# A Natural Generalization of Jeffrey Conditioning

Stefan Lukits

## 1 Introduction

Updating on conditionals is a swiftly developing field in probability kinematics and the belief revision literature in general. 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 one is informed of a conditional rather than an event, 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 constraint 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, 1971).

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 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 flawed. It is based on a denial of principle (L), the Laplacean principle about the assignment of determinate prior probabilities to well-defined events, which we will specify in a moment.

While Wagner is welcome to deny (L), advocates of (M) universally accept it. Wagner is correct in pointing out that denying (L) and accepting (M) is inconsistent. Since all advocates of (M) already accept (L), this is not going to dissuade anyone from continuing their allegiance, neither does it do anything to deepen criticism of (M). This paper shows how elegantly (M) generalizes not only standard conditioning and Jeffrey conditioning, but also Wagner conditioning.

To broaden the horizon for a moment, the criticism of (M) in the 1970s and 1980s succeeded in tarnishing the reputation of (M) as a generally valid, objective updating method in epistemological circles (although many statistical physicists still hold it in high regard). I find that it has mostly done so by pointing out weaknesses in Jaynes' "right-wing totalitarian" exposition of (M) (see Zabell, 2005, 28). A tempered and differentiated account of (M) is not only largely immune to the criticisms, but often insightfully 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 states that a rational agent is able to assign a probability to a well-defined event. In contrast to frequentism and in agreement with Bayesians, (L) holds an epistemological view of probabilities where the probability distribution represents an agent's uncertainty or lack of information and not a fact in the world. When you change your probabilities you are not admitting that you have been wrong, you are only adjusting to new information. Sometimes you know very little about an event, but knowing little is no barrier to assigning a probability—the whole point of probabilities is to reflect uncertainty.

There are many caveats here. To avoid excessive apriorism, as Teddy Seidenfeld calls it (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 Prior Probabilities I and Prior Probabilities II. The former precede any information at all. 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 Prior Probabilities I, has failed.

The case for objectivity in probability kinematics, where prior probabilities are of type II, is different. Again, 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, (L) in conjunction with (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.

(M) requires that we do so 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, green, and red tokens. You know that (C1) at least 50% of the tokens are blue. Then you learn that (C2) 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 (C1) and (C2). The three-dimensional vector containing the probabilities for blue, green, and red is $(\frac{1}{2}, \frac{1}{5}, \frac{3}{10})$. *Infomin*, on the other hand, takes in its last step $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ as its prior probability distribution and then diachronically updates 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 (C2) before (C1) ('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 (C1) and (C2) 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.

While I have emphasized so far that (L) is a weaker claim compared to what Jaynes may have claimed (perhaps contrasting a Laplacean realism of sorts with Jaynes' Laplacean idealism), (L) is a stronger claim than what some Bayesians are prepared to accept. Classically, Bayesians were committed to determinate prior probabilities, but in recent time this commitment has softened. There are now Bayesians who contend that rational agents typically lack determinate prior subjective probabilities and that their opinions are characterized by imprecise credal states in response to unspecific and equivocal evidence. Laplacean realism is weaker than Laplacean idealism, but stronger than Bayesianism, if Bayesianism now includes this kind of softening.

While I appreciate the equivocality of evidence, I would separate the disambiguation of the evidence in articulating formal constraints from bringing to bear a helpful formalism to probability updating which requires numerically precise priors. When we apply mathematics to daily life, we do this by measuring imprecisely and then processing the disambiguated measurements using calculus. One particularly strong advocate of imprecise credal states is James Joyce (see Joyce, 2005, 156f), with the unfortunate consequence that the updating strategies that Joyce proposes for these credal states are impotent.

No amount of evidence can modify the imprecise credal state, because each member of the set of credal states that an agent accepts has a successor with respect to updating that is also a member of these credal states and that is consistent with its predecessor and the evidence. Although the feeling is that the imprecise credal state is narrowed by evidence towards more precision, set theory clearly indicates that the credal state remains static, no matter what the evidence is, unless we introduce a higher-level distribution over these sets—but then the same problems arise on the higher level.

