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. The generalization rests on an ad hoc but plausible intuition. Wagner shows how the principle of maximum entropy disagrees with this intuition, thus casting further doubt on the principle of maximum entropy as a generally valid updating mechanism that generalizes Jeffrey conditioning. This article demonstrates that Wagner's application of the principle of maximum entropy is incorrect and that a correct application agrees with his intuition. It presents a formal proof that the principle of maximum entropy seamlessly and elegantly generalizes not only standard conditioning and Jeffrey conditioning (as is well-documented in the literature) but also Wagner's generalization.

1 Introduction

There are problems of probability update which cannot be addressed effectively by means of standard conditioning or Jeffrey conditioning. Usually, the evidence (or observation, or new information) arises in the form of an event (standard conditioning) or the redistribution of probabilities over a complete partition of events (Jeffrey conditioning). Sometimes the evidence arises in the form of a constraint which falls into neither of the above categories. Bas van Fraassen's *Judy Benjamin* problem and E.T. Jaynes' *Brandeis Dice* problem are two examples.

To solve these cases, Jaynes has suggested the principle of maximum entropy, which extends the idea of optimal information processing (which standard conditioning and Jeffrey conditioning obey) to a larger class of constraints. The principle of maximum entropy was originally developed as a synchronic norm (call this synchronic norm MAXENT), where constraints were coordinated with non-informative prior probabilities to result in a probabilities distribution or density which was maximally entropic (using Shannon's information entropy) but at the same time obeyed all the constraints.

Soon, however, the principle of maximum entropy was applied to problems of probability update where the word 'prior' is used comparatively and refers to a probability distribution which precedes a piece of new information and therefore the posterior probability distribution; it is not used superlatively

and does not refer to a probability distribution which precedes any information at all or has any ambitions to be non-informative. The principle of maximum entropy was used to articulate a diachronic norm (call this diachronic norm *Infomin*), where we are concerned with what Richard Jeffrey terms 'probability kinematics,' the rules or guidelines when moving from a given prior probability distribution to a posterior probability distribution in the wake of new information (using the Kullback-Leibler divergence).

Sometimes the case is made that MAXENT and *Infomin* are two different norms prescribing different probability distributions in certain cases. 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, would ignore the diachronic dimension and prescribe 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 would be $(\frac{1}{2}, \frac{1}{5}, \frac{3}{10})$. *Infomin*, on the other hand, would take as its prior probability distribution $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ and then diachronically update to $(\frac{8}{15}, \frac{1}{5}, \frac{4}{15})$.

While it is helpful to distinguish between synchronic and diachronic norms ... xxyxx

While MAXENT initially attracted attention and was confirmed especially by the work of Shore and Johnston (MAXENT uniquely solves probability update problems provided one signs on to relatively intuitive axioms, see Shore and Johnson, 1980), there were vexing counterexamples. Although there is formal proof that MAXENT generalizes standard conditioning and Jeffrey conditioning, there are problems where standard conditioning and Jeffrey conditioning do not apply.

The consensus emerged in the 1980s that MAXENT was a helpful algorithm, but not a generally valid rule of probability update. There are dissenting voices to this day, mostly among statistical physicists, but the consensus has been strong enough to be incorporated in important textbooks.

- Brian Skyrms states, "MAXENT is not a generally valid updating rule" (Skyrms, 1987b, 237), based primarily on the counterexample provided by Abner Shimony (for example in Friedman and Shimony, 1971, Dias and Shimony, 1981, and Shimony, 1985). Skyrms makes this view known in several articles (see also Skyrms, 1985 and Skyrms, 1987a).

In his textbook *The Dynamics of Rational Deliberation* (1990), there is a section on probability kinematics, but no reference to MAXENT.

- Joseph Halpern argues in his textbook *Reasoning About Uncertainty* (2003) that MAXENT is a promising candidate which delivers unique updated probability distributions, but there is counterintuitive behaviour in one specific case, the *Judy Benjamin* case (see Halpern, 2003, 110, 119). Therefore, MAXENT is a valuable tool that should be used with care and not be applied across the board (see Grove and Halpern, 1997, 110).

