Abstract

When we come to know a conditional, we cannot straightforwardly apply Jeffrey conditioning to gain an updated probability distribution. Carl Wagner has proposed a natural generalization of Jeffrey conditioning to accommodate this case (Wagner conditioning). The generalization rests on an ad hoc but plausible intuition (W). Wagner shows how the principle of maximum entropy (M) disagrees with intuition (W) and therefore considers (M) to be undermined. This article presents a natural generalization of Wagner conditioning which is derived from (M) and implied by it. (M) is therefore not only consistent with (W), it seamlessly and elegantly generalizes it (just as it generalizes standard conditioning and Jeffrey conditioning). Wagner's inconsistency result for (W) and (M) is instructive. It rests on the assumption (I) that the credences of a rational agent may be indeterminate. While many Bayesians now hold (I) it is difficult to articulate (M) on its basis because to date there is no proposal how to measure indeterminate probability distributions in terms of information theory. Most, if not all, advocates of (M) resist (I). If they did not they would be vulnerable to Wagner's inconsistency result. Wagner has therefore not, as he believes, undermined (M) but only demonstrated that advocates of (M) must accept that rational agents ought to have sharp credences.

1 Introduction

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 (or densities, from now on omitted) 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 is more sweeping than Wagner's and includes partial information cases (using the moment(s) of a distribution as evidence), such as Bas van

Fraassen's *Judy Benjamin* problem and Jaynes' own *Brandeis Dice* problem. It uses information theory to suggest that one should (a) always choose prior probabilities which are minimally informative, and (b) update to the probability distribution which is minimally informative relative to the prior probability distribution while obeying the constraints imposed by the observation or the evidence. Again, there is a plausible intuition at work, but (M) soon ran into counter-examples (e.g. *Judy Benjamin*, see van Fraassen, 1981) and conceptual difficulties (e.g. Abner Shimony's Lagrange multiplier problem, see Friedman and Shimony 1972; or more recently, Joseph Halpern's and Peter Grünwald's coarsening at random, 2003).

The question for Wagner was therefore whether his generalization (W) agreed with (M) or not. Wagner found that it did not. Wagner then used 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 that (M) not only generalizes Jeffrey conditioning (as is well known, for a formal proof see Caticha aand 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 assumed. Below, we 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 permitting mushy credences
•	•	•		×	formally shown in Wagner's article
•	•		•	✓	formally shown in this article

While Wagner is welcome to deny (L), my sense is that advocates of (M) usually accept it because they are already to the right of Sandy L. Zabell's spectrum between left-wing dadaists and right-wing totalitarians (see Zabell 2005; Zabell's representative of right-wing totalitarianism is E.T. Jaynes). If there were an advocate of (M) sympathetic to (I), Wagner's result would indeed force her to choose, but my emphasis is that Wagner's criticism of

(M) is misplaced since it rests on an assumption that someone who believes in (M) would naturally not hold. Wagner certainly does not give independent arguments for (I). This paper shows how elegantly (M) generalizes not only standard conditioning and Jeffrey conditioning but also Wagner conditioning, once we accept (L).

A tempered and differentiated account of (M) (contrasted with Jaynes' earlier version) is not only largely immune to criticisms, but often illuminates the problems that the criticisms pose (for example in the *Judy Benjamin* case see Lukits 2014). This account rests on principles such as (L) and a reasonable interpretation of what we mean by objectivity. To make this more clear, and before we launch into the formalities of generalizing 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 is able to assign a probability. Newton did not need to have a prior probability for Einstein's theory in order to have a posterior probability for his theory of gravity.

(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. One of Jaynes' projects, the project of objectivity for type I prior probabilities, has failed.

The case for objectivity in probability kinematics, where prior probabilities are of type II, is consistent with and dependent on a subjectivist interpretation of probabilities, making for some terminological confusion. Interpretations of the evidence and how it is to be cast in terms of formal constraints may vary. Once we agree on a prior distribution (type II), however, and on a set of formal constraints representing our evidence, (M) claims that posterior probabilities follow mechanically. Just as is the case in deductive logic,

we may come to a tentative and voluntary agreement on an interpretation, a set of rules and presuppositions and then go part of the way together. 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 simple 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*).

