# Maximum Entropy and Probability Kinematics Constrained by Conditionals

Stefan Lukits

## 1 Introduction

In his paper "Marginal Probability Distribution Determined by the Maximum Entropy Method" (see Majerník, 2000), Vladimír Majerník asks the following question: If we had two partitions of an event space and knew all the conditional probabilities (any conditional probability of one event in the first partition conditional on another event in the second partition), would we be able to calculate the marginal probabilities for the two partitions? The answer is yes, if we commit ourselves to the principle of maximum entropy, from now on PME:

> [PME] Keep the information entropy of your probability distribution maximal within the constraints that the evidence provides (in the synchronic case), or your cross-entropy minimal (in the diachronic case).

For Majernik's question, PME provides us with a unique and plausible answer (see Majernik's paper). We may also be interested in the obverse question: if the marginal probabilities of the two partitions were given, would we similarly be able to calculate the conditional probabilities? The answer is yes: given PME, Theorems 2.2.1. and 2.6.5. in *Elements of Information Theory* (see Cover and Thomas, 2006) reveal that the joint probabilities are the product of the marginal probabilities. Once the joint probabilities and the marginal probabilities are available, it is trivial to calculate the conditional probabilities.

There is an older problem by Carl Wagner (see Wagner, 1992), which can be cast in similar terms. If we were given some of the marginal probabilities in an updating problem as well as some logical relationships between the two partitions, would we be able to calculate the remaining marginal probabilities? This problem is best understood by example (see Wagner's *Linguist* problem in section 3). Wagner solves it with a natural generalization of Jeffrey conditioning, which we will call Wagner conditioning. It is not based on PME, but on what we may call Jeffrey's updating principle, or JUP for short:

> [JUP] In a diachronic updating process, keep the ratio of probabilities constant as long as they are unaffected by the constraints that the evidence poses.

Richard Jeffrey made this principle famous when he introduced probability kinematics (see Jeffrey, 1965), an updating method which operates on uncertain evidence. As is the case for PME, there is a debate whether updating on evidence by rational agents is bound by JUP (for a defence see Teller, 1973; for detractors see Howson and Franklin, 1994).

Our interest in this article is the relationship between PME and JUP, both of which are updating principles. Wagner contends that his natural generalization of Jeffrey conditioning, based on JUP, contradicts PME. Among formal epistemologists, there is a widespread view that, while PME is a generalization of Jeffrey conditioning, it is an inappropriate updating method in certain cases and does not enjoy the generality of Jeffrey conditioning. Wagner's claims support this view inasmuch as Wagner conditioning is based on the relatively plausible JUP and naturally generalizes Jeffrey conditioning, but according to Wagner it contradicts PME, which gives wrong results in these cases.

I am generally suspicious of the widespread view that there are problems with PME which go beyond the problems of a more general Bayesian viewpoint with respect to probability updating. Although a dominant majority of Bayesians does not accept PME to be a generally valid updating method, I believe that there are persuasive arguments that Bayesian commitments, especially if they are coupled with commitments to JUP, should lead to adherence to PME. Once one accepts JUP, counterexamples to PME and their attendant conceptual problems can be successfully addressed. This is a larger project, which receives support in the more specific claims advanced in this paper, although the more specific claims can be independently and profitably evaluated without reference to the larger project.

Here is this paper's more specific claim: PME generalizes both Jeffrey conditioning and Wagner conditioning, providing a much more integrated approach to probability updating. This integrated approach also gives a coherent answer to the obverse Majernik problem posed above. A more epistemologically oriented companion paper shows that Wagner's argument about the contradiction between JUP and PME is specious.

## 2 Jeffrey Conditioning

Richard Jeffrey proposes an updating method for cases in which the evidence is uncertain, generalizing standard probabilistic conditioning. I will present this method in very unusual notation, anticipating using my notation to solve Wagner's *Linguist* problem and to give a general solution for the obverse Majernik problem. Let $\Omega$ be an event space with finitely many elements and $\{\theta_j\}_{j=1,\ldots,n}$ a partition of $\Omega$. Let $\kappa$ be an $m \times n$ matrix for which each column contains exactly one 1, otherwise 0. Let $P = P_{\text{prior}}$ and $\hat{P} = P_{\text{posterior}}$. Then $\{\omega_i\}_{i=1,\ldots,m}$, for which

$$\omega_i = \bigcup_{j=1,\ldots,n} \theta^*_{ij}, \tag{1}$$

is likewise a partition of $\Omega$ (the $\omega$s are basically a more coarsely grained partition than the $\theta$s). $\theta^*_{ij} = \emptyset$ if $\kappa_{ij} = 0$, $\theta^*_{ij} = \theta_j$ otherwise. Let $\beta$ be the vector of prior probabilities for $\{\theta_j\}_{j=1,\ldots,n}(P(\theta_j) = \beta_j)$ and $\hat{\beta}$ the vector of posterior probabilities $(\hat{P}(\theta_j) = \hat{\beta}_j)$; likewise for $\alpha$ and $\hat{\alpha}$ corresponding to the prior and posterior probabilities for $\{\omega_i\}_{i=1,\ldots,m}$, respectively.

