Submitted to *Entropy*. Pages 1 - 12.

OPEN ACCESS

Article

Maximum Entropy and Probability Kinematics Constrained by Conditionals

Stefan Lukits

[1] Philosophy Department, University of British Columbia, 1866 Main Mall, Buchanan E370, Vancouver BC V6T 1Z1, Canada

Version March 23, 2015 submitted to Entropy. *Typeset by LaTeX using class file mdpi.cls*

Abstract: Two open questions of inductive reasoning are solved: (1) does the principle of maximum entropy (PME) give a solution to the obverse Majerník problem; and (2) is Wagner correct when he claims that Jeffrey's updating principle (JUP) contradicts PME? Majerník shows that PME provides unique and plausible marginal probabilities, given conditional probabilities. The obverse problem posed here is whether PME also provides such conditional probabilities, given certain marginal probabilities. The theorem developed to solve the obverse Majerník problem demonstrates that in the special case introduced by Wagner PME does not contradict JUP, but elegantly generalizes it and offers a more integrated approach to probability updating.

Keywords: Probability update; Jeffrey conditioning; principle of maximum entropy; formal epistemology; conditionals; probability kinematics.

1. Introduction

Jeffrey conditioning is a method of update (recommended first by Richard Jeffrey in [13]) which generalizes standard conditioning and operates in probability kinematics where evidence is uncertain ($P(E) \neq 1$). Sometimes, when we reason inductively, outcomes that are observed have entailment relationships with partitions of the possibility space that pose challenges that Jeffrey conditioning cannot meet. As we will see, it is not difficult to resolve these challenges by generalizing Jeffrey conditioning. There are claims in the literature that the principle of maximum entropy, from now on PME, conflicts with this generalization. I will show under which conditions this conflict obtains. Since proponents of PME are unlikely to subscribe to these conditions, the position of PME in the larger debate over inductive logic and reasoning is not undermined.

In Section 2, I will introduce the obverse Majerník problem and sketch how it ties in with two natural generalizations of Jeffrey conditioning: Wagner conditioning and the PME. In Section 3, I will introduce Jeffrey conditioning in a notation that will later help us to solve the obverse Majerník problem. In Section 4, I will introduce Wagner conditioning and show how it naturally generalizes Jeffrey conditioning. In Section 5, I will show that PME does so as well under conditions that are straightforward to accept for proponents of PME. This solves the obverse Majerník problem and makes Wagner conditioning unnecessary as a generalization of Jeffrey conditioning, since the PME seamlessly incorporates it. The conclusion in Section 6 summarizes my claims and briefly refers to epistemological consequences. An appendix gives proofs how PME generalizes standard conditioning and Jeffrey conditioning, providing a template for a simplified proof of the claim in the body of the paper.

2. Jeffrey's Updating Principle and the Principle of Maximum Entropy

In his paper "Marginal Probability Distribution Determined by the Maximum Entropy Method" (see [21]), 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 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 Majerník's question, PME provides us with a unique and plausible answer (see Majerník'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 [2] reveal that the joint probabilities are the product of the marginal probabilities (see also [4, 1670]). Once the joint probabilities and the marginal probabilities are available, it is trivial to calculate the conditional probabilities.

It is important to note that these joint probabilities do not legislate independence, even though they allow it [4, 1670]. Mérouane Debbah and Ralf Müller correctly describe these joint probabilities as a model with as many degrees of freedom as possible, which leaves free degrees for correlation to exist or not [4, 1674]. This avoids the introduction of unjustified information [4, 1672], corresponding to the simple intuition behind PME: when updating your probabilities, waste no useful information and do not gain information unless the evidence compels you to gain it (see [30, 376], [12, 280], [35, 278], [4, 1685f], and [22, 186]). The principle comes with its own formal apparatus, not unlike probability theory itself: Shannon's information entropy [25], the Kullback-Leibler divergence (see [20], [19], [7, 308ff], [24, 262ff]), the use of Lagrange multipliers (see [7, 327ff], [24, 281], [2, 409ff]), and the log-inverse relationship between information and probability (see [17], [16], [15], [18]).