- In his textbook about ranking functions, called *The Laws of Belief* (2012), Wolfgang Spohn admiringly refers to Wagner's "generalization of Jeffrey conditionalization" (Spohn, 2012, 41) and shows how well it accords with his own conditions for the dynamics of belief in terms of ranking functions (see Spohn, 2012, 196ff). Spohn cites Jeffrey's *Linguist* problem at length and points out that MAXENT, another generalization of Jeffrey conditioning, is not compatible with Wagner's (and, we are led to conclude, also not compatible with Spohn's theory of ranking functions).

We will consider the above three counterexamples to be the main arguments against MAXENT in practice (there are also more theoretical objections, such as Skyrms' claim that MAXENT is not an inductive rule but rather a rule for supposing; or the widely shared non-Bayesian view, for example by Spohn and Halpern, that numerical probabilities are even formally not always the most helpful way to represent beliefs; or alleged incompatibilities between epistemic entrenchment and MAXENT—all of these would need to be addressed for a more systematic defence of MAXENT against its competitors). Here is a fourth example from a textbook, however, which underlines the influence that the three counterexamples above have had in the literature.

- Without citing or engaging any one of the counterexamples, David MacKay writes in his textbook *Information Theory, Inference, and Learning Algorithms* (2003), "maximum entropy is also sometimes proposed as a method for solving inference problems [...] I think it is a bad idea to use maximum entropy in this way; it can give very silly answers" (MacKay, 2003, 308).

Against the tide of scholarly consensus, I maintain that MAXENT is defensible across all counterexamples and survives as a unifying normative principle

in probability update based on a very simple intuition: that one should not illegitimately gain information where the evidence does not provide it, and one should not refuse to incorporate available information in updated beliefs. Bayes' theorem works because it conforms to MAXENT, not vice versa. I have set myself the task of responding to Wagner's *Linguist* counterexample in this paper as part of a systematic effort to revive interest among philosophers in MAXENT as an objective updating method.

Wagner's counterexample is in many ways the easiest counterexample to respond to, as it is based on some fundamental misunderstandings and as MAXENT provides an elegant solution if properly applied. As such a response is not available in the literature, however, it is worth spelling out here, especially in view of the fact that Wagner's counterexample is used in textbooks to support the scholarly consensus that MAXENT is not a generally valid updating rule.

2 Wagner's Natural Generalization of Jeffrey Conditioning

Wagner claims that he has found a relatively common case of probability update in which MAXENT delivers the wrong result so that we must develop an ad hoc generalization of Jeffrey conditioning. This is best explained by example. Wagner refers to Richard Jeffrey's *Linguist* problem (see Jeffrey, 1990).

> *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 ad hoc 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 (often, we will write lower case p or q for the marginal probabilities without spelling out the margin to which they refer; we will write upper case P and Q for the joint probabilities). 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. 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. Other joint probability distributions would use information not provided in the problem or discount information that is provided in the problem (especially the conditionals). 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, although we are surprised about these contortions when a Bayesian approach, combined with MAXENT, gives us exactly the same results. Wagner's paper is a paradigmatic example for the 'anti-Bayesian ad hockeries' addressed in E.T. Jaynes' diatribe (see Jaynes and Bretthorst, 1998, 143). In Wagner's case, however, the motivation of the anti-Bayesian (over which Jaynes despairs) is on the surface: Wagner is mistaken about the correct application of MAXENT, and so he considers the ad hockery necessary to come to an acceptable result, which his incorrect application of MAXENT clearly does not provide.

When Wagner considers MAXENT towards the end of his article, his verdict is scathing.

