Teddy Seidenfeld [1]

ENTROPY AND UNCERTAINTY

ABSTRACT

This essay is, primarily, a discussion of four results about the principle of maximizing entropy (MAXENT) and its connections with Bayesian theory. Result 1 provides a restricted equivalence between the two where the Bayesian model for MAXENT inference uses an *a priori* probability that is

(Frieden, 1972) and estimating missing proportions in contingency tables for socio-economic survey data (Denzau *et al.*, 1984). But I doubt there is a more staunch defender of the generality of entropy as a basis for quantifying [illegible] (probabilistic) uncertainty than the physicist E. T. Jaynes.

Almost thirty years ago, Jaynes (1957) offered his celebrated papers [illegible] on "Information Theory and Statistical Mechanics." There he argued that [illegible]

$1/6$ $(i = 1, \ldots, 6)$.[2] If, instead the constraint specifies

$$E[\text{ number of spots on next roll }] = 4.5 \tag{3}$$

instead of the value 3.5 (for a fair die), the MAXENT solution (Jaynes, 1978) is (to five places):

$$\{p_1, \ldots, p_6\} = \{.05435, .07877, .11416, .16545, .23977, .34749\}. \tag{4}$$

Note that in (4) the probabilities are shifted away from the uniform distribution to lie on a smooth (convex) curve, increasing (decreasing) in

Not all who have examined these supporting arguments find them convincing. (See especially Dias and Shimony, 1981; Frieden, 1984; Friedman and Shimony, 1971; Rowlinson, 1970; Shimony, 1973. Jaynes offers selected rebuttal in (1978).) In what follows I present concerns I have primarily [illegible]

(S_1) U_S is a continuous function of the p_i's.

(S_2) When $P = \{1/n, \ldots, 1/n\}$ is the uniform distribution on n-states, U_S is monotonically increasing in n, the number of states over which one is uncertain.

(S_3) U_S is additive over decomposition of the sample space of possible outcomes. That is, let $\Omega = \{s_1, \ldots, s_n\}$ be the set of (n) possible outcomes, and let Ω be partitioned into $m \leq n$ disjoint subsets $\Omega' = \{r_1, \ldots, r_m\}$ with r_i a subset of Ω. If P is a

Now, it is clear that $P_1(\cdot) = P_1(\cdot \mid e)$, since P_1 satisfies c_{k+1}. Define a prob-

[redacted]

dicting the assumption that P_0 is the MAXENT solution for constraints C_0. To verify that P_0' satisfies C_0, note that the class of distributions satisfying outcomes $\{1, 3, 5\}$—which is the conditional probability $P_0'(\cdot \mid e_1)$—but instead is the distribution (see Appendix).

$$P_1'(i) = \{.21624, .31752, .46624\} \qquad (i = 1, 3, 5) \tag{5a}$$

when P^0 is discrete, and by the ...

In the case of discrete distributions (3) ...

variables, some (Bayesian) conditionalizations do not agree with the revision from P^0 to P^1 by [illegible]

Besides generalizing U_S with discrete distributions, I_K affords a consistent extension of entropy to continuous distributions, unlike the (natural) [illegible] of a cubical die (with spots from 1 to 6 arranged in conventional order), then

Table 1. *('Yes'/'No' identifies which arrangements are possible.)*

# spots showing			
2	Yes	No	Yes
3	Yes	Yes	Yes
4	Yes	No	Yes
5	Yes	Yes	Yes
6	Yes	Yes	Yes

priority in the application of Insufficient Reason. Of course, what is lackin where p_i ($i = 1, \ldots, 6$) is the probability of i-spots showing up.

However, since the alternative partition (Table 1) is a refinement of the six-fold partition used above, the constraint (10) applies there too. Specifically, define $f(\text{ state }_j)$ ($j = 1, \ldots, 14$—counting across possible states in Table 1) as follows:

$$f(\text{state}_1) = 1, f(\text{state}_2) = f(\text{state}_3) = 2,$$
$$f(\text{state}_4) = f(\text{state}_5) = f(\text{state}_6) = 3,$$
$$f(\text{state}_7) = f(\text{state}_8) = 4,$$
$$f(\text{state}_9) = f(\text{state}_{10}) = f(\text{state}_{11}) = 5,$$
and
$$f(\text{state}_{12}) = f(\text{state}_{13}) = f(\text{state}_{14}) = 6$$

Then (10) is equivalent to the constraint:

$$E[f] = 55/14. \tag{12}$$

But the distribution over the 14 states which maximizes entropy, subject to a constant and constants are (vacuously) probabilistically independent of other variables. In Section 4, where MAXENT is contrasted with Bayesian inference, the device of using a degenerate 0–1 distribution with nuisance factors is key to understanding an important

Summary: The question addressed in this section is

conditional probability $P_{BK}(\cdot \mid \cdot)$—coherence.

(B_2) $P_{BK}(\cdot \mid \cdot)$ is relativized to background evidence BK (consistent and closed under entailment), where BK depicts the agent's *total background evidence*.

