Review: BJPS-2013-414

Summary: the author explores a topic of broad interest to formal epistemologists and philosophers of science, *viz.*, how to revise one's credences when new evidential constraints fail to take the form needed to update via Jeffrey conditionalisation. Unfortunately, the manuscript contains serious flaws, and is not suitable for publication.

The author takes her main task to be "responding to Wagner's Linguist counterexample... as part of a systematic effort to revive interest among philosophers in MaxEnt as an objective updating method" (p. 5) In addition, she purports to provide a "general theorem which seamlessly incorporates Wagner's 'natural generalization of Jeffrey conditionalization' (Wagner, 1992, 250) into MaxEnt orthodoxy" (p. 10). While the manuscript contains interesting bits, it is far from clear that it achieves either goal.

First, a few remarks on setup/framing and tone. Regarding setup/framing: The author fails to provide a clear characterisation of MaxEnt. She says: "The idea is to update a prior probability distribution, given constraint, by finding a posterior probability distribution which fulfils the constraint and is information-theoretically speaking as a close as possible to the prior probability distribution" (p. 2). This is not quite right. MaxEnt is a synchronic norm, which says:

> *MaxEnt.* If an agent's total evidence is given by evidential constraints $C_1, ..., C_n$, and $\mathcal{C}$ is the set of credence functions that satisfy $C_1, ..., C_n$, then she ought to have those credences c that maximise Shannon entropy $H(c) = -\sum_X c(X) \cdot log(c(X))$ on $\mathcal{C}$.

The author, rather than describing MaxEnt, describes Williams' (1980) principle of minimum information (Infomin), a diachronic norm:

> *Infomin.* Suppose that an agent has prior credences b, and has a learning experience in which she receives information in the form of a new constraint C on her credences (and no further information). Let $\mathcal{C}$ be the set of credence functions that satisfy C. Then her posterior should be the credences c in $\mathcal{C}$ that minimise Kullback-Leibler divergence from her prior, $\mathcal{D}_{KL}(c, b) = \sum_X c(X) \cdot log(c(X)/b(X))$.

Of course, the two are intimately related.

$$\begin{aligned}\mathcal{D}_{KL}(c, b) &= \sum_X c(X) \cdot log(c(X)/b(X)) \\ &= \sum_X c(X) \cdot log(c(X)) - \sum_X c(X) \cdot log(b(X)) \\ &= H(c, b) - H(c)\end{aligned}$$

That is, the Kullback-Leibler divergence of b from c is just the difference between the cross entropy of b and c and the entropy of c. And, in certain circumstances of course, learning via Infomin is equivalent to simply taking stock of your evidence after your learning experience and applying MaxEnt. In particular, if your prior credence function $b : \Omega \to \mathbb{R}$ is the uniform distribution, *i.e.*, $b(X) = 1/n$ for all n atoms X of Ω, Infomin and MaxEnt agree. For in that case:

$$\begin{aligned}\mathcal{D}_{KL}(c, b) &= \sum_X c(X) \cdot log(c(X)/b(X)) \\ &= \sum_X c(X) \cdot log(c(X)) - \sum_X c(X) \cdot log(b(X)) \\ &= \sum_X c(X) \cdot log(c(X)) - \sum_X c(X) \cdot log(1/n) \\ &= [\sum_X c(X) \cdot log(c(X))] - log(1/n) \\ &= -log(1/n) - H(c)\end{aligned}$$

And $-log(1/n) - H(c)$ takes a minimum exactly when $H(c)$ takes a maximum. But, in general, agents do not come to inference problems with a total absence of prior information. When that's so (when they have relevant prior information), prior credences b will often not be uniformly distributed. And when *that's* so, MaxEnt and Infomin fail, in general, to agree.