A Jeffrey-type problem is when $\beta$ and $\hat{\alpha}$ are given and we are looking for $\hat{\beta}$. A mathematically more concise characterization of a Jeffrey-type problem is the triple $(\kappa, \beta, \hat{\alpha})$. The solution, using Jeffrey conditioning, is

$$\hat{\beta}_j = \beta_j \sum_{i=1}^{m} \frac{\kappa_{ij}\hat{\alpha}_i}{\sum_{\kappa_{il}=1} \beta_l} \text{ for all } j = 1, \ldots, n. \tag{2}$$

I will soon introduce an example which makes this notation more perspicuous. The notation is at first glance off-putting, but we will in the following take full advantage of it to present a generalization where the $\omega_i$ do not range over the $\theta_j$.

## 3 Wagner Conditioning

Carl Wagner uses JUP (explained in more detail in Wagner, 2002) to solve a problem which cannot be solved by Jeffrey conditioning. Here is the narrative (call this 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 us call a problem of this type a Wagner-type problem. It is an instance of the more general obverse Majernik problem where partitions are given with logical relationships between them as well as some marginal probabilities. Wagner-type problems seek as a solution missing marginals, while obverse Majernik problems seek the conditional probabilities as well, both of which we will eventually provide using PME.

Wagner's solution for such problems (from now on Wagner conditioning) rests on JUP and a formal apparatus established by Arthur Dempster (see Dempster, 1967), which is quite different from our notational approach. Wagner legitimately calls his solution a "natural generalization of Jeffrey conditioning" (see Wagner, 1992, 250). There is, however, another natural generalization of Jeffrey conditioning, E.T. Jaynes' principle of maximum entropy. PME does not rest on JUP, but rather claims that one should keep one's entropy maximal within the constraints that the evidence provides (in the synchronic case) and one's cross-entropy minimal (in the diachronic case). Some distinguish between MAXENT, the synchronic rule, and *Infomin*, the diachronic rule, but I have shown elsewhere that the two are compatible and both follow PME (see also Wagner, 2002).

It turns out that PME elegantly generalizes Jeffrey conditioning and therefore absorbs JUP on the more narrow domain of problems that we can solve using Jeffrey conditioning (for a proof see Caticha and Giffin, 2006). Wagner's contention is that on the wider domain of problems where we must use Wagner conditioning, JUPand PME contradict each other. We are now in the awkward position of being confronted with two plausible intuitions, JUP and PME, and it appears that we have to let one of them go. Wagner adduces other conceptual problems for PME (e.g. Bas van Fraassen's *Judy Benjamin* problem and Abner Shimony's Lagrange multiplier problem, see Friedman and Shimony, 1971) to reinforce his conclusion that PME is not a principle on which we should rely in general.

We will see that Wagner's conclusion is incorrect. JUP and PME are compatible. Wagner's formal apparatus, although inspiring, is unnecessary and ad hoc, as the much more integrated maximum entropy approach seamlessly generalizes JUP. There are now two distinctive tasks at hand. One is to show how Wagner construes a contradiction between JUP and PME and where this construction is misleading. I will do this in a more epistemological companion paper, because Wagner's mistake is more epistemological in nature than formal, attributing implausible assumptions to adherents of PME. The other more general and more formal task, which we will pursue here, is to show how PME generalizes Jeffrey conditioning and Wagner conditioning to boot.

## 4 A Natural Generalization of Jeffrey and Wagner Conditioning

To achieve the second task, we use the notation that we have already introduced for Jeffrey conditioning. We can characterize Wagner-type problems analogously to Jeffrey-type problems by a triple $(\kappa, \beta, \hat{\alpha})$. $\{\theta_j\}_{j=1,\ldots,n}$ and $\{\omega_i\}_{i=1,\ldots,m}$ now refer to independent partitions of $\Omega$, i.e. (1) need not be true. Besides the marginal probabilities $P(\theta_j) = \beta_j, \hat{P}(\theta_j) = \hat{\beta}_j, P(\omega_i) = \alpha_i, \hat{P}(\omega_i) = \hat{\alpha}_i$, we therefore also have joint probabilities $m_{ij} = P(\omega_i \cap \theta_j)$ and $\hat{m}_{ij} = \hat{P}(\omega_i \cap \theta_j)$.