> 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)

Here is what we will do to demonstrate that Wagner's conclusions are off-target. First, we will focus on *Linguist* and show that Wagner's solution is plausible. Then we will introduce what Wagner considers to be the solution of MAXENT for this problem and show, in much greater detail than Wagner does, why this result is implausible. Next, we will uncover the fundamental mistake that Wagner makes in his application of MAXENT. In summary, Wagner forgets that advocates of MAXENT are Bayesians and do not buy into Wagner's apparently non-Bayesian convictions. Non-Bayesian assumptions and MAXENT result in counterintuitive conclusions. Once the non-Bayesian assumptions are replaced by Bayesian assumptions (for example, that probabilities represent the uncertainty of those who hold them and not objective probabilities out there in the world), not only do the conclusions become plausible, they also (luckily for Wagner and, by extension, for Spohn's theory of ranking functions) agree with Wagner's solution. Hence, the patchwork solution is unnecessary and only provides intuitive support for MAXENT. We will top off our considerations with a more general theorem which seamlessly incorporates Wagner's "natural generalization of Jeffrey conditionalization" (Wagner, 1992, 250) into MAXENT orthodoxy.

3 The Linguist

According to Wagner (thus the index w in P_w), the prior probabilities for *Linguist* are as follows:

P_w	ω_1	ω_2	ω_3	ω_4	P_Θ
θ_1	?	?	?	?	0.20
θ_2	?	?	?	?	0.30
θ_3	?	?	?	?	0.40
θ_4	?	?	?	?	0.10
P_Ω	0.40	0.30	0.20	0.10	1.00

In Wagner's diction, we are not possessed of the joint probability measure P on $\Theta \times \Omega$, only of the marginal probabilities. This betrays Wagner's non-

Bayesian convictions. A Bayesian would determine these probabilities as representations of prior assumptions or lack of information. Even though many Bayesians think that they are not objectively determined, they never leave prior probabilities unspecified as if it were a matter of epistemological access to objective probabilities to specify them. Much more about this later. Wagner's posterior probabilities Q_w (in contrast to MAXENT's posterior probabilities Q_m later on) for *Linguist* are as follows:

Q_w	ω_1	ω_2	ω_3	ω_4	Q_Θ
θ_1	?	?	0.00	?	0.30
θ_2	?	?	?	?	0.60
θ_3	0.00	0.00	0.00	0.04	0.04
θ_4	0.00	0.00	?	?	0.06
Q_Ω	0.40	0.30	0.20	0.10	1.00

(3) dictates the zero joint probabilities, (4) dictates the marginal probabilities in the last row, and (6) dictates the marginal probabilities in the last column. The posterior probability that the native encountered by the linguist is a northerner, for example, is 34%.

To solve this problem from the perspective of MAXENT, Wagner assumes that the constraint is 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 $b(E_{12})$. In keeping with the idea of MAXENT, we choose from all probability distributions fulfilling the constraint the one which is information-theoretically closest to the prior probability distribution.

Wagner applies this idea to the marginal probability distribution on Θ. He does not provide the numbers, but refers to other examples where b is already a lower bound for the prior probability distribution to make his point that MAXENT does not generally agree with his solution. To aid the discussion, I want to make Wagner's claims more concrete. Using proposition 1.29 in Dimitri Bertsekas' book *Constrained Optimization and Lagrange Multiplier Methods* (see Bertsekas, 1982, 71) and some non-trivial calculations, the MAXENT solution for *Linguist*, given Wagner's assumptions (indexed $Q_{w/m}$) is

$Q_{w/m}$	ω_1	ω_2	ω_3	ω_4	Q_Θ
θ_1	?	?	0.00	?	0.30
θ_2	?	?	?	?	0.45
θ_3	0.00	0.00	0.00	0.10	0.10
θ_4	0.00	0.00	?	?	0.15
Q_Ω	0.40	0.30	0.20	0.10	1.00

$D_{\mathrm{KL}}(Q_{w/m}, P) \approx 0.0823$ is indeed significantly smaller than $D_{\mathrm{KL}}(Q_w, P) \approx 0.4148$. In both cases, $Q_{w/m}$ and Q_w are first marginalized to Θ in order to calculate the Kullback-Leibler Divergence

$$D_{\mathrm{KL}}(q, p) = \sum_{i=1}^{4} q(\theta_i) \log_2 \frac{q(\theta_i)}{p(\theta_i)}. \tag{7}$$