Returning from a defence of (L) to updating on conditionals, Wagner's method and his plausible intuition (W) provide an interesting generalization of Jeffrey conditioning, but contrary to Wagner's claims they do nothing to vitiate Laplacean realism or (M). Some advocates of (M) may find Laplacean realism 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 produce a formalism which seamlessly generalizes Wagner conditioning. A welcome side-effect of reinstating harmony between (M) and (W) is that this formalism provides an inverse procedure to Vladimír Majerník's method of finding marginals based on given conditional probabilities (see Majerník, 2000).

## 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 ($\theta_1$), a Catholic southerner ($\theta_2$), a Protestant northerner ($\theta_3$), or a Protestant southerner ($\theta_4$). 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 ($\omega_1$), an epithet uncomplimentary to Protestants ($\omega_2$), an innocuous southern regionalism ($\omega_3$), or a slang expression used throughout the country in question ($\omega_4$). 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 $\omega$ to $\Gamma(\omega)$, the narrowest event in $\Theta$ 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 $Q_\Theta$ the marginalization of $Q$ to $\Theta$, $Q_\Omega$ the marginalization of $Q$ to $\Omega$. Wagner plausibly suggests that $Q$ is compatible with $u$ and $\Gamma$ 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 an additional formal apparatus. 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 $\Theta$ 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, although a Laplacean approach, combined with (M), gives us exactly the same results, as we will show below, and moreover easily generalizes over both Jeffrey conditioning and Wagner conditioning. Wagner's scathing verdict about (M) towards the end of his article is not really a verdict about (M) in the Laplacean tradition, where most (if not all) of its advocates are at home, but about the curious conjunction of (M) with a non-Laplacean Bayesianism that renounces any commitment to determinate priors on specified partitions of the event space:

> 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 $\omega_1$ and $\omega_2$ 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 $\Theta$. 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

$$\tilde{\beta} = (Q_{wm}(\theta_1) = 0.30, Q_{wm}(\theta_2) = 0.45, Q_{wm}(\theta_3) = 0.10, Q_{wm}(\theta_4) = 0.15) \tag{7}$$

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

$$\beta = (P(\theta_1) = 0.20, P(\theta_2) = 0.30, P(\theta_3) = 0.40, P(\theta_4) = 0.10) \tag{8}$$

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

$$\hat{\beta} = ((Q(\theta_1) = 0.30, Q(\theta_2) = 0.60, Q(\theta_3) = 0.04, Q(\theta_4) = 0.06)) \tag{9}$$

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

$$D_{\text{KL}}(q,p) = \sum q(\theta_i) \log_2 \frac{q(\theta_i)}{p(\theta_i)}. \tag{10}$$

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 $\Omega$. 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 is (at least prima facie) not a problem: Bayes' theorem is used retrospectively all the time, 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. Ariel Caticha and Adom Giffin make the following appeal:

> Bayes' theorem requires that $P(\omega, \theta)$ be defined and that assertions such as '$\omega$ *and* $\theta$' be meaningful; the relevant space is neither $\Omega$ nor $\Theta$ 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, $\beta$, 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 $\omega_5$ 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). 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 $\kappa$ represents the logical relationships between the $\theta_j$'s and the $\omega_i$'s. In Wagner-type problems, the conditionals imply that some of the joint probabilities are zero. The observation of $\omega_i$ for $i = 1, \ldots, m-1$ implies that the last row of $\kappa$, which consists of 1's, becomes a row of 0's in the posterior representation $\hat{\kappa}$ of these relationships. Thus,

$$\kappa = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \hat{\kappa} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad (13)$$

The triple $(\beta, \hat{\alpha}, \kappa)$ corresponds to Wagner's conditions (3) (dictating the zero joint probabilities or $\kappa$), (4) (dictating the marginal probabilities $\hat{\alpha}$ or $Q(\omega_i)$), and (6) (dictating the marginal probabilities $\beta$ or $P(\theta_j)$). The marginal prior probabilities $\alpha = (P(\omega_1), \ldots, P(\omega_m)^\intercal)$ and $\hat{\beta} = (Q(\theta_1), \ldots, Q(\theta_n)^\intercal)$ are unknown. We do not need to know $\alpha$, 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:

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

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 $R_{il} = r_i \delta_{il}, S_{kj} = s_j \delta_{kj}, \hat{R}_{il} = \hat{r}_i \delta_{il}, \hat{S}_{kj} = \hat{s}_j \delta_{kj}$ ($\delta$ is Kronecker delta). For those who want to investigate this more closely, here are the Lagrangian functions:

$$\begin{aligned} &\Lambda(m_{ij}, \lambda, \mu) = \\ &\sum_{\kappa_{ij}=1} m_{ij} \log m_{ij} + \sum_{i=1}^{m} \lambda_i \left( \alpha_i - \sum_{\kappa_{il}=1} m_{il} \right) + \\ &\sum_{j=1}^{n} \mu_j \left( \beta_j - \sum_{\kappa_{kj}=1} m_{kj} \right) \end{aligned} \tag{20}$$

$$\hat{\Lambda}(\hat{m}_{ij},\hat{\lambda},\hat{\mu}) =$$

$$\sum_{\hat{\kappa}_{ij}=1}\hat{m}_{ij}\log\frac{\hat{m}_{ij}}{m_{ij}}+\sum_{i=1}^{m}\hat{\lambda}_i\left(\hat{\alpha}_i-\sum_{\hat{\kappa}_{il}=1}\hat{m}_{il}\right)+$$

$$\sum_{j=1}^{n}\hat{\mu}_j\left(\hat{\beta}_j-\sum_{\hat{\kappa}_{kj}=1}\hat{m}_{kj}\right) \tag{21}$$

As is easy to check, 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):

$$\begin{bmatrix} ? & ? & 0 & 0 & \hat{\alpha}_1=0.4 \\ ? & ? & 0 & 0 & \hat{\alpha}_2=0.3 \\ 0 & ? & 0 & ? & \hat{\alpha}_3=0.2 \\ ? & ? & ? & ? & \hat{\alpha}_4=0.1 \\ \hat{\beta}_1=0.3 & \hat{\beta}_2=0.6 & \hat{\beta}_3=0.04 & \hat{\beta}_4=0.06 & 1.00 \end{bmatrix} \tag{22}$$

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.

$$\begin{bmatrix} 0.16 & 0.24 & 0 & 0 & \hat{\alpha}_1=0.4 \\ 0.12 & 0.18 & 0 & 0 & \hat{\alpha}_2=0.3 \\ 0 & 0.15 & 0 & 0.05 & \hat{\alpha}_3=0.2 \\ 0.02 & 0.03 & 0.04 & 0.01 & \hat{\alpha}_4=0.1 \\ \hat{\beta}_1=0.3 & \hat{\beta}_2=0.6 & \hat{\beta}_3=0.04 & \hat{\beta}_4=0.06 & 1.00 \end{bmatrix} \tag{23}$$

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, Dimitri. *Constrained Optimization and Lagrange Multiplier Methods.* Boston, MA: Academic, 1982.

Caticha, Ariel, and Adom Giffin. "Updating Probabilities." In *MaxEnt 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods.* 2006.

Dempster, Arthur. "Upper and Lower Probabilities Induced by a Multivalued Mapping." *The Annals of Mathematical Statistics* 38, 2: (1967) 325–339.

Friedman, Kenneth, and Abner Shimony. "Jaynes's Maximum Entropy Prescription and Probability Theory." *Journal of Statistical Physics* 3, 4: (1971) 381–384.

Joyce, James. "How Probabilities Reflect Evidence." *Philosophical Perspectives* 19, 1: (2005) 153–178.

Lukits, Stefan. "The Principle of Maximum Entropy and a Problem in Probability Kinematics." *Synthese* (forthcoming) (2014).

Majerník, Vladimír. "Marginal Probability Distribution Determined by the Maximum Entropy Method." *Reports on Mathematical Physics* 45, 2: (2000) 171–181.

Seidenfeld, Teddy. "Why I Am Not an Objective Bayesian; Some Reflections Prompted by Rosenkrantz." *Theory and Decision* 11, 4: (1979) 413–440.

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

Spohn, Wolfgang. *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford, 2012.

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

Zabell, Sandy. *Symmetry and Its Discontents: Essays on the History of Inductive Probability*. Cambridge University Press, 2005.