Here is a brief excursion to dispel this worry. Consider a bag with blue, red, and green tokens. You know that (C') at least 50% of the tokens are blue. Then you learn that (C'') at most 20% of the tokens are red. The synchronic norm MAXENT, on the one hand, ignores the diachronic dimension and prescribes the probability distribution which has the maximum entropy and obeys both (C') and (C''). The three-dimensional vector containing the probabilities for blue, red, and green is $(\frac{1}{2}, \frac{1}{5}, \frac{3}{10})$. *Infomin*, on the other hand, processes (C') and (C'') sequentially, taking in its second step $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ as its prior probability distribution and then diachronically updating to $(\frac{8}{15}, \frac{1}{5}, \frac{4}{15})$.

The information provided in a problem calling for MAXENT and the information provided in a problem calling for *Infomin* is different, as temporal relations and their implications for dependence between variables clearly matter. In the above case, we might have relevantly received information (C'') before (C') ('before' may be understood logically rather than temporally) so that *Infomin* updates in its last step $(\frac{2}{5}, \frac{1}{5}, \frac{2}{5})$ to $(\frac{1}{2}, \frac{1}{6}, \frac{1}{3})$. Even if ($C'$) and ($C''$) are received in a definite order, the problem may be phrased in a way that indicates independence between the two constraints. In this case, MAXENT is the appropriate norm to use. *Infomin* does not assume such independence and therefore processes the two pieces of information separately. Disagreement arises when observations are interpreted differently, not because MAXENT and *Infomin* are inconsistent with each other. In the following I will assume that MAXENT and *Infomin* are compatible and part of the toolkit at the disposal of (M), the principle of maximum entropy.

Returning now to the issue of updating on conditionals, Wagner's method

and his plausible intuition (W) provide a generalization of Jeffrey conditioning, but contrary to Wagner's claims they do nothing to vitiate the principle of maximum entropy. Some advocates of (M) may find (L) too weak in its claims, but none think it is too strong. Once (L) is assumed, however, Wagner's diagnosis of disagreement between (W) and (M) fails. Moreover, (M) and (L) together seamlessly generalize Wagner conditioning. In the remainder of this paper I will provide a sketch of a formal proof for this claim. A welcome side-effect of reinstating harmony between (M) and (W) is that it provides an inverse procedure to Vladimír Majerník's method of finding marginals based on given conditional probabilities (see Majerník 2000 and my more technical companion paper to this one).

2 Wagner's Natural Generalization of Jeffrey Conditioning

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. This is best explained by using Wagner's example, the *Linguist* problem.

> You encounter the native of a certain foreign country and wonder whether he is a Catholic northerner (θ_1), a Catholic southerner (θ_2), a Protestant northerner (θ_3), or a Protestant southerner (θ_4). Your prior probability p over these possibilities (based, say, on population statistics and the judgment that it is reasonable to regard this individual as a random representative of his country) is given by $p(\theta_1) = 0.2, p(\theta_2) = 0.3, p(\theta_3) = 0.4,$ and $p(\theta_4) = 0.1$. The individual now utters a phrase in his native tongue which, due to the aural similarity of the phrases in question, might be a traditional Catholic piety (ω_1), an epithet uncomplimentary to Protestants (ω_2), an innocuous southern regionalism (ω_3), or a slang expression used throughout the country in question (ω_4). After reflecting on the matter you assign subjective probabilities $u(\omega_1) = 0.4, u(\omega_2) = 0.3, u(\omega_3) = 0.2,$ and $u(\omega_4) = 0.1$ to these alternatives. In the light of this new evidence how should you revise p? (See Wagner 1992, 252, and Spohn 2012, 197.)

