Comments on Carl Wagner's *Jeffrey Conditioning and External Bayesianity*

Stephen Petersen
Niagara University
steve@stevepetersen.net

May 1, 2008*

In "Jeffrey conditioning and external Bayesianity", Carl Wagner provides further support from mathematical elegance for what he calls the *uniformity rule*: namely, that Bayesians should represent "identical learning" by sameness of odds ratios across atomic events. Put another way, if something prompts both you and me to change our subjective probabilities, and as a result we *learn* the *same* thing, then it must be that for each atomic event, our subjective odds have expanded or shrunk by the same factor.

In particular, Wagner argues, suppose n of us have our own probability mass functions (pmfs) over some probability space, and we are seeking to *pool* this n-tuple of pmfs into one (as a compromise among our views, or summary of them, or whatever). Suppose further that as we are about to pool our pmfs, we all share the same prompt for probability revision. Should we pool our pmfs first, and then revise based on this new learning, or vice-versa? Intuitively such a question should be the least of our concerns. That is, we should prefer a pooling operator where this order doesn't matter; our pooling should commute with the probability revision. Wagner shows our chosen pooling operator *will* commute with some likelihood revision if and only if it commutes with Jeffrey conditioning, *when parameterized as the uniformity rule would dictate*.

The math looks good to me. (Whether because it *is* good, or because I'm not an adequate judge, I leave to you.) In my comments on the paper I'll instead focus on providing context for this result, first by summarizing some of the other mathematical elegance Wagner has adduced in favor of his uniformity rule, and then by sketching the philosophical motivations and issues behind the mathematics.

1 Mathematical considerations

Ensuring external Bayesianity is just one of the advantages behind the uniformity rule. The key result, which goes back to Field (1978), is that when the "input factor" of probability revision is represented by the ratios of new to old evidential odds, then Jeffrey conditioning will be commutative; learning one thing and then another will not

*This is a *draft* (version: *few-wagner-comments.tex,v1.3*), and should not be circulated or cited without permission.

result in a different final pmf than learning the same things in the reverse order. This is emphatically *not* the case when "identical learning" is taken to mean coming to the same posterior probabilities for evidential events. Thus the uniformity rule rescues the plausibility of Jeffrey conditioning, which in turn allows for revising based on uncertain evidence, and thus also for the possibility of *reversing* one's learning should such evidence later be undermined.

Wagner modestly fails to emphasize the extensive work he has already done building on this main idea. Wagner (2002) extended Field's result from finite samples spaces to infinite ones with countable partitions, and Wagner (2003) ties this result more explicitly to his uniformity rule. That paper also argues (in conjunction with other papers of his) that the uniformity rule can help solve the historical old evidence problem—the problem of how to allow a hypothesis to gain credence not because of new evidence, but because we newly learn that it implies old evidence. Wagner shows that Jeffrey's suggestion to reconstruct a pre-evidential "ur-distribution" can be modified to allow for uncertain revision, *if* we do it *à la* the uniformity rule (this time changing new *conditional* probabilities with identical odds ratios). Thus we needn't be concerned whether we revise the uncertain evidence or the explanatory relation first. All along Wagner has shown that rival representations of identical learning, such as relevance quotients or probability differences, will not give the same results as neatly.

One final mathematical consideration to which Wagner has pointed me is worth special note. Chan and Darwiche (2002) have suggested the following nice metric for probability measures on a finite sample space Ω:

$$\mathrm{CD}(p,q) = \log \max_{\omega\in\Omega} \frac{q(\omega)}{p(\omega)} - \log \min_{\omega\in\Omega} \frac{q(\omega)}{p(\omega)}$$