(B_3) As regulated by Bayes' theorem for conditional probability $P_{BK}(\cdot \mid \cdot \& e)$ is the agent's hypothetical belief state for the hypothesis that he accepts only the new (consistent) evidence e, i.e., under the hypoth-

[redacted]

given BK)—conditionalization.[10]

[redacted]

a restricted equivalence between revising probabilities through MAXENT

[redacted]

– the MAXENT principle is coherent (from a Bayesian point of view) returns us to the question of the previous section. Under which conditions can we extend (refine) the field of events, while preserving MAXENT solutions for a given set of constraints?

It is from this perspective I propose we consider the [illegible] tion Jaynes wants for the "constraint" in his Brandeis Dice problem.[16] So, [illegible]

In (1978) he writes,

[illegible]

exchangeable P is a mixture if i.i.d. multinomial distributions (each on a sample space of n-outcomes) for some "mixing" prior π on the multinomial parameter. Recall, when $r = (n+1)/2$, that is when the "sample average" equals the average of the number of spots showing on the n faces of the

5. COMMENTS ON THE CONCENTRATION THEOREM (JAYNES, 1979 AND SEE (1963, PP. 51–52))

Theorem (Jaynes): Consider N repetitions of an experiment with n possi-

ble outcomes on a given trial. Let f_i $(1 \leq i \leq n)$ be the observed relative frequency of the ith outcome in these N trials. Then the class of sequences the MAXENT probability is determined by the (asymptotic) *proportion* of these states with frequencies close to the MAXENT distribution. Why is this a problem? It is because, if the concentration about the MAXENT solution

Support for this research came from the Department of Preventive Medicine, Washington University (St. Louis), and NSF grant SES 8607300.

paradoxes" (due to Dawid *et al.*, 1973). As Dawid *et al.* use their anoma-
lies to question this [illegible]

[redacted]

(1971) formulation, it is supposed there is one state whose magnitude a_m equals the average of the n magnitudes $a_m = (1/n) \cdot \sum_{i=1}^{n} a_i$. This condi-

[redacted]

(Recall. $\lambda = \infty$ corresponds to the point

I thank Prof. E. Greenberg for alerting me to Frieden's recent work.

APPENDIX A: ON THE MAXENT FORMALISM

Here we review some of the

[redacted]

Then, by (A3) and (A4)

$$-\frac{\partial}{\partial \lambda_1} \log Z = \left(1 - 7x^6 + 6x^7\right) / \left[(1-x)(1-x^6)\right] = c_1. \qquad \text{(A6)}$$

In the problem discussed on p. 269, (10) sets the constraint: $c_1 = 55/14$. Solving (A6) for this value yields:

[redacted]

(as obtained on my TI 58C). This results in the MAXENT distribution

[redacted]

In other words, Result 4 establishes that the MAXENT and use the argument which follows to establish the desired property for each P^1_{iX}. By continuity of cross-entropy shifts, the desired property obtains for their limit, P^1_X.) Refine X to Y so that P^0_Y is uniform on Y. (This is possible [illegible]

[illegible]

c in the measure space $(Y, \mathcal{Y})$. By the lemma (above), the minimum cross-entropy shift from P^0_Y to P^1_Y agrees with the minimum cross-entropy shift from P^0_X to P^1_X on X. But, with P^0_Y uniform on Y, the minimum cross-entropy shift is just the MAXENT distribution P, in the measure space $(Y, \mathcal{Y})$, subject to the constraint c. Then apply Result 4 to show that P^1_Y [illegible]

[illegible]

The inequality (B6) is demonstrated as follows. Let

$$k = rm, \tag{B7}$$

so $P^U(E_1)/P^U(E_2) = r$. Substituting (B7) into (B5), we obtain

$$P(E_1 \cup E_2) \cdot Z = mr^{1-\alpha}(1/[\alpha^\alpha + (1-\alpha)^{1-\alpha}]). \tag{B8}$$

The inequality (B6) obtains just in case

$$1/[\alpha^\alpha + (1-\alpha)^{1-\alpha}] < (1 \ldots$$

Taking the derivative (with respect to r) of the r.h.s. of (B9) and setting it equal to 0 ...

$$r = (1-\alpha)/\alpha \{= P(E_1)/P(E_2)\} \tag{B10}$$

[redacted]

University Press.

[redacted]

4th printing. New York: Interscience Publishers.

Dawid, A. P., M. Stone, and J. V. Zidek (1973), "Marginalization paradoxes in Bayesian and structural inference." *Journal of the Royal Statistical Society*, Series B **35**, 189–233 (with discussion).

[redacted]

Jaynes, E. T. (1983), *Papers on Probability, Statistics and Statistical Physics*, ed. R. Rosenkrantz. Dordrecht: D. Reidel.

Jaynes, E. T. (1983), "Highly informative priors." In *Bayesian Statistics 2*, ed. J.M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, pp. 329–