Entropy and Uncertainty
Author(s): Teddy Seidenfeld
Source: *Philosophy of Science*, Vol. 53, No. 4 (Dec., 1986), pp. 467-491
Published by: The University of Chicago Press on behalf of the Philosophy of Science Association
Stable URL: http://www.jstor.org/stable/187902
Accessed: 29/09/2014 16:46

Philosophy of Science

December, 1986

ENTROPY AND UNCERTAINTY*

TEDDY SEIDENFELD†

Department of Philosophy
Carnegie-Mellon University

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 uniform, and where all MAXENT constraints are limited to 0-1 expectations for simple indicator-variables. The other three results report on an inability to extend the equivalence beyond these specialized constraints. Result_2 established a sensitivity of MAXENT inference to the choice of the algebra of possibilities even though all empirical constraints imposed on the MAXENT solution are satisfied in each measure space considered. The resulting MAXENT distribution is not invariant over the choice of measure space. Thus, old and familiar problems with the Laplacian principle of Insufficient Reason also plague MAXENT theory. Result_3 builds upon the findings of Friedman and Shimony (1971; 1973) and demonstrates the absence of an exchangeable, Bayesian model for predictive MAXENT distributions when the MAXENT constraints are interpreted according to Jaynes's (1978) prescription for his (1963) Brandeis Dice problem. Lastly, Result_4 generalizes the Friedman and Shimony objection to cross-entropy (Kullback-information) shifts subject to a constraint of a new odds-ratio for two disjoint events.

1. Introduction. Thirty-six years after Shannon (1948) and Wiener (1948) introduced their now familiar expression for the uncertainty captured in

*Received January 1985; revised January 1986.

†I thank J. Kadane, I. Levi, and A. Shimony for their detailed, constructive comments on an earlier draft of this paper. Also, I have benefited from discussions with: A. Denzau, C. Genest, P. Gibbons, E. Greenberg, E. Jaynes, M. Schervish, G. Tsebelis, B. Wise, the members of the Philosophy Department Colloquium at Carnegie-Mellon University, and other helpful critics at the 29th NBER-NSF Seminar on Bayesian Inference in Economics.

Support for this research came from the Department of Preventive Medicine, Washington University (St. Louis), and N.S.F. Grant #SES-8607300.

Philosophy of Science, 53 (1986) pp. 467–491.

a probability distribution, entropy formalism is a thriving enterprise. Its advocates find applications in diverse settings, including problems of image restoration (Frieden 1972), and estimating missing proportions in contingency tables for socio-economic survey data (Denzau, Gibbons, and Greenberg 1984). But I doubt there is a more staunch defender of the generality of entropy as a basis for quantifying (probabilistic) uncertainty than the physicist E. T. Jaynes.

Almost thirty years ago, Jaynes (1957) offered his celebrated papers on "Information Theory and Statistical Mechanics." There he argued that statistical mechanics is best understood as an instance of "inference," subject to inductive principles for maximizing uncertainty (measured by entropy), rather than as a "physical theory" in which, for example, the results of ergodic theory depend upon equations of motion and suspect assumptions about appropriateness of time-intervals (for use in identifying time frequencies and phase averages). In one fell swoop, Jaynes's approach reproduced a host of computational rules for determining statistical distributions, grounded on a simple rule for maximing entropy. The conceptual innovation was to give this rule a wide scope, elevating it to a principle of inductive logic for assigning (subjective) probabilities in an observer-invariant (objective) fashion. Investigators holding the same "evidence" agree in their determination of probabilities, provided they adhere to Jaynes's program for selecting a probability distribution that maximizes entropy subject to the constraints of the shared "evidence."

Consider a simple illustration, used by Jaynes (1963) in his Brandeis Lectures. Suppose we are faced with an ordinary six-sided die whose "bias" is stipulated to constrain our expectation for the next roll:

$$E[\text{number of spots on next roll}] = 3.5. \tag{1}$$

The problem is to determine a (subjective) probability distribution for the set $X = \{1, \ldots, 6\}$ of possible outcomes. Shannon's formula for the uncertainty (entropy) in a discrete distribution (over n-states) is:

$$U_S = -\sum_{i=1}^{n} p_i \cdot \log(p_i). \tag{2}$$

Jaynes's principle of Maximizing Entropy (MAXENT) directs us to choose that distribution over X ($p_i \geq 0$, $\Sigma p_i = 1$) which maximizes (2) subject to the constraint (1). That is, from among those distributions satisfying:

$$\sum_{i=1}^{6} i \cdot p(i) = 3.5,$$

maximize uncertainty. The solution is the uniform distribution, $p(i) =$

$^1/_6$ (i =1,. . .,6).[1] 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):

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

Note that in (4) the probabilities are shifted away from the uniform distribution to lie on a smooth (convex) curve, increasing (decreasing) in p_i whenever the constraint fixes an expectation greater than (less than) 3.5—corresponding to the uniform distribution.

Why does Jaynes find the MAXENT principle compelling? Why should a rational person pick the uniform distribution from among the continuum of distributions satisfying (1), or choose the distribution (4) from among the continuum of distributions satisfying (3)? I can identify five reasons proposed by various authors:

(i) A pragmatic justification—in an impressive variety of empirical problems, researchers find MAXENT solutions useful. (See Frieden 1984.)

(ii) An argument for the long run—asymptotically, a MAXENT distribution is the focus of concentration among all distributions satisfying the given constraints. That is, if we use entropy to gauge "distance" between distributions, asymptotically, the class of distributions satisfying the given constraints concentrates sharply about the MAXENT solution. (See Jaynes 1979.)

(iii) An a priori analysis—MAXENT is justified by axiomatic considerations of (necessary) conditions for representing uncertainty. (See Shore and Johnson 1980 and 1981.)