From the perspective of a MAXENT advocate, there are only two explanations for this difference in cross-entropy. Either Wagner's solution illegitimately uses information not contained in the problem, or Wagner's MAXENT solution has failed to include information contained in the problem. I will radically simplify *Linguist* 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 solution and, surely, the correct solution). Using b as a lower bound and MAXENT, Wagner's MAXENT solution for this radically simplified problem is 60:40, clearly a more entropic solution than Wagner's. The problem, of course, is that Wagner's MAXENT solution is both counter-intuitive and false. It is our task to find out why Wagner's interpretation of how MAXENT should proceed is misguided.

Wagner's core mistake is that for a Bayesian, the prior joint probability distribution on $\Theta \times \Omega$ is not left unspecified. It represents the subjective priors or the lack of information on part of the agent who holds these probabilities. Whether it is the former or the latter is a matter of considerable debate among 'subjectivist' and 'objectivist' Bayesians (for example, Bruno

de Finetti and E.T. Jaynes), both of which, confusingly, are adherents of interpreting probabilities as subjective probabilities.

This debate aside, all Bayesians agree that well-defined events have prior probabilities. Advocates of MAXENT, who form a strict subset of the set of all Bayesians, do not only believe (as Bayesians do and non-Bayesians do not) that the prior joint probability distribution is numerically populated in accordance with basic probability axioms; they also believe that the prior joint probability distribution is numerically populated with the probabilities that have the highest entropy compatible with the available information (in our case, the marginal probability distributions).

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

Advocates of MAXENT, who are objectivist Bayesians, think that determining these probabilities is not solely a matter of consistency with probability axioms. If the marginal probabilities are available, then the joint distribution is equal to the product of the marginal distributions

$$P(\theta_i, \omega_j) = p(\theta_i)p(\omega_j) \text{ for all } i = 1, \ldots, n \text{ and } j = 1, \ldots, m \tag{8}$$

to achieve maximum entropy (for a proof, see exercise 12.4 in Cover and Thomas, 2006, 421).

Therefore, the prior probability table for *Linguist*, from a MAXENT Bayesian's perspective, looks as follows:

P_m	ω_1	ω_2	ω_3	ω_4	Q_Θ
θ_1	0.08	0.06	0.04	0.02	0.20
θ_2	0.12	0.09	0.06	0.03	0.30
θ_3	0.16	0.12	0.08	0.04	0.40
θ_4	0.04	0.03	0.02	0.01	0.10
Q_Ω	0.40	0.30	0.20	0.10	1.00

Given the constraints (3) and (4) and a formal result provided in the next section, the MAXENT solution minimizing the cross-entropy to this prior probability distribution is

Q_m	ω_1	ω_2	ω_3	ω_4	Q_Θ
θ_1	0.16	0.12	0.00	0.02	0.30
θ_2	0.24	0.18	0.15	0.03	0.60
θ_3	0.00	0.00	0.00	0.04	0.04
θ_4	0.00	0.00	0.05	0.01	0.06
Q_Ω	0.40	0.30	0.20	0.10	1.00

Q_m agrees with Q_w in the marginalization to Θ (the last column). Consequently, MAXENT delivers a reasonable solution conforming to the ad hoc intuitive condition imposed by Wagner. The solution that Wagner foists on MAXENT is simply a victim of coarsening at random (see Grünwald and Halpern, 2003, who show how coarsening at random is responsible for misguided intuitions in the *Monty Hall* and the *Three Prisoners* problem). Just as the misguided interpretation of the *Monty Hall* problem operates on a coarsening of the event space, Wagner's MAXENT solution operates on a coarsening of the event space, which then fails to process the totality of the information provided in the wording of the problem. Wagner's MAXENT solution uses a probability distribution that is unnecessarily coarse—a more finely grained prior probability distribution delivers the right results.