Suppose, for example, that you are going to draw a ball from an urn containing red (R), blue (B) and green (G) balls. Your prior information: you're at least 50% likely to draw a red. Suppose you maximise entropy on your prior evidence $\mathcal{C} = \{p \mid p(R) \geq 1/2\}$ to arrive at your prior credences b. Then $b(R) = .5$ and $b(B) = b(G) = .25$. Now you learn that you're at most 20% likely to draw a green. Compare MaxEnt and Infomin. If you arrive at your posterior credences c by maximising entropy on your total evidence $\mathcal{C}' = \{p \mid p(R) \geq .5 \;\&\; p(G) \leq .2\}$, then $c(R) = .5$, $c(B) = .3$ and $c(G) = .2$. If instead you arrive at your posterior credences c by minimising KL-divergence from your prior b on your new evidence $\mathcal{C}'' = \{p \mid p(G) \leq .2\}$, then $c(R) = .533$, $c(B) = .267$ and $c(G) = .2$.

Providing clean characterisations of both MaxEnt and Infomin, and saying a bit about their relationship, would bring clarity to the manuscript. (The author does, after all, *use* both MaxEnt and Infomin in providing a 'proper MaxEnt solution' to the Linguist problem.) It would also likely have helped avoid the manuscript's most serious mistake. The author misapplies MaxEnt. Indeed, she reproduces Wagner's results only because she misapplies MaxEnt. More on this in a bit.

One additional framing note: The author cites Friedman and Shimony's (1971) results, as well as van Fraassen's Judy Benjamin problem, as two of the three "main arguments against MaxEnt in practice" (p. 4). She claims to be able to defend MaxEnt against these problems, with the result that MaxEnt "survives as a unifying normative principle in probability update" (p. 4). But she offers no indication of how her treatment of the Linguist case might help resolve either problem. The manuscript

would benefit greatly from her doing so.

Regarding tone: the author makes, at times, snarky remarks unaccompanied by in-depth discussion. For example, she says, "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)" (p. 8). There is no reason to rehash Jaynes' diatribe here. It only detracts from the paper.

Issues regarding setup/framing and tone aside, my primary concerns are with the substance of the author's treatment of the Linguist case. She claims that Wagner pairs "non-Bayesian assumptions and MaxEnt" to derive counterintuitive conclusions (p. 10). But "once the non-Bayesian assumptions are replaced by Bayesian assumptions... not only do the conclusions become plausible, they also agree with Wagner's solution" (p. 10). The problem: the author reproduces Wagner's results only by misapplying MaxEnt. Her defence, then, offers no solace to proponents of MaxEnt.

To see this, let's walk slowly through Wagner's treatment of the Linguist case, on behalf of the proponent of Infomin, and the author's complaints about that treatment. Wagner imagines that a proponent of Infomin would treat the Linguist case as follows. "You encounter a native of a 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)" (Wagner 1992, p. 252). Your prior credences b are given by:

	Catholic	*Protestant*
Northerner	$b(\theta_1) = 0.2$	$b(\theta_3) = 0.4$
Southerner	$b(\theta_2) = 0.3$	$b(\theta_4) = 0.1$

Then you learn something. He "utters a phrase in his native tongue" (Wagner 1992, p. 252). This learning experience (for reasons to be explained shortly) imposes the following constraints on your posterior credences c:

- C_1: $c(\theta_1 \vee \theta_2) \geq 0.7$
- C_2: $c(\theta_1 \vee \theta_2 \vee \theta_4) \geq 0.9$
- C_3: $c(\theta_2 \vee \theta_4) \geq 0.2$

Let $\mathcal{C}$ be the set of credence functions that satisfy these constraints. Then Infomin says: your posterior should be the c in $\mathcal{C}$ that minimises $\mathcal{D}_{KL}(c, b) = \sum_X c(X) \cdot log(c(X)/b(X))$. Hence, c should be (as the author correctly reports):

	Catholic	*Protestant*
Northerner	$c(\theta_1) = 0.3$	$c(\theta_3) = 0.1$
Southerner	$c(\theta_2) = 0.45$	$c(\theta_4) = 0.15$

Where do $C_1 - C_3$ come from? Well, imagine expanding the space of possibilities that you have credences over. Imagine in particular coming to recognise the following possibilities: the native's utterance is 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). Suppose further that, upon recognising these possibilities, you would come to have the following credences: $c(\omega_1) = 0.4$, $c(\omega_2) = 0.3$, $c(\omega_3) = 0.2$ and $c(\omega_4) = 0.1$.