Let $\Theta = \{\theta_i : i = 1, \ldots, 4\}, \Omega = \{\omega_i : i = 1, \ldots, 4\}$. Let $\Gamma : \Omega \to 2^{\Theta} - \{\emptyset\}$ be the function which maps ω to $\Gamma(\omega)$, the narrowest event in Θ entailed by the outcome $\omega \in \Omega$. Here are two definitions that take advantage of the apparatus established by Arthur Dempster (see Dempster 1967). We

will need m and b to articulate Wagner's (W) solution for *Linguist* type problems.

$$\text{For all } E \subseteq \Theta, m(E) = u(\{\omega \in \Omega : \Gamma(\omega) = E\}). \tag{1}$$

$$\text{For all } E \subseteq \Theta, b(E) = \sum_{H \subseteq E} m(H) = u(\{\omega \in \Omega : \Gamma(\omega) \subseteq E\}). \tag{2}$$

Let Q be the posterior joint probability measure on $\Theta \times \Omega$, and Q_Θ the marginalization of Q to Θ, Q_Ω the marginalization of Q to Ω. Wagner plausibly suggests that Q is compatible with u and Γ if and only if

$$\text{for all } \theta \in \Theta \text{ and for all } \omega \in \Omega, \theta \notin \Gamma(\omega) \text{ implies that } Q(\theta, \omega) = 0 \tag{3}$$

and

$$Q_\Omega = u. \tag{4}$$

The two conditions (3) and (4), however, are not sufficient to identify a "uniquely acceptable revision of a prior" (Wagner 1992, 250). Wagner's proposal includes a third condition, which extends Jeffrey's rule to the situation at hand. We will call it (W). To articulate the condition, we need some more definitions. For all $E \subseteq \Theta$, let $E_\bigstar = \{\omega \in \Omega : \Gamma(\omega) = E\}$, so that $m(E) = u(E_\bigstar)$. For all $A \subseteq \Theta$ and all $B \subseteq \Omega$, let "A" $= A \times \Omega$ and "B" $= \Theta \times B$, so that $Q(\text{"A"}) = Q_\Theta(A)$ for all $A \subseteq \Theta$ and $Q(\text{"B"}) = Q_\Omega(B)$ for all $B \subseteq \Omega$. Let also $\mathcal{E} = \{E \subseteq \Theta : m(E) > 0\}$ be the family of evidentiary focal elements.

According to Wagner only those Q satisfying the condition

$$\text{for all } A \subseteq \Theta \text{ and for all } E \in \mathcal{E}, Q(\text{"A"} \,|\, \text{"E}_\bigstar\text{"}) = p(A|E) \tag{5}$$

are eligible candidates for updated joint probabilities in *Linguist* type problems. To adopt (5), says Wagner, is to make sure that the total impact of

the occurrence of the event $E_\bigstar$ is to preclude the occurrence of any outcome $\theta \notin E$, and that, within E, p remains operative in the assessment of relative uncertainties (see Wagner 1992, 250). While conditions (3), (4) and (5) may admit an infinite number of joint probability distributions on $\Theta \times \Omega$, their marginalizations to Θ are identical and give us the desired posterior probability, expressible by the formula

$$q(A) = \sum_{E \in \mathcal{E}} m(E)p(A|E). \tag{6}$$

So far we are in agreement with Wagner. Wagner's scathing verdict about (M) towards the end of his article, however, is not really a verdict about (M) in the Laplacean tradition but about the curious conjunction of (M) and (I):

> Students of maximum entropy approaches to probability revision may [...] wonder if the probability measure defined by our formula (6) similarly minimizes [the Kullback-Leibler information number] $D_{\text{KL}}(q, p)$ over all probability measures q bounded below by b. The answer is negative [...] convinced by Skyrms, among others, that MAXENT is not a tenable updating rule, we are undisturbed by this fact. Indeed, we take it as additional evidence against MAXENT that (6), firmly grounded on [...] a considered judgment that (5) holds, might violate MAXENT [...] the fact that Jeffrey's rule coincides with MAXENT is simply a misleading fluke, put in its proper perspective by the natural generalization of Jeffrey conditioning described in this paper. [References to formulas and notation modified.] (Wagner 1992, 255.)