The calculations for *Linguist* are awkward, because the 4×4 joint probability matrix is too large to deal with on the back of a napkin (and handling constraints on upper and lower probabilities using Wagner's faulty assumptions is much more complicated than handling constraints on joint probability distributions as shown in the next section). Using my simplified linguist problem above (where we only need to attend to a 2×2 matrix) makes the point just as beautifully, and again the MAXENT solution concurs both with Wagner's condition and with our intuitions: the posterior probability that the native is a Protestant is 80% versus a 20% probability that he or she is a Catholic.

Here are the relevant tables for the simplified linguist problem (θ'_1 means that the native is Catholic; θ'_2 that the native is Protestant; ω'_1 means that the native's utterance excludes the possibility that the native is Catholic; ω'_2 means that the native's utterance provides no information about the native's religious affiliation). As before, P_w is Wagner's prior probability distribution (with lacunae), Q_w his posterior probability distribution. $Q_{w/m}$ is Wagner's (faulty) interpretation of MAXENT, based on P_w. P_m and Q_m are the correct versions of applying MAXENT to the simplified linguist problem. The prior probability distributions are prior in the sense that they are prior to the information about what the two utterances entail.

P_w	ω'_1	ω'_2	$P_{\Theta'}$		Q_w	ω'_1	ω'_2	$Q_{\Theta'}$
θ'_1	?	?	0.50		θ'_1	0.60	0.20	0.80
θ'_2	?	?	0.50		θ'_2	0.00	0.20	0.20
$P_{\Omega'}$	0.60	0.40	1.00		$Q_{\Omega'}$	0.60	0.40	1.00
					$Q_{w/m}$	ω'_1	ω'_2	$Q_{\Theta'}$
					θ'_1	0.60	0.00	0.60
					θ'_2	0.00	0.40	0.40
					$Q_{\Omega'}$	0.60	0.40	1.00
P_m	ω'_1	ω'_2	$P_{\Theta'}$		Q_m	ω'_1	ω'_2	$Q_{\Theta'}$
θ'_1	0.30	0.20	0.50		θ'_1	0.60	0.20	0.80
θ'_2	0.30	0.20	0.50		θ'_2	0.00	0.20	0.20
$P_{\Omega'}$	0.60	0.40	1.00		$Q_{\Omega'}$	0.60	0.40	1.00

As in *Linguist*, Q_w and Q_m agree on the solution of the simplified linguist problem where Q_w has specified probabilities, especially in the margins where all of Q_w's probabilities are specified in both problems.

4 MAXENT is the Natural Generalization of Jeffrey Conditioning

We will show, using Lagrange multipliers, that MAXENT conforms to the intuition voiced by Wagner that the prior probability distribution remains operative in the assessment of relative uncertainties with respect to the consequents of observed conditionals. Wagner denies that MAXENT conforms to this intuition, thus disparaging MAXENT, but we have shown in the previous section that Wagner's judgment relies on an interpretation of MAXENT in terms of non-Bayesian assumptions, which is nonsensical as MAXENT is a more specific version of Baysianism.

We will show that MAXENT conforms to the intuitions behind standard conditioning; then that it conforms to the intuitions behind Jeffrey conditioning; and finally that it conforms to Wagner's ad hoc 'natural generalization of Jeffrey conditioning.' Throughout, to make formalities comprehensible for non-mathematical folk, we will refer to finite (and therefore discrete) probability distributions. For countable and continuous probability distributions, the reasoning is largely analogous (for a mathematically rigorous introduction to continuous entropy see Guiaşu, 1977, 16ff; for an example of how to

do a proof of this section for continuous probability densities see Caticha and Giffin, 2006, 11; for a proof that the stationary points of the Lagrange function are indeed the desired extrema see Zubarev et al., 1996, 55 and Cover and Thomas, 2006, 410; for the pioneer of the method applied in this section see Jaynes, 1979, 241ff).