There is an older problem by Carl Wagner [31] which can be cast in similar terms as Majerník's. 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

4). Wagner solves it using a natural generalization of Jeffrey conditioning, which I will call Wagner conditioning. It is not based on PME, but on what I 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.

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 [28]; for detractors see [9]). Our interest in this paper 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.

This paper resists Wagner's conclusions and shows that 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 Majerník problem posed above.

3. 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 unusual notation, anticipating using my notation to solve Wagner's *Linguist* problem and to give a general solution for the obverse Majerník problem. Let Ω be a finite event space and $\{\theta_j\}_{j=1,\ldots,n}$ a partition of Ω. Let κ 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 Ω (the ω are basically a more coarsely grained partition than the θ). $\theta^*_{ij} = \emptyset$ if $\kappa_{ij} = 0$, $\theta^*_{ij} = \theta_j$ otherwise. Let β 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 α and $\hat{\alpha}$ corresponding to the prior and posterior probabilities for $\{\omega_i\}_{i=1,\ldots,m}$, respectively.

A Jeffrey-type problem is when β 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

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

The notation is more complicated than it needs to be for Jeffrey conditioning. In Section 5, however, I will take full advantage of it to present a generalization where the ω_i do not range over the θ_j. In the meantime, here is an example to illustrate (2).

> A token is pulled from a bag containing 3 yellow tokens, 2 blue tokens, and 1 purple token. You are colour blind and cannot distinguish between the blue and the purple token when you see it. When the token is pulled, it is shown to you in poor lighting and then obscured again. You come to the conclusion based on your observation that the probability that the pulled token is yellow is $1/3$ and that the probability that the pulled token is blue or purple is $2/3$. What is your updated probability that the pulled token is blue?

Let $P(\text{blue})$ be the prior subjective probability that the pulled token is blue and $\hat{P}(\text{blue})$ the respective posterior subjective probability. Jeffrey conditioning, based on JUP (which mandates, for example, that $\hat{P}(\text{blue}|\text{blue or purple}) = P(\text{blue}|\text{blue or purple})$) recommends

$$\begin{aligned}\hat{P}(\text{blue}) \quad &= \hat{P}(\text{blue}|\text{blue or purple})\hat{P}(\text{blue or purple}) + \\ &\quad \hat{P}(\text{blue}|\text{neither blue nor purple})\hat{P}(\text{neither blue nor purple}) \\ &= P(\text{blue}|\text{blue or purple})\hat{P}(\text{blue or purple}) = 4/9 \end{aligned} \tag{3}$$

In the notation of (2), the example is calculated with $\beta = (1/2, 1/3, 1/6)^t, \hat{\alpha} = (1/3, 2/3)^t$,

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

and yields the same result as (3) with $\hat{\beta}_2 = 4/9$.

4. Wagner Conditioning

Carl Wagner uses JUP (explained in more detail in [32]) 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 (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 [31, 252], and [27, 197].)

Let us call a problem of this type a Wagner-type problem. It is an instance of the more general obverse Majerník 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 Majerník problems seek the conditional probabilities as well, both of which I 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 in [5], which is quite different from our notational approach.

Wagner legitimately calls his solution a "natural generalization of Jeffrey conditioning" [31, 250]. There is, however, another natural generalization of Jeffrey conditioning, E.T. Jaynes' principle of maximum entropy in [10]. 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).

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. Although Jaynes' original claims are concerned with type I prior probabilities, this paper works on the assumptions of Jaynes' later work focusing on type II prior probabilities. Some distinguish between MAXENT, the synchronic rule, and *Infomin*, the diachronic rule. The understanding here is that both operate on type II prior probabilities: MAXENT considers uniform prior probabilities (however this uniformity may have arisen) and a set of synchronic constraints on them; *Infomin*, in a more standard sense of updating, considers type II prior probabilities that are not necessarily uniform and updates them given evidence represented as new (diachronic) constraints on acceptable posterior probability distributions. Some say that MAXENT and *Infomin* contradict each other, but I disagree and maintain that they are compatible. I will have to defer this problem to future work, but a core argument for compatibility is already accessible in [32].