In the next section, we will contrast what Wagner considers to be the solution of (M) for this problem, 'Wagner's (M) solution,' and Wagner's solution presented in this section, 'Wagner's (W) solution,' and show, in much greater detail than Wagner does, why Wagner's (M) solution misrepresents (M).

3 Wagner's (M) Solution

Wagner's (M) solution assumes the constraint that b must act as a lower bound for the posterior probability. Consider $E_{12} = \{\theta_1 \vee \theta_2\}$. Because both ω_1 and ω_2 entail E_{12}, according to (2), $b(E_{12}) = 0.70$. It makes sense to

consider it a constraint that the posterior probability for E_{12} must be at least $b(E_{12})$. Then we choose from all probability distributions fulfilling the constraint the one which is closest to the prior probability distribution, using the Kullback-Leibler divergence.

Wagner applies this idea to the marginal probability distribution on Θ. He does not provide the numbers, but refers to simpler examples to make his point that (M) does not generally agree with his solution. To aid the discussion, I want to populate Wagner's claim for the *Linguist* problem with numbers. Using proposition 1.29 in Dimitri Bertsekas' book *Constrained Optimization and Lagrange Multiplier Methods* (see Bertsekas 1982, 71) and some non-trivial calculations, Wagner's (M) solution for the *Linguist* problem (indexed Q_{wm}) is

$$\tilde{\beta} = (Q_{wm}(\theta_j))^\intercal = (0.30, 0.45, 0.10, 0.15)^\intercal. \tag{7}$$

A brief remark about notation: I will use α for vectors expressing ω_i probabilities and β for vectors expressing θ_j probabilities. I will use a tilde as in $\tilde{\beta}$ or a hat as in $\hat{\beta}$ for posteriors, while priors remain without such ornamentation. The tilde is used for Wagner's (M) solution (which, as we will see, is incorrect) and the hat for the correct solution (both (W) and (M)).

The cross-entropy between $\tilde{\beta}$ and the prior

$$\beta = (P(\theta_j))^\intercal = (0.20, 0.30, 0.40, 0.10)^\intercal \tag{8}$$

is indeed significantly smaller than the cross-entropy between Wagner's (W) solution

$$\hat{\beta} = (Q(\theta_j))^\intercal = (0.30, 0.60, 0.04, 0.06)^\intercal \tag{9}$$

and the prior β (0.0823 compared to 0.4148). For the cross-entropy, we use the Kullback-Leibler Divergence

$$D_{\mathrm{KL}}(q, p) = \sum_j q(\theta_j) \log_2 \frac{q(\theta_j)}{p(\theta_j)}. \tag{10}$$

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. I will simplify the *Linguist* problem in order to show that the latter is the case.

> The *Simplified Linguist Problem.* Imagine the native is either Protestant or Catholic (50:50). Further imagine that the utterance of the native either entails that the native is a Protestant (60%) or provides no information about the religious affiliation of the native (40%).

Using (6), the posterior probability distribution is 80:20 (Wagner's (W) solution and, surely, the correct solution). Using b as a lower bound and (M), Wagner's (M) solution for this radically simplified problem is 60:40, clearly a more entropic solution than Wagner's (W) solution. The problem, as we will show, is that Wagner's (M) solution does not take into account (L), which an (M) advocate would naturally accept.

For a Laplacean, the prior joint probability distribution on $\Theta \times \Omega$ is not left unspecified for the calculation of the posteriors. Before the native makes the utterance, the event space is unspecified with respect to Ω. After the utterance, however, the event space is defined (or brought to attention) and populated by prior probabilities according to (L). That this happens retrospectively may or may not be a problem: Bayes' theorem is frequently used retrospectively, for example when the anomalous precession of Mercury's perihelion, discovered in the mid-1800s, was used to confirm Albert Einstein's General Theory of Relativity in 1915. I shall bracket for now that this procedure is controversial and refer the reader to the voluminous literature on Old Evidence.