Given the specific nature of Wagner-type problems, there are a few constraints on the triple $(\kappa, \beta, \hat{\alpha})$. The last row $(m_{mj})_{j=1,\ldots,n}$ is special because it represents the probability of $\omega_m$, which is the negation of the events deemed possible after the observation. In the *Linguist* problem, for example, $\omega_5$ is the event (initially highly likely, but impossible after the observation of the native's utterance) that the native does not make any of the four utterances. The native may have, after all, uttered a typical Buddhist phrase, asked where the nearest bathroom was, complimented your fedora, or chosen to be silent. $\kappa$ will have all 1s in the last row. Let $\hat{\kappa}_{ij} = \kappa_{ij}$ for $i = 1, \ldots, m-1$ and $j = 1, \ldots, n$; and $\hat{\kappa}_{mj} = 0$ for $j = 1, \ldots, n$. $\hat{\kappa}$ equals $\kappa$ except that its last row are all 0s, and $\hat{\alpha}_m = 0$. Otherwise the 0s are distributed over $\kappa$ (and equally over $\hat{\kappa}$) so that no row and no column has all 0s, representing the logical relationships between the $\omega_i$s and the $\theta_j$s. We set $P(\omega_m) = x$ ($\hat{P}(\omega_m) = 0$), where $x$ depends on your prior knowledge. Fortunately, the value of $x$ cancels out nicely and will play no further role. For convenience, we define $\zeta = (0, \ldots, 0, 1)^\intercal$ with $\zeta_m = 1$.

The best way to visualize such a problem is by providing the joint probability matrix $M = (m_{ij})$ together with the marginals $\alpha$ and $\beta$ in the last column/row, here for example as for the *Linguist* problem with $m = 5$ and $n = 4$,

$$\begin{bmatrix} m_{11} & m_{12} & 0 & 0 & \alpha_1 \\ m_{21} & m_{22} & 0 & 0 & \alpha_2 \\ 0 & m_{32} & 0 & m_{34} & \alpha_3 \\ m_{41} & m_{42} & m_{43} & m_{44} & \alpha_4 \\ m_{51} & m_{52} & m_{53} & m_{54} & x \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & 1.00 \end{bmatrix}. \tag{3}$$

The $m_{ij} \neq 0$ where $\kappa_{ij} = 1$. Ditto, mutatis mutandis, for $\hat{M}, \hat{\alpha}, \hat{\beta}$. To make this a little less abstract, Wagner's *Linguist* problem is characterized by the triple $(\kappa, \beta, \hat{\alpha})$,

$$\kappa = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \hat{\kappa} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{4}$$

$$\beta = (0.2, 0.3, 0.4, 0.1)^\intercal \text{ and } \hat{\alpha} = (0.4, 0.3, 0.2, 0.1, 0)^\intercal. \tag{5}$$

Wagner's solution, based on JUP, is

$$\hat{\beta}_j = \beta_j \sum_{i=1}^{m-1} \frac{\hat{\kappa}_{ij}\hat{\alpha}_i}{\sum_{\hat{\kappa}_{il}=1} \beta_l} \text{ for all } j = 1, \ldots, n. \tag{6}$$

In numbers,

$$\hat{\beta}_j = (0.3, 0.6, 0.04, 0.06)^{\intercal}. \tag{7}$$

The posterior probability that the native encountered by the linguist is a northerner, for example, is 34%. Wagner's notation is completely different and never specifies or provides the joint probabilities, but I hope the reader appreciates both the analogy to (2) underlined by this notation as well as its efficiency in delivering a correct PME solution for us (compared to Wagner's incorrect PME solution, which is to some extent misleadingly suggested by Wagner's Dempsterian setup). That (6) follows from JUP is well-documented in Wagner's article.

For the PME solution for this problem, we will not use (6) or JUP, but maximize the entropy for the joint probability matrix $M$ and then minimize the cross-entropy between the prior probability matrix $M$ and the posterior probability matrix $\hat{M}$. The PME solution, despite its seemingly different ancestry in principle, formal method, and assumptions, agrees with (6). This completes our argument.