Standard Conditioning

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

$$x_i = \frac{y_i}{\sum_{k \in I''} y_k}. \tag{9}$$

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

$$\Lambda(x, \lambda) = \sum_{i \in I''} x_i \ln \frac{x_i}{y_i} + \lambda \left(1 - \sum_{i \in I''} x_i \right). \tag{10}$$

Differentiating the Lagrange function with respect to x_i and setting the result to zero gives us

$$x_i = y_i e^{\lambda - 1} \tag{11}$$

with λ normalized to

$$\lambda = -1 + \ln \sum_{i \in I''} y_i. \tag{12}$$

(9) follows immediately. MAXENT and standard conditioning are consistent with each other.

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 x_{ij} be the posterior probability distribution derived from Jeffrey conditioning with

$$\sum_{i=1}^{n} x_{ij} = \alpha_j \text{ for all } j = 1, \ldots, m \tag{13}$$

where the $\alpha_j \neq 0, \sum \alpha_j = 1$ are the observed redistribution of the marginal probability for ω_j. Jeffrey conditioning requires that for all $i = 1, \ldots, n$

$$q(\theta_i) = \sum_{j=1}^{m} p(\theta_i|\omega_j)q(\omega_j) = \sum_{j=1}^{m} \frac{y_{ij}}{p(\omega_j)} q(\omega_j) \tag{14}$$

where p and q are the marginal probabilities for $(\theta_i)_{i=1,\ldots,n}$ and $(\omega_j)_{j=1,\ldots,m}$, prior and posterior respectively (just as P and Q are the joint probabilities on $\Theta \times \Omega$), for example $\alpha_j = q(\omega_j)$, $P(\theta_i, \omega_j) = y_{ij}$ and

$$q(\theta_i) = \sum_{j=1}^{m} x_{ij}. \tag{15}$$

Using MAXENT to get the posterior distribution (x_{ij}), the Lagrange function is (writing in vector form $\lambda = (\lambda_1, \ldots, \lambda_m)$ and $x = (x_{11}, \ldots, x_{n1}, \ldots, x_{nm})$)

$$\Lambda(x, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \ln \frac{x_{ij}}{y_{ij}} + \sum_{j=1}^{m} \lambda_j \left(\alpha_j - \sum_{i=1}^{n} x_{ij} \right). \tag{16}$$

Consequently,

$$x_{ij} = y_{ij} e^{\lambda_j - 1} \tag{17}$$

with the Lagrangian parameters λ_j normalized by

$$\sum_{i=1}^{n} y_{ij} e^{\lambda_j - 1} = \alpha_j \tag{18}$$

(14) follows immediately. MAXENT and Jeffrey conditioning are consistent with each other.

Wagner Conditioning

Let $\theta_i, i = 1, \ldots, n$ and $\omega_j, j = 1, \ldots, m$ be finite partitions of the event space. Let the joint prior probability matrix be (y_{ij}) and the joint posterior probability matrix be (x_{ij}). The elements of this matrix may be unknown or may have to be inferred from one's state of ignorance or uncertainty or may be chosen any other way according to basic probability axioms and your favourite interpretation of probability. The marginal probabilities are $p(\theta_i), p(\omega_j), q(\theta_i), q(\omega_j)$ so that, for example,

$$p(\theta_i) = \sum_{j=1}^{m} y_{ij} \tag{19}$$

if the pertinent y_{ij} exist. According to (6), Wagner conditioning determines the posterior marginal probability to be

$$q(\theta_i) = \sum_{E \in \mathcal{E}} m(E) p(\theta_i | E) \text{ for all } i = 1, \ldots, n \tag{20}$$

given the conditions in (3) and (4). (3) means for Q that depending on the observed conditionals there is a partition of the indices $\{1, \ldots, n\} \times \{1, \ldots, m\}$ into sets K' and K'' such that $Q(\theta_i, \omega_j) = 0$ (i.e. $x_{ij} = 0$) for $(i, j) \in K'$ and $Q(\theta_i, \omega_j) \neq 0$ (i.e. $x_{ij} \neq 0$) for $(i, j) \in K''$.