Ariel Caticha and Adom Giffin make the following appeal:

> Bayes' theorem requires that $P(\omega, \theta)$ be defined and that assertions such as 'ω *and* θ' be meaningful; the relevant space is neither Ω nor Θ but the product $\Omega \times \Theta$ [notation modified] (Caticha and Giffin 2006, 9)

Following (L) we shall populate the joint probability matrix P on $\Omega \times \Theta$, which is a perfect task for MAXENT, as updating the joint probability P to Q on $\Omega \times \Theta$ will be a perfect task for *Infomin*. For the *Simplified Linguist*

Problem, this procedure gives us the correct result, agreeing with Wagner's (W) solution (80:20).

There is a more general theorem which incorporates Wagner's (W) method into Laplacean realism and MAXENT orthodoxy. The proof of this theorem will be in a more technical companion paper, but its validity is confirmed by how well it works for the *Linguist* problem (as well as the *Simplified Linguist Problem*).

4 The Linguist

The *Linguist* problem is a specific case of a more general Wagner-type problem characterized by two vectors and one matrix $(\beta, \hat{\alpha}, \kappa)$ (the dimensions are n, m, and $m \times n$, respectively). The first vector, β, represents the marginal prior probability $P(\theta_j)$. For the *Linguist problem*,

$$\beta = (0.2, 0.3, 0.4, 0.1)^\intercal. \tag{11}$$

The second vector, $\hat{\alpha}$, represents the marginal posterior probability $Q(\omega_i)$. For the *Linguist problem*,

$$\hat{\alpha} = (0.4, 0.3, 0.2, 0.1, 0)^\intercal. \tag{12}$$

Whereas Wagner only considers four dimensions, corresponding to the four utterances of the native, we have to add a fifth dimension corresponding to the case in which the native does not make any of those utterances, i.e. $\omega_5 = \neg(\bigvee_{i=1,\ldots,4} \omega_i)$. Presumably, the prior probability of ω_5 is very high, nearly 1 (the native may have uttered a typical Buddhist phrase, asked where the nearest bathroom was, complimented your fedora, or chosen to be silent, as a commenter pointed out to me). By the principle of regularity, however, it does not equal 1 (for a defence of the principle of regularity, that one should not assign probability 0 to any possibility, see Edwards et. al. 1963). The posterior probability is 0, as the *Linguist* problem specifies that one of the four possibilities was uttered by the native. $\hat{\alpha}_m$ is therefore always 0 for Wagner-type problems.

The matrix κ represents the logical relationships between the θ_j's and the ω_i's. In Wagner-type problems, the conditionals imply that some of the joint probabilities are zero. The observation of ω_i for $i = 1, \ldots, m-1$ implies that the last row of κ, which consists of 1's, becomes a row of 0's in the posterior representation $\hat{\kappa}$ of these relationships. Thus,

$$\kappa = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \hat{\kappa} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{13}$$

The triple $(\beta, \hat{\alpha}, \kappa)$ corresponds to Wagner's conditions (3) (dictating the zero joint probabilities or κ), (4) (dictating the marginal probabilities $\hat{\alpha}$ or $Q(\omega_i)$), and (6) (dictating the marginal probabilities β or $P(\theta_j)$). The marginal prior probabilities $\alpha = (P(\omega_1), \ldots, P(\omega_m)^\intercal)$ and posterior probabilities $\hat{\beta} = (Q(\theta_1), \ldots, Q(\theta_n)^\intercal)$ are unknown. We do not need to know α, but the point of the exercise is to determine $\hat{\beta}$. According to (W), $\hat{\beta} = (0.3, 0.6, 0.04, 0.06)$.