What follows may be accessible largely to PME cognoscenti, since it involves the Lagrange multiplier method (see Guiaşu, 1977, 327ff, and Jaynes, 1978, 244). Others may want to skip to the Conclusion. To maximize the Shannon entropy of $M$ and minimize the Kullback-Leibler divergence between $\hat{M}$ and $M$, consider the Lagrangian functions:

$$\Lambda(m_{ij}, \mu) =$$
$$\sum_{\kappa_{ij}=1} m_{ij} \log m_{ij} + \sum_{j=1}^{n} \mu_j \left( \beta_j - \sum_{\kappa_{kj}=1} m_{kj} \right) +$$
$$\lambda_m \left( x - \sum_{j=1}^{n} m_{mj} \right) \tag{8}$$

and

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

$$\sum_{\hat{\kappa}_{ij}=1} \hat{m}_{ij} \log \frac{\hat{m}_{ij}}{m_{ij}} + \sum_{i=1}^{m} \hat{\lambda}_i \left( \hat{\alpha}_i - \sum_{\hat{\kappa}_{il}=1} \hat{m}_{il} \right). \tag{9}$$

For the optimization, we set the partial derivatives to 0, which results in

$$M = rs^{\mathsf{T}} \circ \kappa \tag{10}$$

$$\hat{M} = \hat{r}s^{\mathsf{T}} \circ \hat{\kappa} \tag{11}$$

$$\beta = S\kappa^{\mathsf{T}} r \tag{12}$$

$$\hat{\alpha} = \hat{R}\kappa s \tag{13}$$

where $r_i = e^{\zeta_i \lambda_m}, s_j = e^{-1-\mu_j}, \hat{r}_i = e^{-1-\hat{\lambda}_i}$ represent factors arising from the Lagrange multiplier method. The operator $\circ$ is the entry-wise Hadamard product in linear algebra. $r, s, \hat{r}$ are the vectors containing the $r_i, s_j, \hat{r}_i$, respectively. $R, S, \hat{R}$ are the diagonal matrices with $R_{il} = r_i \delta_{il}, S_{kj} = s_j \delta_{kj}, \hat{R}_{il} = \hat{r}_i \delta_{il}$ ($\delta$ is Kronecker delta).

Note that

$$\frac{\beta_j}{\sum_{\hat{\kappa}_{il}=1} \beta_l} = \frac{s_j}{\sum_{\hat{\kappa}_{il}=1} s_l} \text{ for all } (i,j) \in \{1, \ldots, m-1\} \times \{1, \ldots, n\}. \tag{14}$$

(13) implies

$$\hat{r}_i = \frac{\hat{\alpha}_i}{\sum_{\hat{\kappa}_{il}=1} s_l} \text{ for all } i = 1, \ldots, m-1. \tag{15}$$

Consequently,

$$\hat{\beta}_j = s_j \sum_{i=1}^{m-1} \frac{\hat{\kappa}_{ij}\hat{\alpha}_i}{\sum_{\kappa_{il}=1} s_l} \text{ for all } j = 1, \ldots, n. \tag{16}$$

(16) gives us the same solution as (6), taking into account (14). Therefore, Wagner conditioning and PME agree.

## 5 Conclusion

Wagner-type problems (but not obverse Majernik-type problems) can be solved using JUP and Wagner's ad hoc method. Obverse Majernik-type problems, and therefore all Wagner-type problems, can also be solved using PME and its established and integrated formal method. What at first blush looks like serendipitous coincidence, namely that the two approaches deliver the same result, reveals that JUP is safely incorporated in PME. Not to gain information where such information gain is unwarranted and to process all the available and relevant information is the intuition at the foundation of PME. My results show that this more fundamental intuition generalizes the more specific intuition that ratios of probabilities should remain constant unless they are affected by observation or evidence. Wagner's argument that PME conflicts with JUP is ineffective because, as my more epistemological companion paper demonstrates, it rests on assumptions that advocates of PME naturally reject.

## 6 References


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

Cover, T.M., and J.A. Thomas. *Elements of Information Theory*, volume 6. Hoboken, NJ: Wiley, 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.

Guiaşu, Silviu. *Information Theory with Application.* New York: McGraw-Hill, 1977.

Howson, Colin, and Allan Franklin. "Bayesian Conditionalization and Probability Kinematics." *The British Journal for the Philosophy of Science* 45, 2: (1994) 451–466.

Jaynes, E.T. "Where Do We Stand on Maximum Entropy." In *The Maximum Entropy Formalism*, edited by R.D. Levine, and M. Tribus, Cambridge, MA: MIT, 1978, 15–118.

Jeffrey, Richard. *The Logic of Decision.* New York, NY: Gordon and Breach, 1965.

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

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

Teller, Paul. "Conditionalization and Observation." *Synthese* 26, 2: (1973) 218–258.

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

———. "Probability Kinematics and Commutativity." *Philosophy of Science* 69, 2: (2002) 266–278.