The Principle of Maximum Entropy and a Problem in Probability Kinematics

Stefan Lukits

August 2, 2011

This summarizes the labours of the first half of the red book, numbers in parentheses refer to the red book. We begin with a quote:

> Probability kinematics resembles ethics in the sense that there are all kinds of things we are able to say about the relations between our intuitions and the prescriptions or rules we propose. We never cease to be vulnerable, however, to the question why the states of affairs we describe should entail that we have one set of probability assignments and updating strategies and not another. That an observation or a piece of evidence should change our assessment of uncertainty with respect to relevant propositions and events in particular ways cannot be a matter of logical consistency. Even a Dutch Book argument rests on assumptions that are entangled with the relations and intuitions we are supposed to explain. Our task in this paper, then, is to show what were to flow from certain assumptions being made and certain intuitions being accepted, and to articulate them clearly and well so that we understand where they are reasonable, arbitrary, or subject to criticism. In all this, we never lose a sense of need for what ethicists in Rawl's tradition call a reflective equilibrium, as it is not the intuitions about particular cases alone, nor the general judgments they sometimes inspire, that carry away the prize, but rather a balance between them. The principle of maximum entropy is a poster child for this method: it is a principle with great generality and scope, arguably outperforming all others, but it also raises worries in particular cases. There is beauty in the fact that, as sweeping as the principle is, it cannot accommodate everything we think and feel about how conditionalization (another term for probability update) should proceed. This paper cautions, however, against undue enthusiasm about the

> full employment theorem, the view that ultimately all rules and methods of conditionalization are tools in the hand of a human inquirer, expressing that which one to use must always be based on the intuitive and creative labour of the user. Probability kinematics is not a sit-down dinner: various approaches mingle, easily shift positions, and have access to the buffet table from different angles. There is no throne even for the view that, when all is said and done, a special place remains for the art of human inquiry.

Let's assume a partition $\{B_i\}_{i=1,\ldots,4n}$ of $A_1 \cup A_2 \cup A_3$ into sets that are of equal measure and whose intersection with A_i is either the empty set or the whole set itself. (MAP) dictates that the number of sets covering A_3 equals the number of sets covering $A_1 \cup A_2$. For convenience, we assume the number of sets covering A_1 to be n. Let $\mathcal{C}$, a subset of the powerset of $\{B_i\}_{i=1,\ldots,4n}$, be the collection of sets which agree with the constraint imposed by (HDQ), i.e.

$$C \in \mathcal{C} \text{ iff } C = \{C_j\} \text{ and } t\mu\left(\bigcup C_j \cap A_1\right) = \mu\left(\bigcup C_j \cap A_2\right)$$

(Provide diagrams.) Let X be the random variable that corresponds to the ratio of the number of sets that are in A_3 and the total number of sets for a randomly chosen $C \in \mathcal{C}$. We would now anticipate that the expectation of X (which we will call EX) gives us an indication of the posterior probability that Judy is in A_3. Grove and Halpern's uniformity approach (based on retrospective conditioning, see Diaconix and Zabell, 822) results in this quantity, which we have loosely called q_3, being $q_3 = .5$, whereas the maximum entropy approach (either using Kullback-Leibler directly or Lagrange Multipliers) results in $q_3 = .53$ for $t = 3$. In which direction is the powerset approach outlined in the last couple of paragraphs going to take us? The powerset approach is often superior to the uniformity approach: if you have played Monopoly, you will know that the frequencies for rolling a 2, a 7, or a 10 with two dice tend to conform more closely to the binomial distribution (based on a powerset approach) rather than to the uniform distribution with $P(X = i) = 1/11$ for $i = 2, \ldots, 12$.

The binomial distribution dictates the value of EX, using simple combinatorics. In this case we require, again for convenience, that n be divisible by t and the 'grain' of the partition A be $s = n/t$. We introduce a few variables which later on will help for abbreviation:

$$n = ts \qquad 2m = n \qquad 2j = n - 1 \qquad \psi = t^2 + 1$$

EX, of course, depends both on the grain of A and the value of t. It makes sense to make it independent of the grain by letting the grain become increasingly finer and by determining EX as $s \to \infty$. This cannot be done for the binomial distribution, as it is notoriously uncomputable for large numbers (even with a powerful computer things get dicy around $s = 20$). But, equally notorious, the normal distribution provides a good approximation of the binomial distribution and will help us arrive at a value q_3 suggested by the powerset approach.