This metric is closely related to the odds ratios $\beta_{q,p}(A:B)$ between events $A,B \in \mathscr{P}(\Omega)$ in the revision from p to q:

$$\beta_{q,p}(A:B) = \frac{q(A)/q(B)}{p(A)/p(B)}$$

The definition of $\mathrm{CD}(p,q)$ above gives

$$\begin{aligned}
\mathrm{CD}(p,q) &= \log \frac{\max_{\omega\in\Omega} q(\omega)/p(\omega)}{\min_{\omega'\in\Omega} q(\omega')/p(\omega')} \\
&= \log \max_{\omega,\omega'\in\Omega} \frac{q(\omega)/p(\omega)}{q(\omega')/p(\omega')} \\
&= \max_{\omega,\omega'\in\Omega} \log \frac{q(\omega)/p(\omega)}{q(\omega')/p(\omega')} \\
&= \max_{\omega,\omega'\in\Omega} \log \frac{q(\omega)/q(\omega')}{p(\omega)/p(\omega')} \\
&= \max_{\omega,\omega'\in\Omega} \log \beta_{q,p}(\{\omega\}:\{\omega'\}) \\
&= \max_{A,B\in\mathscr{P}(\Omega)-\emptyset} \log \beta_{q,p}(A:B)
\end{aligned}$$

(This last step is not trivial, but not hard.) As a result, when identical learning is taken to be sameness of odds ratios, any two probability measures undergoing identical learning will move the same CD-distance. This fails to hold of relative entropy measures like the Kullback-Leibler divergence (which is not properly a metric anyway).[1]

2 Philosophical hesitations

As Wagner is well aware, there remains a reasonable philosophical dispute over the notion of identical learning and the attendant appropriateness of Jeffrey conditioning—despite this pile of mathematical elegance. I can give here only an overview of the issue.

Let's start with the problem that Daniel Garber raised for Field's initial proposal of using odds ratios for learning. Garber (1980) points out that learning by odds ratios cannot be a way to represent learning by *sensory experience*, since otherwise repeating the same uncertain sensory experience over and over will, by repeatedly applying the same odds ratios, drive your probabilities toward certainty. Wagner (2002) holds that we should therefore divorce identical learning from sense experiences; "we learn nothing new from repeated glances and so all [odds ratios] beyond the first are equal to one" (p. 276).

This notion brings counterintuitive results, though, it seems—as an example from Döring (1999) points out. Suppose I give a very high prior probability that some shirt is blue, and you give a very low one. We then get an identical glimpse at the shirt under a neon light, and it looks the same shade of bluish-green to both of us. My posterior for the shirt's blueness should be lower, while yours should be higher. But then, since our odds ratios clearly differ, we did not (by the uniformity rule) undergo identical learning. In a sense perhaps this is right, but in another sense—one naturally wedded to our similarity of sensory experience—this seems clearly wrong.

Part of the temptation of odds ratios is in its factoring out of differing priors, measuring only the change both distributions undergo when learning. But these examples seem to show that in an important sense priors are quite relevant to whether one undergoes "identical learning" *à la* the uniformity rule. Put metaphorically, one way to neglect your starting place is to measure only how far you moved; another is to look only at the force that's pushing you. These can come apart—if, for example, you can weigh different amounts. Field evidently hoped that factoring out priors in the former sense (that of the uniformity rule) was the same as factoring out priors in the latter sense (that of looking only at physical sensory stimuli, independent of background beliefs). This hope, it seems, is fruitless.

Sometimes, on the other hand, tying a learning experience to priors works in our favor, and may even be a plausible way to read sensory experience. Consider this worry from Brian Skyrms about updating on uncertain evidence: if I catch a dim fleeting

[1] Wagner notes that, historically, Alan Turing, I. J. Good, and their Bletchley Park gang used exactly $\log\beta(A:B)$ as a measure of plausibility—or "weight of evidence"—gained in one hypothesis over another as a result of new learning. They used $\log_{10}$, and called their units "bans" for cute reasons (preferring to speak in "decibans"), but of course it would be just as appropriate to take the $\log_2$, resulting in the information unit of "bits" later preferred by Claude Shannon.

glimpse of a crow, and thereby assign it a relatively low probability of its being black, then it seems by updating on this uncertainty I can thereby disconfirm my hypothesis that all crows are black. In general, Skyrms says, "I could disconfirm lots of theories just by running around at night" (quoted in Lange (2000), p. 397). Marc Lange suggests that if "the raven looks about the way that any dusky colored object would be expected to look under those conditions" (p. 397), then we should perhaps instead think of this sensory experience as only slightly inflating the prior odds that the particular crow would be black—as opposed to the comparably large odds ratio change resulting from getting a good look. If so, the uniformity rule looks appropriate.

