A Natural Generalization of Jeffrey Conditioning

Standard conditioning in Bayesian probability theory gives us a relatively well-accepted tool to update on the observation of an event. Jeffrey conditioning provides another tool which updates probability distributions given uncertain evidence. Jeffrey conditioning generalizes standard conditioning. Evidence can be viewed as imposing a constraint on acceptable probability distributions, often one with which the prior probability distribution is inconsistent. If it is a conditional which constitutes this constraint, standard conditioning and Jeffrey conditioning do not always apply. Carl Wagner presents such a case (see Wagner, 1992) together with a solution based on a plausible intuition. We will call this intuition (W). Wagner's (W) solution, or Wagner conditioning, in its turn generalizes Jeffrey conditioning.

Twenty years earlier, E.T. Jaynes had already proposed a generalization of Jeffrey conditioning, the principle of maximum entropy (M). This generalization was more sweeping than Wagner's and included partial information cases (using the moments of a distribution as evidence). Even though there was a plausible intuition at work as well, (M) soon ran into counter-examples and conceptual difficulties. The question for Wagner is therefore whether his generalization (W) agrees with (M) or not. Wagner notes that it does not. Wagner then uses his method not only to present a "natural generalization of Jeffrey conditioning" (see Wagner, 1992, 250), but also to deepen criticism of (M).

I will show, by contrast, that (M) not only generalizes Jeffrey conditioning (as is well known, for a formal proof see Caticha and Giffin, 2006) but also Wagner conditioning. Wagner's intuition (W) is plausible, and his method works. His derivation of a disagreement with (M), however, is conceptually more complex than he assumes. Below, I will show that (M) and (W) are consistent given (L). (L) is what I call the Laplacean principle which requires a rational agent, besides other standard Bayesian commitments, to hold sharp credences with respect to well-defined events under consideration. (I), which is inconsistent with (L) and which some Bayesians accept, allows a rational agent to have indeterminate or imprecise credences (see Ellsberg, 1961; Levi, 1985; Walley, 1991; and Joyce, 2010).

(M)	(W)	(I)	(L)		
•	•			×	according to Wagner's article
•	•			✓	according to this article
		•	•	×	disagree over indeterminate credences
•	•	•		×	formally shown in Wagner's article
•	•		•	✓	formally shown in this article

Before we generalize Wagner conditioning by using (M), let us articulate (L) and (M). (L) is what I call the Laplacean principle and in addition to standard Bayesian commitments states that a rational agent assigns a determinate precise probability to a well-defined event under consideration (for a defence of (L) against (I) see White, 2010; and Elga, 2010).

To avoid excessive apriorism (see Seidenfeld, 1979), (L) does not require that a rational agent has probabilities assigned to all events in an event space, only that, once an event has been brought to attention, and sometimes retrospectively, the rational agent assigns a sharp probability. (L) also does not require objectivity in the sense that all rational agents must agree in their probability distributions if they have the same information.

It is important to distinguish between type I and type II prior probabilities. The former precede any information at all (so-called ignorance priors). The latter are simply prior relative to posterior probabilities in probability kinematics. They may themselves be posterior probabilities with respect to an earlier instance of probability kinematics. Once we agree on a prior distribution (type II) and on a set of formal constraints representing our evidence, (M) claims that posterior probabilities follow mechanically. To standard Bayesian commitments and (L), (M) adds

> Update type II prior distributions under formalized constraints in accordance with information theory and a commitment to keep the entropy maximal,

> if constraints are synchronic, and the cross-entropy minimal, if they are diachronic.

This corresponds to the intuition that we ought not to gain information where the additional information is not warranted by the evidence. Some want to drive a wedge between the synchronic rule to keep the entropy maximal (MAXENT) and the diachronic rule to keep the cross-entropy minimal (*Infomin*) (for this objection see Walley, 1991, 270f). I will dispel this worry elsewhere and assume here that MAXENT and *Infomin* are compatible.

Wagner claims that he has found a relatively common case of probability kinematics in which (M) delivers the wrong result so that we must develop an ad hoc generalization of Jeffrey conditioning. The example involves a linguist who encounters a native and tries to determine on one utterance whether the native is Catholic, Protestant, a northerner, or a southerner (see Wagner, 1992, 252 and Spohn, 2012, 197). I am in agreement with Wagner about his solution, but his scathing verdict about (M) (see Wagner, 1992, 255) is not really a verdict about (M) in the Laplacean tradition but about the curious conjunction of (M) and (I). Let us look at the contrast between what Wagner considers to be the solution of (M) for a *Simplified Linguist* problem, 'Wagner's (M) solution,' and Wagner's own, much more plausible solution, 'Wagner's (W) solution.' I will show why Wagner's (M) solution misrepresents (M).