According to (M), we use Lagrange multipliers and first maximize the entropy of M, the joint prior probability matrix; then we use Lagrange multipliers again to minimize the cross-entropy from M to the joint posterior probability matrix $\hat{M}$. The situation can be visualized like this for the *Linguist* problem:

$$\begin{bmatrix} m_{11} & m_{12} & 0 & 0 & \alpha_1 \\ m_{21} & m_{22} & 0 & 0 & \alpha_2 \\ 0 & m_{32} & 0 & m_{34} & \alpha_3 \\ m_{41} & m_{42} & m_{43} & m_{44} & \alpha_4 \\ m_{51} & m_{52} & m_{53} & m_{54} & \alpha_5 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & 1.00 \end{bmatrix} \tag{14}$$

where the last column and the last row are the row and column sums of $M = (m_{ij})$. Similarly for the posterior joint probability matrix $\hat{M} = (\hat{m}_{ij})$

$$\begin{bmatrix} \hat{m}_{11} & \hat{m}_{12} & 0 & 0 & \hat{\alpha}_1 \\ \hat{m}_{21} & \hat{m}_{22} & 0 & 0 & \hat{\alpha}_2 \\ 0 & \hat{m}_{32} & 0 & \hat{m}_{34} & \hat{\alpha}_3 \\ \hat{m}_{41} & \hat{m}_{42} & \hat{m}_{43} & \hat{m}_{44} & \hat{\alpha}_4 \\ 0 & 0 & 0 & 0 & \hat{\alpha}_5 \\ \hat{\beta}_1 & \hat{\beta}_2 & \hat{\beta}_3 & \hat{\beta}_4 & 1.00 \end{bmatrix}. \tag{15}$$

The Lagrange multiplier method (for details, see the more technical companion paper) yields:

$$M = \frac{1}{e} r s^{\intercal} \circ \kappa \tag{16}$$

$$\hat{M} = \frac{1}{e} \hat{r} \hat{s}^{\intercal} \circ \hat{\kappa} \circ M \tag{17}$$

$$e\beta = S\kappa^{\intercal} r \tag{18}$$

$$e^2 \hat{\alpha} = \hat{R} \kappa \hat{s} \tag{19}$$

where $r_i = e^{\lambda_i}, s_j = e^{\mu_j}, \hat{r}_i = e^{\hat{\lambda}_i}, \hat{s}_j = e^{\hat{\mu}_j}$ represent factors arising from the Lagrange multiplier method. The operator $\circ$ is the entry-wise Hadamard product in linear algebra. $r, s, \hat{r}, \hat{s}$ are the vectors containing the $r_i, s_j, \hat{r}_i, \hat{s}_j$, respectively. $R, S, \hat{R}, \hat{S}$ are the diagonal matrices with $R_{il} = r_i \delta_{il}, S_{kj} = s_j \delta_{kj}, \hat{R}_{il} = \hat{r}_i \delta_{il}, \hat{S}_{kj} = \hat{s}_j \delta_{kj}$ (δ is Kronecker delta).

Wagner's (W) solution $\hat{\beta}$ solves this system of equation (but not Wagner's (M) solution $\tilde{\beta}$). Because maximum entropy and minimum cross-entropy solutions are unique (see Shore and Johnson 1980), (M) agrees with (W). To get there, we have assumed (L), namely that the joint probability matrices are populated by determinate probabilities. Wagner ostensibly disagrees with (L), as he represents the joint probability matrix $\hat{M}$ like this (visualized here with the marginals):

$$\begin{bmatrix} ? & ? & 0 & 0 & \hat{\alpha}_1 = 0.4 \\ ? & ? & 0 & 0 & \hat{\alpha}_2 = 0.3 \\ 0 & ? & 0 & ? & \hat{\alpha}_3 = 0.2 \\ ? & ? & ? & ? & \hat{\alpha}_4 = 0.1 \\ \hat{\beta}_1 = 0.3 & \hat{\beta}_2 = 0.6 & \hat{\beta}_3 = 0.04 & \hat{\beta}_4 = 0.06 & 1.00 \end{bmatrix}. \tag{20}$$

The posterior probability that the native encountered by the linguist is a northerner, for example, is 34%. (L) in conjunction with (M), by contrast, provides the joint probability matrix in full without lacunae.

