq Q(A) < 1$ , then  $E$  polarizes  $P$  and  $Q$  iff

$$\frac{P(E|A)}{P(E|\neg A)} < 1 < \frac{Q(E|A)}{Q(E|\neg A)}$$

"The proof of this result uses only the probability axioms and algebra. We omit it, assured the reader can furnish it herself should she so desire." Lol.

**Upshot:** No reason to expect learning the same evidence to reduce disagreement.

And since learning any finite stream of evidence is equivalent to learning a big conjunction, no reason to expect any finite stream of evidence to reduce rational disagreement.

→ So, say N&S, there's little reason to expect that increasing evidence will lead rational people to converge in local opinions.

**Q:** Is this a good argument?

## II. Global Polarization

But come to agree on  $E$ ! Global disagreement?

We can measure the overall disagreement between  $P$  and  $Q$  using their **total variational distance**, i.e. the maximum degree to which they disagree about any proposition.

$$d(P, Q) = \max_{A \subseteq W} |P(A) - Q(A)|$$

Is this a good measure of overall disagreement?

- When  $P$  and  $Q$  agree on everything,  $d(P, Q) = 0$ .
- When  $P$  and  $Q$  disagree maximally on something,  $d(P, Q) = 1$ .

In particular, the areas they assign positive credence to are disjoint.

Increasing  $d$  can also be perfectly rational.

E.g. now we agree that Abby's stronger than Bill and that Christa is stronger than Dan, but we disagree on how much stronger:

	$a$	$b$	$c$	$d$
$P$ :	1/4	1/8	1/2	1/8
$Q$ :	1/2	1/12	1/4	1/6
$P(\cdot E)$ :	1/3	0	2/3	0
$Q(\cdot E)$ :	2/3	0	1/3	0

**Q:** Is this a good argument?

## III. Infinite disagreement

Consider an infinite set of refined partitions  $\{\mathcal{E}_n\}_{n \in \mathbb{N}}$ . E.g. initial-segments of an infinite coin toss. Call this *increasing* evidence. Update by conditioning.

Suppose the question  $Q$  the agent is interested in is generated by the filtration, so that the evidence is *increasing and complete* wrt  $Q$ .

So 'it lands heads at least  $k$  times' or 'the  $4238^{922123}$ th toss lands heads' are about  $Q$ . But (importantly!) so are events the agent never observes, like 'the long-run relative frequency of heads is  $\frac{1}{2}$ .'

**Convergence to the truth theorem:** In this setup, the prior  $P$  assigns probability 1 to the event that her posteriors will get arbitrarily close to the truth of every proposition about  $Q$ .

$P$  shares evidence with  $Q$  if: for all  $E \in \mathcal{E}_n$ , if  $P(E) > 0$ ,  $Q(E) > 0$ .

$P$  is *absolutely continuous* wrt  $Q$  if  $Q(A) = 0$  implies  $P(A) = 0$ .  $P$  merges with  $Q$  if  $d(P_n, Q_n) \rightarrow 0$  as  $n \rightarrow \infty$ .

**Merging of opinions theorem:** In this setup,  $P$  assigns probability 1 to merging with  $Q$ .

Let  $H$  = the set of worlds that  $P$  assigns higher probability to than  $Q$  =  $\{w : P(w) > Q(w)\}$ .

Then  $d(P, Q) = P(H) - Q(H)$ .

Easier calculation:  $\frac{1}{2} \sum_w |P(w) - Q(w)|$

$$\begin{aligned} d(P, Q) &= P(\{b, c\}) - Q(\{b, c\}) \\ &= \frac{5}{8} - \frac{4}{12} = 7/24 \approx 0.29, \\ \text{while } d(P(\cdot|E), Q(\cdot|E)) & \\ &= P(\{c\}|E) - Q(\{c\}|E) = \frac{1}{3} \approx 0.33 \end{aligned}$$

A *filtration*.

$Q$  is the smallest sigma-algebra containing  $\bigcup_{n=1}^{\infty} \mathcal{E}_n$ .

Obvious for finitely-settled claims. Surprising for infinite ones.

Gaps:

- If not mutually absolutely continuous, needn't converge. ("Consensus or polarization law")
- Says nothing about events outside the evidence-generated algebra