One advantage of PME is that it works on the wide domain of updating problems where the evidence corresponds to an affine constraint (for affine constraints see [3]; for problems with evidence not in the form of affine constraints see [23]). Updating problems where standard conditioning and Jeffrey conditioning are applicable are a subset of this domain. Some partial information cases (using the moment(s) of a distribution as evidence), such as Bas van Fraassen's *Judy Benjamin* problem and Jaynes' *Brandeis Dice* problem, are not amenable to either standard conditioning or Jeffrey conditioning. PME generalizes Jeffrey conditioning (and, a fortiori, standard conditioning) and therefore absorbs JUP on the more narrow domain of problems that we can solve using Jeffrey conditioning (for a proof see the appendix, although it can also be gleaned from [1]).

Wagner's contention is that on the wider domain of problems where we must use Wagner conditioning (and which he does not cast in terms of affine constraints), JUP and 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 (see [6], [26], [24], [33, 270], [29], [8, 107]) to reinforce his conclusion that PME is not a principle on which we should rely in general.

5. A Natural Generalization of Jeffrey and Wagner Conditioning

In order to show how PME generalizes Jeffrey conditioning (in the appendix) and Wagner conditioning to boot, I use the notation that I 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 Ω, 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 $\mu_{ij} = P(\omega_i \cap \theta_j)$ and $\hat{\mu}_{ij} = \hat{P}(\omega_i \cap \theta_j)$.

Given the specific nature of Wagner-type problems, there are a few constraints on the triple $(\kappa, \beta, \hat{\alpha})$. The last row $(\mu_{mj})_{j=1,\ldots,n}$ is special because it represents the probability of ω_m, which is the negation of the events deemed possible after the observation. In the *Linguist* problem, for example, ω_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. κ 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 κ except that its last row are all 0s, and $\hat{\alpha}_m = 0$. Otherwise the 0s are distributed over κ (and equally over $\hat{\kappa}$) so that no row and no column has all 0s, representing the logical relationships between the ω_is and the θ_js ($\kappa_{ij} = 0$ if and only if $\hat{P}(\omega_i \cap \theta_j) = \mu_{ij} = 0$). We set $P(\omega_m) = x$ ($\hat{P}(\omega_m) = 0$), where x depends on the specific prior knowledge. Fortunately, the value of x cancels out nicely and will play no further role. For convenience, we define

$$\zeta = (0, \ldots, 0, 1)^t \tag{5}$$

with $\zeta_m = 1$ and $\zeta_i = 0$ for $i \neq m$.

The best way to visualize such a problem is by providing the joint probability matrix $M = (\mu_{ij})$ together with the marginals α and β in the last column/row, here for example as for the *Linguist* problem with $m = 5$ and $n = 4$ (note that this is not the matrix M, which is $m \times n$, but M expanded with the marginals in improper matrix notation):

$$\begin{bmatrix} \mu_{11} & \mu_{12} & 0 & 0 & \alpha_1 \\ \mu_{21} & \mu_{22} & 0 & 0 & \alpha_2 \\ 0 & \mu_{32} & 0 & \mu_{34} & \alpha_3 \\ \mu_{41} & \mu_{42} & \mu_{43} & \mu_{44} & \alpha_4 \\ \mu_{51} & \mu_{52} & \mu_{53} & \mu_{54} & x \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & 1.00 \end{bmatrix}. \tag{6}$$

The $\mu_{ij} \neq 0$ 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 $(\kappa, \beta, \hat{\alpha})$,

$$\kappa = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } \hat{\kappa} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{7}$$

$$\beta = (0.2, 0.3, 0.4, 0.1)^t \text{ and } \hat{\alpha} = (0.4, 0.3, 0.2, 0.1, 0)^t. \tag{8}$$

Wagner's solution, based on JUP, is

$$\hat{\beta}_j = \beta_j \sum_{i=1}^{m-1} \frac{\hat{\kappa}_{ij}\hat{\alpha}_i}{\sum_{\hat{\kappa}_{il}=1} \beta_l} \text{ for all } j = 1, \ldots, n. \tag{9}$$

In numbers,

$$\hat{\beta}_j = (0.3, 0.6, 0.04, 0.06)^t. \tag{10}$$

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. The solution that Wagner attributes to PME is misleading because of Wagner's Dempsterian setup which does not take into account that proponents of PME are likely to be proponents of the classical Bayesian position that type II prior probabilities are specified and determinate once the agent attends to the events in question. Some Bayesians in the current discussion explicitly disavow this requirement for (possibly retrospective) determinacy (especially James Joyce in [14] and other papers). Proponents of PME (a proper subset of Bayesians), however, are unlikely to follow Joyce—if they did, they would indeed have to address Wagner's example to show that their allegiances to PME and to indeterminacy are compatible.