You could then reason as follows: ω_1 and ω_2 are inconsistent with θ_3 and θ_4 (Protestants don't utter catholic pieties or anti-Protestant remarks); ω_3 is inconsistent with θ_1 and θ_3 (northerners don't utter southern regionalisms). So your (hypothetically expanded) posterior credences c must satisfy:

	ω_1	ω_2	ω_3	ω_4
θ_1	???	???	$c(\theta_1 \& \omega_3) = 0$	???
θ_2	???	???	???	???
θ_3	$c(\theta_3 \& \omega_1) = 0$	$c(\theta_3 \& \omega_2) = 0$	$c(\theta_3 \& \omega_3) = 0$	???
θ_4	$c(\theta_4 \& \omega_1) = 0$	$c(\theta_4 \& \omega_2) = 0$	???	???
	$c(\omega_1) = 0.4$	$c(\omega_2) = 0.3$	$c(\omega_3) = 0.2$	$c(\omega_4) = 0.1$

This table, together with the probability axioms, places exactly our constraints $C_1 - C_3$ on your credences for $\theta_1 - \theta_4$.

The author claims that Wagner makes three mistakes in deriving the MaxEnt solution to the Linguist problem. First, Wagner misdescribes your prior credences. They are not given by:

	Catholic	*Protestant*
Northerner	$b(\theta_1) = 0.2$	$b(\theta_3) = 0.4$
Southerner	$b(\theta_2) = 0.3$	$b(\theta_4) = 0.1$

but rather by some b defined over this larger space:

	ω_1	ω_2	ω_3	ω_4	
θ_1	$b(\theta_1 \& \omega_1) =?$	$b(\theta_1 \& \omega_2) =?$	$b(\theta_1 \& \omega_3) =?$	$b(\theta_1 \& \omega_4) =?$	$b(\theta_1) = 0.2$
θ_2	$b(\theta_2 \& \omega_1) =?$	$b(\theta_2 \& \omega_2) =?$	$b(\theta_2 \& \omega_3) =?$	$b(\theta_2 \& \omega_4) =?$	$b(\theta_2) = 0.3$
θ_3	$b(\theta_3 \& \omega_1) =?$	$b(\theta_3 \& \omega_2) =?$	$b(\theta_3 \& \omega_3) =?$	$b(\theta_3 \& \omega_4) =?$	$b(\theta_3) = 0.4$
θ_4	$b(\theta_4 \& \omega_1) =?$	$b(\theta_4 \& \omega_2) =?$	$b(\theta_4 \& \omega_3) =?$	$b(\theta_4 \& \omega_4) =?$	$b(\theta_4) = 0.1$
	$b(\omega_1) = 0.4$	$b(\omega_2) = 0.3$	$b(\omega_3) = 0.2$	$b(\omega_4) = 0.1$	

"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" (p.16). Wagner's second mistake: he should not only focus on priors and posteriors defined over this larger space, but also fill in your prior using MaxEnt (p. 14). When you do this, you arrive at the following prior:

	ω_1	ω_2	ω_3	ω_4	
θ_1	$b(\theta_1\&\omega_1) = 0.08$	$b(\theta_1\&\omega_2) = 0.06$	$b(\theta_1\&\omega_3) = 0.04$	$b(\theta_1\&\omega_4) = 0.02$	$b(\theta_1) = 0.2$
θ_2	$b(\theta_2\&\omega_1) = 0.12$	$b(\theta_2\&\omega_2) = 0.09$	$b(\theta_2\&\omega_3) = 0.06$	$b(\theta_2\&\omega_4) = 0.03$	$b(\theta_2) = 0.3$
θ_3	$b(\theta_3\&\omega_1) = 0.16$	$b(\theta_3\&\omega_2) = 0.12$	$b(\theta_3\&\omega_3) = 0.08$	$b(\theta_3\&\omega_4) = 0.04$	$b(\theta_3) = 0.4$
θ_4	$b(\theta_4\&\omega_1) = 0.04$	$b(\theta_4\&\omega_2) = 0.03$	$b(\theta_4\&\omega_3) = 0.02$	$b(\theta_4\&\omega_4) = 0.01$	$b(\theta_4) = 0.1$
	$b(\omega_1) = 0.4$	$b(\omega_2) = 0.3$	$b(\omega_3) = 0.2$	$b(\omega_4) = 0.1$	