First, we express the random variable X by the two independent random variables X_{12} and X_3. X_{12} is the number of sets in the randomly chosen C which cover either A_1 or A_2 (the random variable of the number of sets covering A_1 and the random variable of the number of sets covering A_2 are decisively not independent, because they need to obey (HDQ)); X_3 is the number of sets in the randomly chosen C which cover A_3. A relatively simple calculation shows that $EX_3 = n$, which is just what we would expect (either the powerset approach or the uniformity approach would give us this result).

Using combinatorics,

$$EX_3 = 2^{-2n} \sum_{i=0}^{2n} i \binom{2n}{i} = n \text{ (use } \binom{n}{k} = \frac{n}{k} \binom{n-1}{k-1} \text{)}$$

The expectation of X, X being the random variable expressing the ratio of the number of sets covering A_3 and the number of sets covering $A_1 \cup A_2 \cup A_3$, is

$$EX = \frac{EX_3}{EX_{12} + EX_3} = \frac{n}{EX_{12} + n}$$

Using the uniformity approach, $EX_{12} = n$ and $EX = 1/2$, just as Grove and Halpern suggest (although their uniformity approach is admittedly less crude than the one used here). Will the powerset approach concur with the uniformity approach, will it support the principle of maximum entropy, or will it make another suggestion on how to update the prior probabilities? To answer this question, we must find out what EX_{12} is, for a given value t and $s \to \infty$, using the binomial distribution and its approximation by the normal distribution.

Using combinatorics,

$$EX_{12} = (t+1) \frac{\sum_{i=1}^{s} i \binom{ts}{i} \binom{ts}{ti}}{\sum_{i=0}^{s} \binom{ts}{i} \binom{ts}{ti}}$$

Let us call the numerator of this fraction NUM and the denominator DEN. According to the de Moivre-Laplace Theorem,

$$\mathrm{DEN} = \sum_{i=0}^{s} \binom{ts}{i}\binom{ts}{ti} \approx 2^{2n} \sum_{i=0}^{s} \int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \mathcal{N}(\frac{n}{2},\frac{n}{4})(i)\mathcal{N}(\frac{n}{2},\frac{n}{4})(ti)di$$

where

$$\mathcal{N}(\mu,\sigma^2)(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Substitution yields

$$\mathrm{DEN} \approx 2^{2n}\frac{1}{\pi m} \sum_{i=0}^{s} \int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{m} - \frac{t^2\left(x-\frac{m}{t}\right)^2}{m}\right) dx$$

Consider briefly the argument of the exponential function:

$$-\frac{(x-m)^2}{m} - \frac{t^2\left(x-\frac{m}{t}\right)^2}{m} = -\frac{t^2}{m}(a''x^2+b''x+c'') = -\frac{t^2}{m}\left(a''(x-h'')^2+k''\right)$$

with (the double prime sign corresponds to the simple prime sign for the numerator later on)

$$a'' = \frac{1}{t^2}\psi \qquad b'' = (-2m)\frac{1}{t^2}(t+1) \qquad c'' = 2m^2\frac{1}{t^2}$$

$$h'' = -b''/2a'' \qquad k'' = a''h''^2 + b''h'' + c''$$

Consequently,

$$\mathrm{DEN} \approx 2^{2n}\exp\left(-\frac{t^2}{m}k''\right)\sqrt{\frac{1}{\pi a'' m t^2}}\int_{-\infty}^{s+\frac{1}{2}} \mathcal{N}\left(h'', \frac{m}{2a''t^2}\right) dx$$

And, using the error function for the cumulative density function of the normal distribution,

$$\mathrm{DEN} \approx 2^{2n-1}\sqrt{\frac{1}{\pi a'' m t^2}}\exp\left(-\frac{k''t^2}{m}\right)\left(1-\mathrm{erf}(w'')\right) \tag{1}$$

with

$$w'' = \frac{t\sqrt{a''}\left(s+\frac{1}{2}-h''\right)}{\sqrt{m}}$$

We proceed likewise with the numerator, although the additional factor introduces a small complication:

$$\begin{aligned} \text{NUM} &= \sum_{i=1}^{s} i\binom{ts}{i}\binom{ts}{ti} = \sum_{i=1}^{s} s\binom{ts}{i}\binom{ts-1}{ti-1} \\ &\approx s2^{2n-1}\sum_{i=1}^{s}\mathcal{N}\left(m,\frac{m}{2}\right)(i)\mathcal{N}\left(j,\frac{j}{2}\right)(ti-1) \end{aligned}$$