That (9) follows from JUP is well-documented in Wagner's paper. For the PME solution for this problem, I will not use (9) 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 (9). This completes our argument.

What follows may only be accessible to PME cognoscenti, since it involves the Lagrange multiplier method (see [7, 327ff] and [11, 244]). Others may read the conclusion and find a sketch for an easier, but much less rigorous proof in the appendix. To maximize the Shannon entropy of M and minimize the Kullback-Leibler divergence between $\hat{M}$ and M, consider the Lagrangian functions:

$$\Lambda(\mu_{ij}, \xi) = \sum_{\kappa_{ij}=1} \mu_{ij} \log \mu_{ij} + \sum_{j=1}^{n} \xi_j \left(\beta_j - \sum_{\kappa_{kj}=1} \mu_{kj} \right) + \lambda_m \left(x - \sum_{j=1}^{n} \mu_{mj} \right) \tag{11}$$

and

$$\hat{\Lambda}(\hat{\mu}_{ij}, \hat{\lambda}) = \sum_{\hat{\kappa}_{ij}=1} \hat{\mu}_{ij} \log \frac{\hat{\mu}_{ij}}{\mu_{ij}} + \sum_{i=1}^{m} \hat{\lambda}_i \left(\hat{\alpha}_i - \sum_{\hat{\kappa}_{il}=1} \hat{\mu}_{il} \right). \tag{12}$$

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

$$M = rs^t \circ \kappa \tag{13}$$

$$\hat{M} = \hat{r}s^t \circ \hat{\kappa} \tag{14}$$

$$\beta = S\kappa^t r \tag{15}$$

$$\hat{\alpha} = \hat{R}\kappa s \tag{16}$$

where $r_i = e^{\zeta_i \lambda_m}, s_j = e^{-1-\xi_j}, \hat{r}_i = e^{-1-\hat{\lambda}_i}$ represent factors arising from the Lagrange multiplier method (ζ was defined in (5)). 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 $R_{il} = r_i\delta_{il}, S_{kj} = s_j\delta_{kj}, \hat{R}_{il} = \hat{r}_i\delta_{il}$ (δ is Kronecker delta).

Note that

$$\frac{\beta_j}{\sum_{\hat{\kappa}_{il}=1} \beta_l} = \frac{s_j}{\sum_{\hat{\kappa}_{il}=1} s_l} \text{ for all } (i,j) \in \{1,\ldots,m-1\} \times \{1,\ldots,n\}. \tag{17}$$

(16) implies

$$\hat{r}_i = \frac{\hat{\alpha}_i}{\sum_{\hat{\kappa}_{il}=1} s_l} \text{ for all } i = 1,\ldots,m-1. \tag{18}$$

Consequently,

$$\hat{\beta}_j = s_j \sum_{i=1}^{m-1} \frac{\hat{\kappa}_{ij}\hat{\alpha}_i}{\sum_{\kappa_{il}=1} s_l} \text{ for all } j = 1,\ldots,n. \tag{19}$$

(19) gives us the same solution as (9), taking into account (17). Therefore, Wagner conditioning and PME agree.

6. Conclusion

Wagner-type problems (but not obverse Majerník-type problems) can be solved using JUP and Wagner's ad hoc method. Obverse Majerník-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 it rests on assumptions that proponents of PME naturally reject.

A. Appendix: PME generalizes Jeffrey Conditioning

A proof that PME generalizes standard conditioning is in [34]. A proof that PME generalizes Jeffrey conditioning is in [1]. I will give my own simple proofs here that are more in keeping with the notation in the paper. An interested reader can also apply these proofs to show that PME generalizes Wagner conditioning, but not without simplifications that compromise mathematical rigour. The more rigorous proof for the generalization of Wagner conditioning is in the body of the paper.