We want to show that MAXENT comes to the same conclusion as Wagner conditioning, provided that the probability matrices $(y_{ij}), (x_{ij})$ are fully quantified—to which we commit ourselves since MAXENT only makes sense in

a Bayesian framework. Because the matrices are fully quantified, we rewrite (20) as

$$q(\theta_i) = p(\theta_i) \sum_{j=1}^{m} \frac{\sum_{k=1}^{n} y_{kj}}{\sum_{r=1}^{n \circledast j} p(\theta_r)} \tag{21}$$

where for $j = 1, \ldots, m$

$$\sum_{r=1}^{n \circledast j} p(\theta_r) = \sum_{r=1,(r,j) \in K''}^{n} p(\theta_r) \tag{22}$$

Note that for all $i, k = 1, \ldots, n$ and $j = 1, \ldots, m$

$$\frac{y_{ij}}{y_{kj}} = \frac{p(\theta_i)}{p(\theta_k)} \tag{23}$$

because of (8). The Lagrange function is, given the above-mentioned constraints,

$$\Lambda(x, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{m \circledast i} x_{ij} \ln x_{ij} + \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{n} y_{ij} - \sum_{i=1}^{n \circledast j} x_{ij} \right) \tag{24}$$

Differentiating Λ with respect to x_{ij} gives us for $(i, j) \in K''$

$$x_{ij} = y_{ij} e^{\lambda_j - 1} \tag{25}$$

with the Lagrangian parameters λ_j normalized such that

$$\mu_j = e^{\lambda_j - 1} = \frac{\sum_{i=1}^{n} y_{ij}}{\sum_{i=1}^{n \circledast j} y_{ij}}. \tag{26}$$

The μ_j provide a simple (and correct) way to arrive at MAXENT solutions for these types of problems, compared to the complicated (and incorrect) way in which Wagner arrives at MAXENT solutions. Comparing the correct MAXENT solution gained with the help of Lagrangian multipliers

$$q(\theta_i) = \sum_{j=1}^{m} \mu_j y_{ij} \tag{27}$$

to Wagner's solution in (21) it is clear that they agree if and only if

$$\frac{y_{ij}}{y_{kj}} = \frac{p(\theta_i)}{p(\theta_k)} \tag{28}$$

which is (23). MAXENT and Wagner conditioning are consistent with each other.

This could not have been any easier, given that the proofs with respect to standard conditioning and Jeffrey conditioning are readily available in the literature. MAXENT, when it is not adulterated by viewing probabilities as properties of the external world rather than representations of uncertainty in the agents who entertain them, seamlessly and elegantly handles the observation of conditionals.

5 References

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.

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.

Dias, Penha Maria Cardozo, and Abner Shimony. "A Critique of Jaynes' Maximum Entropy Principle." *Advances in Applied Mathematics* 2: (1981) 172–211.

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

Grove, A., and J.Y. Halpern. "Probability Update: Conditioning Vs. Cross-Entropy." In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. Providence, Rhode Island: Citeseer, 1997.

Grünwald, P., and J.Y. Halpern. "Updating Probabilities." *Journal of Artificial Intelligence Research* 19: (2003) 243–278.

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

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

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, 1979.

Jaynes, E.T., and G.L. Bretthorst. *Probability Theory: the Logic of Science*. Cambridge, UK: Cambridge University Press, 1998.

Jeffrey, Richard. *The Logic of Decision*. University of Chicago, 1990.

MacKay, David. *Information Theory, Inference and Learning Algorithms*. Cambridge, 2003.

Shimony, Abner. "The Status of the Principle of Maximum Entropy." *Synthese* 63, 1: (1985) 35–53.

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.

Skyrms, Brian. "Maximum Entropy Inference as a Special Case of Conditionalization." *Synthese* 63, 1: (1985) 55–74.

———. "Dynamic Coherence and Probability Kinematics." *Philosophy of Science* 1–20.

———. "Updating, Supposing, and Maxent." *Theory and Decision* 22, 3: (1987b) 225–246.

———. *The Dynamics of Rational Deliberation.* Cambridge, 1990.

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

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

Zubarev, Dmitrii, Vladimir Morozov, and Gerd Röpke. *Statistical Mechanics of Nonequilibrium Processes.* Berlin: Akademie, 1996.