Again, we substitute and get

$$\text{NUM} \approx s2^{2n-1}\left(\pi\sqrt{mj}\right)^{-1}\sum_{0}^{s-1}\int_{i-\frac{1}{2}}^{i+\frac{1}{2}}\exp\left(a'(x-h')^2+k'\right)$$

where the argument for the exponential function is

$$-\frac{1}{mj}\left(j(x-m)^2+mt^2\left(x-\frac{j+1}{t}\right)^2\right)$$

and therefore

$$a' = j+mt^2 \qquad b' = 2j(1-m)+2mt\,(t-j) \qquad c' = j(1-m)^2+m\,(t-j-1)^2$$

$$h' = -b'/2a' \qquad k' = a'h'^2+b'h'+c'$$

Using the error function,

$$\text{NUM} \approx 2^{2n-2}\frac{s}{\sqrt{\pi a'}}\exp\left(-\frac{k'}{mj}\right)\left(1+\text{erf}(w')\right) \tag{2}$$

with

$$w' = \frac{\sqrt{a'}\left(s-\frac{1}{2}-h'\right)}{\sqrt{mj}}$$

Combining (1) and (2),

$$\begin{aligned} EX_1 &= (t+1)\frac{\text{NUM}}{\text{DEN}} \\ &\approx \frac{1}{2}(t+1)\sqrt{\frac{\psi ts}{\psi ts-1}}se^{\alpha_{t,s}} \end{aligned}$$

for large s, because the arguments for the error function w' and w'' escape to positive infinity in both cases (NUM and DEN) so that their ratio goes to 1. The argument for the exponential function is

$$\alpha_{t,s} = -\frac{k'}{mj} + \frac{k''t^2}{m}$$

and, for $s \to \infty$, goes to

$$\alpha_t = \frac{1}{2}\psi^{-2}(2t^3 - 3t^2 + 4t - 5)$$

Notice that, for $t \to \infty$, α_t goes to 0 and

$$EX = \frac{n}{EX_{12} + n} \to \frac{2}{3}$$

in accordance with intuition T2.

We now have a formula for the powerset approach corresponding to the formula for the MAXENT approach, giving us q_3 dependent on t. Notice that this formula is for $t = 2, 3, 4, \ldots$. For $t = 1$ use the Chu-Vandermonde identity to find that

$$EX_{12} = (t+1)\frac{\sum_{i=1}^{s} i\binom{ts}{i}\binom{ts}{ti}}{\sum_{i=0}^{s}\binom{ts}{i}\binom{ts}{ti}} = (t+1)\frac{s}{2}$$

and consequently $EX = 1/2$, as one would expect. For $t = 1/2, 1/3, 1/4, \ldots$ we can simply reverse the roles of A_1 and A_2. These results give us a graph of the normalized odds vector, a bit bumpy around the middle because the t-values are discrete and farther apart in the middle, as $t = q/(1-q)$. (Put a line through $q = 3/4$.) Comparing the graphs of the normalized odds vector under Grove and Halpern's uniformity approach, Jaynes' MAXENT approach, and the powerset approach suggested in this paper, it is clear that the powerset approach supports MAXENT.

The powerset approach cannot stand on its own. The reason behind using MAXENT is that we want our evidence to have just the right influence on our prior probabilities, i.e. neither over-inform nor under-inform. There is no corresponding reason why we should update our probabilities using the powerset approach. What the powerset approach does is lend support to another approach. In this task, it is strangely powerful because it tells us what would happen if we were to divide the event space into infinitesimally

small 'atomic' bits of information, analogous to the way in which the normal distribution captures so many natural phenomena because it generalizes the binomial distribution with its discrete on/off switches (as in a Galton box) to an infinite grain and a smooth distribution.

Going through the calculations, it seems at many places that the powerset approach should give its support to the uniformity approach. It was highly surprising to me that in the mathematical analysis, $\alpha_{t,s}$ turned out to converge to a non-trivial (neither 0 nor $+/-\infty$) factor, enabling a graph of the normalized odds vector that was not of the simple nature of the graph suggested by Grove and Halpern. Most surprisingly, the powerset approach, prima facie unrelated to an approach using information, supports the idea that a set of events about which nothing is known (such as A_3) gains in probability in the posterior probability distribution compared to the set of events about which something is known (such as A_1 and A_2).