Here is a similar example from Daniel Osherson, in Osherson (2002). Suppose some experience of a patch of sky brings my subjective probability for rain from .5 to .7. Now suppose that before this glimpse I first hear a weather report that brings my probability for rain down to .3. If after looking at the sky I still go to .7, Osherson says I must thereby have had a *different* (more rain-like) sensory experience. Here Osherson seems insistent on measuring sensory experience by something like odds ratios.

It is also not obvious that we should seek commutativity when updating on uncertain evidence. In an example from Roger Rosencrantz,

> Consider a child who has just knocked over a jar of paint and is wondering whether he is going to get spanked. In one scenario, a parental scowl is followed by good natured laughing, while, in the other, these responses occur in the opposite sequence! (Rosenkrantz (1981), 3.6-2, as quoted in Lange (2000) p. 396.)

Lange dismisses this case—since it's classical Bayesian conditioning, he says, rather than revision on uncertain evidence, of course it will commute; it only appears not to because they are *not* the same pieces of evidence received in different order. One piece of evidence is a scowl transforming into laughter, and the other piece of evidence a laugh transforming to a scowl.

This answer strikes me as overly simple, though; first, it's not obvious this is a case of classical conditioning; perhaps the spanking probability revision is based on the probability that the parent is angry, for which facial expressions give uncertain evidence. Second, assuming that a video of one such parental transformation could be the reverse video of the other, then clearly they could be seen as the same *sensory* experiences in a different order—think of the stills at 30 FPS or higher. The motivation for calling the differently-ordered conjunctions of such impressions different experiences seems mostly to come from the fact that when the conjunction appears in one order, it nudges posterior spanking probability in a different direction than when taken in the opposite order. This can be done of course; Skyrms showed we can always get the results of Jeffrey conditioning by enriching the sample space enough to do classical Bayesian conditioning. The question is one of how plausible it is to admit such points into our sample space.

In many cases of course it seems clear we do want to maintain commutativity, and without so enriching the sample space. Take for example the Döring case in which my final estimated probability of where an explosion originated in an airplane depends on the order in which I notice the rows of seats left intact (p. S383). Here it seems clear we should *not* say "seeing row of seats x then y is a different piece of evidence from seeing row *y* then *x*." Instead we'd better darn well hope that these individual pieces of evidence commute with our revisions under uncertainty.

In the end, it may simply be that sometimes the order of sensory experience matters (as in the parents' reaction to spilling paint) and sometimes it doesn't (as in the airplane seats); sometimes by "learning the same" we mean "coming to attribute the same posterior probability to observation sentences" (as when we see the cloth in neon light) and sometimes we mean "having our probability distributions disturbed the same amount" (as by the crows). I suspect the problem is in the variability in specifying the sample space, and that such *ad hoc*kery will continue to haunt us at least until we manage to revive a protocol-sentence-like notion of observation that is independent of background theory. (I further suspect such a notion cannot be revived.)

References

Chan, H. and Darwiche, A. (2002). A distance measure for bounding probabilistic belief change. In *Eighteenth National Conference on Artificial Intelligence*, pages 539–545, Menlo Park, CA. American Association for Artificial Intelligence.

Döring, F. (1999). Why Bayesian psychology is incomplete. *Philosophy of Science*, 66 (Proceedings):S379–S389.

Field, H. (1978). A note on Jeffrey conditionalization. *Philosophy of Science*, 45(3):361–367.

Garber, D. (1980). Field and Jeffrey conditionalization. *Philosophy of Science*, 47(1):142–145.

Lange, M. (2000). Is Jeffrey conditionalization defective by virtue of being noncommutative? Remarks on the sameness of sensory experiences. *Synthese*, 123:393–403.

Osherson, D. (2002). Order dependence and Jeffrey conditionalization. `http://www.princeton.edu/~osherson/papers/jeff3.pdf`.

Rosenkrantz, R. (1981). *Foundations and Applications of Inductive Probability*. Ridgeview Press, Atascadero, CA.

Wagner, C. G. (2002). Probability kinematics and commutativity. *Philosophy of Science*, 69:266–278.

Wagner, C. G. (2003). Commuting probability revisions: The uniformity rule. *Erkenntnis*, 59:349–364.