$$\begin{bmatrix} 0.16 & 0.24 & 0 & 0 & \hat{\alpha}_1 = 0.4 \\ 0.12 & 0.18 & 0 & 0 & \hat{\alpha}_2 = 0.3 \\ 0 & 0.15 & 0 & 0.05 & \hat{\alpha}_3 = 0.2 \\ 0.02 & 0.03 & 0.04 & 0.01 & \hat{\alpha}_4 = 0.1 \\ \hat{\beta}_1 = 0.3 & \hat{\beta}_2 = 0.6 & \hat{\beta}_3 = 0.04 & \hat{\beta}_4 = 0.06 & 1.00 \end{bmatrix} \tag{21}$$

We have not formally demonstrated that for all Wagner-type problems $(\beta, \hat{\alpha}, \kappa)$, the correct (M) solution (versus Wagner's deficient (M) solution) agrees with Wagner's (W) solution, although we have established a useful framework and demonstrated the agreement for the *Linguist* problem. The technical companion paper will accomplish the more general proof. As Vladimír Majerník has shown how to derive marginal probabilities from conditional probabilities using (M) (see Majerník 2000), we will inversely show how to derive conditional probabilities (i.e. the joint probability matrices) from the marginal probabilities and logical relationships provided in Wagner-type problems. This technical result together with the claim established in the present paper that Wagner's intuition (W) is consistent with (M), given (L), underlines the formal and conceptual virtue of (M).

5 References

Bertsekas, D. (1982). Constrained Optimization and Lagrange Multiplier Methods. Boston, MA: Academic.

Caticha, A. & Giffin, A. (2006). "Updating Probabilities." In MaxEnt 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods.

Dempster, A. (1967).“Upper and Lower Probabilities Induced by a Multivalued Mapping.” The Annals of Mathematical Statistics, 38 (2), 325–339.

Edwards, W., Lindman H., Savage L.J. (1963). “Bayesian Statistical Inference for Psychological Research.” Psychological Review, 70 (3), 193.

Elga, A. (2010). “Subjective Probabilities Should Be Sharp.” Philosophers' Imprints, 10 (5), 1–11.

Ellsberg, D. (1961). “Risk, Ambiguity, and the Savage Axioms.” The Quarterly Journal of Economics, 75 (4), 643–669.

Friedman, K. & Shimony, A. (1971). “Jaynes's Maximum Entropy Prescription and Probability Theory.” Journal of Statistical Physics, 3 (4), 381–384.

Grünwald, P. & Halpern, J.Y. (2003). “Updating Probabilities.” Journal of Artificial Intelligence Research, 19, 243–278.

Joyce, J. (2010). “A Defense of Imprecise Credences in Inference and Decision Making.” Philosophical Perspectives, 24 (1), 281–323.

Levi, I. (1985). “Imprecision and Indeterminacy in Probability Judgment.” Philosophy of Science, 52 (3), 390–409.

Lukits, S. (2014). “The Principle of Maximum Entropy and a Problem in Probability Kinematics.” Synthese, 191 (7), 1409–1431.

Majerník, V. (2000). “Marginal Probability Distribution Determined by the Maximum Entropy Method.” Reports on Mathematical Physics, 45 (2), 171–181.

Seidenfeld, T. (1979). “Why I Am Not an Objective Bayesian; Some Reflections Prompted by Rosenkrantz.” Theory and Decision, 11 (4), 413–440.

Shore, J. & Johnson, R.W. (1980). “Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy.” IEEE Transactions on Information Theory, 26 (1), 26–37.

Spohn, W. (2012). The Laws of Belief: Ranking Theory and Its Philosophical Applications. Oxford University.

Wagner, C.G. (1992). “Generalized Probability Kinematics.” Erkenntnis, 36 (2), 245–257.

Walley, P. (1991). Statistical Reasoning with Imprecise Probabilities. Chapman and Hall London.

White, R. (2010). "Evidential Symmetry and Mushy Credence." In Oxford Studies in Epistemology, edited by Tamar Gendler, and John Hawthorne, New York, NY: Oxford University, 161–186.

Zabell, S.L. (2005). Symmetry and Its Discontents: Essays on the History of Inductive Probability. Cambridge University.