Wagner's third mistake, according to the author: when you hear the native's utterance, your learning experience imposes the following constraints on your posterior credences c (rather than the constraints that Wagner describes): $c(\theta_1\&\omega_3) = c(\theta_3\&\omega_1) = c(\theta_3\&\omega_2) = c(\theta_3\&\omega_3) = c(\theta_4\&\omega_1) = c(\theta_4\&\omega_2) = 0$, as well as $c(\omega_1) = 0.4$, $c(\omega_2) = 0.3$, $c(\omega_3) = 0.2$ and $c(\omega_4) = 0.1$. (The author does not make this point explicit. But these are the only constraints that allow her to reproduce Wagner's results.)

Let $\mathcal{C}$ be the set of credence functions that satisfy these constraints. Now we can apply Infomin. Your posterior should be the c in $\mathcal{C}$ that minimises $\mathcal{D}_{KL}(c, b) = \sum_X c(X) \cdot log(c(X)/b(X))$, *viz.*,

	ω_1	ω_2	ω_3	ω_4	
θ_1	$c(\theta_1\&\omega_1) = 0.16$	$c(\theta_1\&\omega_2) = 0.12$	$c(\theta_1\&\omega_3) = 0.00$	$c(\theta_1\&\omega_4) = 0.02$	$c(\theta_1) = 0.30$
θ_2	$c(\theta_2\&\omega_1) = 0.24$	$c(\theta_2\&\omega_2) = 0.18$	$c(\theta_2\&\omega_3) = 0.15$	$c(\theta_2\&\omega_4) = 0.03$	$c(\theta_2) = 0.60$
θ_3	$c(\theta_3\&\omega_1) = 0.00$	$c(\theta_3\&\omega_2) = 0.00$	$c(\theta_3\&\omega_3) = 0.00$	$c(\theta_3\&\omega_4) = 0.04$	$c(\theta_3) = 0.04$
θ_4	$c(\theta_4\&\omega_1) = 0.00$	$c(\theta_4\&\omega_2) = 0.00$	$c(\theta_4\&\omega_3) = 0.05$	$c(\theta_4\&\omega_4) = 0.01$	$c(\theta_4) = 0.06$
	$c(\omega_1) = 0.40$	$c(\omega_2) = 0.30$	$c(\omega_3) = 0.20$	$c(\omega_4) = 0.10$	

This posterior, the author emphasises, agrees with Wagner's about your credences for $\theta_1 - \theta_4$.

The problem: in responding to Wagner, the author misapplies MaxEnt. To reiterate, MaxEnt says:

> *MaxEnt.* If an agent's total evidence is given by evidential constraints $C_1, ..., C_n$, and $\mathcal{C}$ is the set of credence functions that satisfy $C_1, ..., C_n$, then she ought to have those credences c that maximise Shannon entropy $H(c) = -\sum_X c(X) \cdot log(c(X))$ on $\mathcal{C}$.

To apply MaxEnt, the author insists that your prior evidence in the Linguist case is given by the following evidential constraints (though again she does not make this explicit):

- C_1: $b(\omega_1) = 0.4$, $b(\omega_2) = 0.3$, $b(\omega_3) = 0.2$ and $b(\omega_4) = 0.1$.
- C_2: $b(\theta_1) = 0.2$, $b(\theta_2) = 0.3$, $b(\theta_3) = 0.4$ and $b(\theta_4) = 0.1$.