I assume finite (and therefore discrete) probability distributions. For countable and continuous probability distributions, the reasoning is largely analogous (for an introduction to continuous entropy see [7, 16ff]; for an example of how to do a proof of this section for continuous probability densities see [1, 11]; for a proof that the stationary points of the Lagrange function are indeed the desired extrema see [36, 55] and [2, 410]; for the pioneer of the method applied in this section see [11, 241ff]).

A.1. Standard Conditioning

Let y_i (all $y_i \neq 0$) be a finite type II prior probability distribution summing to 1, $i \in I$. Let $\hat{y}_i$ be the posterior probability distribution derived from standard conditioning with $\hat{y}_i = 0$ for all $i \in I'$ and $\hat{y}_i \neq 0$ for all $i \in I''$, $I' \cup I'' = I$. I' and I'' specify the standard event observation. Standard conditioning requires that

$$\hat{y}_i = \frac{y_i}{\sum_{k \in I''} y_k}. \tag{20}$$

To solve this problem using PME, we want to minimize the cross-entropy with the constraint that the non-zero $\hat{y}_i$ sum to 1. The Lagrange function is (writing in vector form $\hat{y} = (\hat{y}_i)_{i \in I''}$)

$$\Lambda(\hat{y}, \lambda) = \sum_{i \in I''} \hat{y}_i \ln \frac{\hat{y}_i}{y_i} + \lambda \left(1 - \sum_{i \in I''} \hat{y}_i \right). \tag{21}$$

Differentiating the Lagrange function with respect to $\hat{y}_i$ and setting the result to zero gives us

$$\hat{y}_i = y_i e^{\lambda - 1} \tag{22}$$

with λ normalized to

$$\lambda = -1 + \ln \sum_{i \in I''} y_i. \tag{23}$$

(20) follows immediately. PME generalizes standard conditioning.

A.2. Jeffrey Conditioning

Let $\theta_i, i = 1, \ldots, n$ and $\omega_j, j = 1, \ldots, m$ be finite partitions of the event space with the joint prior probability matrix (y_{ij}) (all $y_{ij} \neq 0$). Let κ be defined as in Section 3, with (1) true (remember that in Section 5, (1) is no longer required). Let P be the type II prior probability distribution and $\hat{P}$ the posterior probability distribution.

Let $\hat{y}_{ij}$ be the posterior probability distribution derived from Jeffrey conditioning with

$$\sum_{i=1}^{n} \hat{y}_{ij} = \hat{P}(\omega_j) \text{ for all } j = 1, \ldots, m \tag{24}$$

Jeffrey conditioning requires that for all $i = 1, \ldots, n$

$$\hat{P}(\theta_i) = \sum_{j=1}^{m} P(\theta_i|\omega_j)\hat{P}(\omega_j) = \sum_{j=1}^{m} \frac{y_{ij}}{P(\omega_j)}\hat{P}(\omega_j) \tag{25}$$

Using PME to get the posterior distribution $(\hat{y}_{ij})$, the Lagrange function is (writing in vector form $\hat{y} = (x_{11}, \ldots, x_{n1}, \ldots, x_{nm})^t$ and $\lambda = (\lambda_1, \ldots, \lambda_m)^t$)

$$\Lambda(\hat{y}, \lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{y}_{ij} \ln \frac{\hat{y}_{ij}}{y_{ij}} + \sum_{j=1}^{m} \lambda_j \left(\hat{P}(\omega_j) - \sum_{i=1}^{n} \hat{y}_{ij} \right). \tag{26}$$

Consequently,

$$\hat{y}_{ij} = y_{ij} e^{\lambda_j - 1} \tag{27}$$

with the Lagrangian parameters λ_j normalized by

$$\sum_{i=1}^{n} y_{ij} e^{\lambda_j - 1} = \hat{P}(\omega_j) \tag{28}$$

(25) follows immediately. PME generalizes Jeffrey conditioning.

B. References

[1] Caticha, Ariel, and Adom Giffin. "Updating Probabilities." In *Max-Ent 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods*. 2006.