> The *Simplified Linguist* problem. Imagine the native from Wagner's *Linguist* example is either Buddhist or Catholic (50:50). Further imagine that the utterance of the native to the linguist either entails that the native is a Buddhist (60%) or provides no information about the religious affiliation of the native (40%).

Wagner's (W) solution (and, surely, the correct solution) is the posterior probability distribution 80:20. Wagner's (M) solution for this radically simplified problem is 60:40, clearly a more entropic solution than Wagner's (W) solution. From the perspective of an (M) advocate, there are only two explanations for this difference in cross-entropy. Either Wagner's (W) solution illegitimately uses information not contained in the problem, or Wagner's (M) solution has failed to include information that is contained in the problem. The problem is that Wagner's (M) solution does not take into account (L). Therefore, the latter is the case.

For a Laplacean, the prior joint probability distribution is not left unspecified for the calculation of the posteriors. Before the native makes the utterance, the event space is unspecified. After the utterance, however, the epistemically accessible event space is populated by prior probabilities according to (L). The following is a distribution matrix for which the last row is the sum of the previous rows, the last column is the sum of the previous columns, and all matrix elements not in the sum rows or sum columns add up to 1. p_{bx}, for example, is the type II prior probability that the native is a Buddhist and that her utterance is x (the one which entails that she is a Buddhist). q_{bx} is the posterior probability. If the native's utterance is y, then the linguist has no information about her religious identity. z is the event where the native says neither x nor y, which as a matter of prior probability is highly likely and is excluded by the evidence so that $q_x = 0$.

$$\begin{pmatrix} p_{bx} & p_{cx} & p_x \\ p_{by} & p_{cy} & p_y \\ p_{bz} & p_{cz} & p_z \\ p_b & p_c & 1.00 \end{pmatrix} \tag{1}$$

Following (L) we shall populate the joint probability matrix by MAXENT (using the marginals) and update it by *Infomin*. For the *Simplified Linguist* problem, this procedure gives us the correct result, agreeing with Wagner's (W) solution (80:20).

$$\begin{pmatrix} q_{bx} = 0.60 & q_{cx} = 0.00 & q_x = 0.60 \\ q_{by} = 0.20 & q_{cy} = 0.20 & q_y = 0.40 \\ q_{bz} = 0.00 & q_{cz} = 0.00 & q_z = 0.00 \\ q_b = 0.80 & q_c = 0.20 & 1.00 \end{pmatrix} \tag{2}$$

More detailed calculations reveal that it also gives us the correct result for Wagner's more involved *Linguist* problem. The mathematically detailed general proof is complex and connected to an inversion of a problem presented by Vladimír Majerník (see Majerník, 2000; and Lukits, 2015). These calculations use the Kullback-Leibler divergence from information theory, not Wagner's (and Jeffrey's) intuition about maintaining ratios that are unaffected by evidence.

It turns out that (M) agrees with (W) on all cases where (W) is applicable. (M) is applicable to a wider range of cases, so that (M) can be said to generalize (W). To get there, we have assumed (L): the joint probability matrices are populated by determinate probabilities. Wagner disagrees with (L) and follows Peter Walley's recommendation in *Statistical Reasoning with Imprecise Probabilities* that marginals do not determine a unique product (see Walley, 1991, 456). Consequently, Wagner's prior probability matrix looks like this:

$$\begin{pmatrix} ? & ? & p_x \\ ? & ? & p_y \\ ? & ? & p_z \\ p_b & p_c & 1.00 \end{pmatrix} \tag{3}$$

This prior probability matrix leads to Wagner's (M) solution, which is clearly implausible:

$$\begin{pmatrix} ? & ? & q_x = 0.60 \\ ? & ? & q_y = 0.40 \\ ? & ? & q_z = 0.00 \\ q_b = 0.60 & q_c = 0.40 & 1.00 \end{pmatrix} \tag{4}$$

While Wagner is welcome to deny (L), my sense is that in general advocates of (M) accept it. Wagner's result makes clear that not to do so would be inconsistent. Wagner's result, however, does not show that (M) on its own is undermined without independent arguments that defend the hidden assumption (I) or impugn (L). The formal proof of the claims outlined in this paper shows that (M) generalizes (W), given (L), so that Wagner's intuition is plausible but unnecessary and can be integrated in the more basic and more far-reaching intuitions we have about information gain in updating.