Then she maximises entropy on $\mathcal{C} = \{p \mid p \textit{ satisfies } C_1 \textit{ and } C_2\}$ to arrive at:

	ω_1	ω_2	ω_3	ω_4	
θ_1	$b(\theta_1\&\omega_1) = 0.08$	$b(\theta_1\&\omega_2) = 0.06$	$b(\theta_1\&\omega_3) = 0.04$	$b(\theta_1\&\omega_4) = 0.02$	$b(\theta_1) = 0.2$
θ_2	$b(\theta_2\&\omega_1) = 0.12$	$b(\theta_2\&\omega_2) = 0.09$	$b(\theta_2\&\omega_3) = 0.06$	$b(\theta_2\&\omega_4) = 0.03$	$b(\theta_2) = 0.3$
θ_3	$b(\theta_3\&\omega_1) = 0.16$	$b(\theta_3\&\omega_2) = 0.12$	$b(\theta_3\&\omega_3) = 0.08$	$b(\theta_3\&\omega_4) = 0.04$	$b(\theta_3) = 0.4$
θ_4	$b(\theta_4\&\omega_1) = 0.04$	$b(\theta_4\&\omega_2) = 0.03$	$b(\theta_4\&\omega_3) = 0.02$	$b(\theta_4\&\omega_4) = 0.01$	$b(\theta_4) = 0.1$
	$b(\omega_1) = 0.4$	$b(\omega_2) = 0.3$	$b(\omega_3) = 0.2$	$b(\omega_4) = 0.1$	

But this seriously misrepresents your prior evidence in the Linguist case. Prior to hearing your interlocutor "utter a phrase in his native tongue" you absolutely do *not* have evidence to the effect that $b(\omega_1) = 0.4$, $b(\omega_2) = 0.3$, $b(\omega_3) = 0.2$ and $b(\omega_4) = 0.1$ (Wagner 1992, p. 252). It is the character of that utterance, together with the "aural similarity of the phrases in question" that result in you having these credences for $\omega_1 - \omega_4$ (Wagner 1992, p. 252). Your *posterior* evidence, not your prior evidence, imposes the relevant constraint. You have little if any relevant prior information about what the native will utter (if he utters anything at all). If you have determinate prior credences regarding the content of that utterance at all, they will be (according to MaxEnt) spread over infinitely many hypotheses, *e.g.*, that he will utter a typical Buddhist phrase (ω_5), that he will ask where the nearest bathroom is (ω_6), that he will compliment your fedora (ω_7), etc. You are absolutely *not* certain, prior to hearing your interlocutor's utterance, that one of $\omega_1 - \omega_4$ is true. You have next to no idea what he will say!

What's more, if we agree with the implausible assumption that you have determinate prior credences for $\omega_1 - \omega_4$, then the one thing that you ought to be sure of, prior to hearing your interlocutor's utterance, is that $\theta_3\&\omega_1$, $\theta_3\&\omega_2$, $\theta_4\&\omega_1$ and $\theta_4\&\omega_2$ are all false (Protestants don't utter catholic pieties or anti-Protestant remarks); $\theta_1\&\omega_3$ and $\theta_3\&\omega_3$ are false as well (northerners don't utter southern regionalisms). This is prior information. You do *not* acquire this information as a result of your learning experience. If you did not know that northerners don't utter southern regionalisms prior to hearing the native utter some vaguely comprehensible phrase, you will not know it afterward.

The issue, then, is that the author seriously misdescribes your prior evidence in the Linguist case, as well as the constraints imposed by your learning experience. The result: she misapplies MaxEnt (and Infomin). When we adjust the author's mistaken assumptions (in any number of ways) MaxEnt/Infomin fails to reproduce Wagner's results.

One final remark: the author says, "Wagner's core mistake is the for a Bayesian, the prior joint probability distribution on $\Theta \times \Omega$ is not left unspecified... all Bayesians agree that well-defined events have prior probabilities" (p. 13). This is false. Many sophisticated Bayesians, including Richard Jeffrey, Isaac Levi, Mark Kaplan, Jim Joyce, and many more, agree that agents typically lack determinate prior subjective probabilities. Their opinions are characterised by *imprecise* credal states, rather than precise ones. And this is quite often the *right* response to their unspecific and equivocal evidence.