[2] Cover, T.M., and J.A. Thomas. *Elements of Information Theory*, volume 6. Hoboken, NJ: Wiley, 2006.

[3] Csiszár, Imre. "Information-Type Measures of Difference of Probability Distributions and Indirect Observations." *Studia Scientiarum Mathematicarum Hungarica* 2: (1967) 299–318.

[4] Debbah, Mérouane, and Ralf Müller. "MIMO Channel Modeling and the Principle of Maximum Entropy." *IEEE Transactions on Information Theory* 51, 5: (2005) 1667–1690.

[5] Dempster, Arthur. "Upper and Lower Probabilities Induced by a Multi-valued Mapping." *Annals of Mathematical Statistics* 38, 2: (1967) 325–339.

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

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

[8] Halpern, Joseph. *Reasoning About Uncertainty*. Cambridge, MA: MIT, 2003.

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

[10] Jaynes, E.T. "Information Theory and Statistical Mechanics." *Physical Review* 106, 4: (1957) 620–630.

[11] 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.

[12] Jaynes, E.T. "Optimal Information Processing and Bayes's Theorem: Comment." *American Statistician* 42, 4: (1988) 280–281.

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

[14] Joyce, James. "A Defense of Imprecise Credences in Inference and Decision Making." *Philosophical Perspectives* 24, 1: (2010) 281–323.

[15] Kampé de Fériet, J., and B. Forte. "Information et probabilité." *Comptes rendus de l'Académie des sciences* A 265: (1967) 110–114.

[16] Ingarden, R. S., and K. Urbanik. "Information Without Probability." *Colloquium Mathematicum* 9: (1962) 131–150.

[17] Khinchin, Aleksandr. *Mathematical Foundations of Information Theory*. New York: Dover, 1957.

[18] Kolmogorov, Andrey. "Logical Basis for Information Theory and Probability Theory." *IEEE Transactions on Information Theory* 14, 5: (1968) 662–664.

[19] Kullback, Solomon. *Information Theory and Statistics*. London, UK: Dover, 1959.

[20] Kullback, Solomon and Leibler, Richard. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22, 1: (1951) 79–86.

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

[22] Palmieri, Francesco, and Domenico Ciuonzo. "Objective Priors from Maximum Entropy in Data Classification." *Information Fusion* 14, 2: (2013) 186–198.

[23] Paris, Jeff. *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge, UK: Cambridge University, 2006.

[24] Seidenfeld, Teddy. "Entropy and Uncertainty." In *Advances in the Statistical Sciences: Foundations of Statistical Inference*, Springer, 1986, 259–287.

[25] Shannon, Claude. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27, 1: (1948) 379–423, 623–656.

[26] Skyrms, Brian. "Updating, Supposing, and Maxent." *Theory and Decision* 22, 3: (1987) 225–246.

[27] Spohn, Wolfgang. *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford, UK: Oxford University, 2012.

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

[29] Uffink, Jos. "Can the Maximum Entropy Principle Be Explained as a Consistency Requirement?" *Studies in History and Philosophy of Science* 26, 3: (1995) 223–261.

[30] Van Fraassen, Bas, R.I.G. Hughes, and Gilbert Harman. "A Problem for Relative Information Minimizers, Continued." *British Journal for the Philosophy of Science* 37, 4: (1986) 453–463.

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

316 [32] Wagner, Carl. "Probability Kinematics and Commutativity." *Philosophy of Science* 69, 2: (2002)
317 266–278.
318 [33] Walley, Peter. *Statistical Reasoning with Imprecise Probabilities*. London, UK: Chapman and
319 Hall, 1991.
320 [34] Williams, Paul. "Bayesian Conditionalisation and the Principle of Minimum Information."
321 *British Journal for the Philosophy of Science* 31, 2: (1980) 131–144.
322 [35] Zellner, Arnold. "Optimal Information Processing and Bayes's Theorem." *American Statistician*
323 42, 4: (1988) 278–280.
324 [36] Zubarev, Dmitrii, Vladimir Morozov, and Gerd Röpke. *Statistical Mechanics of Nonequilibrium*
325 *Processes*. Berlin: Akademie, 1996.

326 Conflicts of Interest

327 The author declares no conflict of interest.