(iv) A defense of MAXENT through Insufficient Reason—MAXENT provides a consistent form of the Laplacian principle of Insufficient Reason; hence, it helps rehabilitate the classical interpretation of probability. (See Jaynes 1978.)

(v) MAXENT justified as an extension of Bayesian theory—the Bayesian program for representing degrees of belief by probabilities and "updating" these through conditional probability (as regulated by Bayes's theorem) is a special case of MAXENT inference. (See Jaynes 1968, 1978, and 1981; Rosenkrantz 1977; and Williams 1980.)

Not all who have examined these supporting arguments find them convincing. (See especially: Dias and Shimony 1981; Frieden 1984; Fried-

[1]The MAXENT formalism is discussed in the appendix.

man and Shimony 1971; Rowlinson 1970; and Shimony 1973. Jaynes 1978 offers selected rebuttal.) In what follows, I present concerns I have primarily with the third, fourth, and fifth claims (above). I fear MAXENT is not as attractive as the advertising suggests. In particular, my doubts center on the assertion that MAXENT avoids the conceptual difficulties that plague simpler versions of Insufficient Reason. (This is discussed in section 3. See also my 1979.) A related argument (given in section 4) undercuts the allegation that canonical applications of MAXENT have Bayesian models; in fact, it shows that all but the most trivial applications of MAXENT are unBayesian. Hence, there is solid ground for disputing the fifth claim (above). All of this is previewed in the discussion (section 2.1) of the relation between Bayesian "conditionalization" and shifts that minimize changes in entropy—connected with an evaluation of claim (iii).

The scope of a single essay is insufficient also to address the first two arguments (justification (i) and (ii)) in the detail they deserve. A pragmatic appeal to successful applications of MAXENT formalism cannot be dismissed lightly. The objections to MAXENT that I raise in this paper are general. Whether (and if so, how) the researchers who apply MAXENT avoid these difficulties remains on open question. Perhaps, by appeal to extra, discipline-specific assumptions they find ways to resolve the conflicts within MAXENT theory. A case-by-case examination is called for.

Justification (ii) introduces a family of issues separate from those relevant to concerns (iii)–(v): when do asymptotic properties of an inductive principle warrant its use in the short run too? I offer some reflections on the "concentration" theorem in section 5.

The reader will observe that throughout this essay I rely on Jaynes's prescriptions for the application and interpretation of the MAXENT formalism. Of course, my intent is to ask serious questions, not to hunt out minor inconsistencies in a scholar's writings spanning thirty years' active work. That is, I take Jaynes's papers on MAXENT to be the most thorough account available.

2. Axiomatic Properties Characterizing MAXENT and Its Generalization through Kullback-Leibler Cross-Entropy.

2.1. Shannon (1948) proved an elegant uniqueness theorem establishing that U_S (2) is characterized by three simple properties:

(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 probability distribution over Ω, P' the corresponding distribution over Ω', and $P(\cdot|r_i)$ the conditional distribution (over Ω) given r_i, then:

$$U_S(P) = U_S(P') + \sum_{i=1}^{m} p_i' \cdot U_S(P(\cdot|r_i)). \tag{5}$$

A few remarks remind the reader why these three conditions are important for the MAXENT program. The property (S_1) is a structural assumption that guarantees MAXENT distributions shift smoothly with smooth changes in constraints. (S_2) is important since the uniform distribution $p_i = 1/n$ ($i = 1, \ldots, n$) maximizes entropy over all distributions on n-states. Hence, (S_2) assures that, subject to MAXENT, uncertainty increases with the number of possibilities about which one is "ignorant." Lastly, (S_3) is reminiscent of the multiplication rule for probabilities:

$$P(A\&B) = P(A|B) \cdot P(B).$$

Condition (S_3) suggests a version of the Bayesian principle of conditionalization is satisfied by MAXENT (as I noted in 1979, p. 438, n. 22). Specifically, we have:

Result$_1$. Let P_0 be a MAXENT solution subject to the constraints $C_0 = \{c_1, \ldots, c_k\}$. If one adds the constraint that event e occurs (assumed consistent with C_0), then the new (updated) MAXENT distribution P_1 is the "old" conditional probability $P_0(\cdot|e)$ if and only if $P_0(\cdot|e)$ satisfies the constraints in C_0.

Proof ("if"). Use (S_3) by setting $\Omega' = \{e, \sim e\}$. Let $C_1 = \{c_1, \ldots, c_k, c_{k+1}\}$, where c_{k+1} is the constraint $E[I_e] = 1$, for the indicator variable

$$I_e = 1 \text{ if } e \text{ occurs}$$
$$= 0 \text{ otherwise.}$$

Contrary to the conclusion, suppose P_1 (the MAXENT solution subject to C_1) is not equal to $P_0(\cdot|e)$. That is, suppose

$$U_S(P_1) > U_S(P_0(\cdot|e)). \tag{*}$$

Now, it is clear that $P_1(\cdot)$ is also a conditional probability of the form $P_1(\cdot) = P_1(\cdot|e)$, since P_1 satisfies c_{k+1}. Define a probability $P_0'(\cdot)$ by $P_0'(\cdot) = P_0(e) \cdot P_1(\cdot|e) + P_0(\sim e) \cdot P_0(\cdot|\sim e)$. Then, by ($S_3$), $U_S(P_0') > U_S(P_0)$, in light of the inequality (*). But P_0' satisfies C_0, contradicting 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 a constraint set is convex (see appendix A), and note that P_1 does (since it satisfies C_1) and that either $P_0(e) = 1$ whence $P_0' = P_1$, or else $P_0(\cdot \mid \sim e)$ satisfies C_0 since P_0 and $P_0(\cdot \mid e)$ do (and constraints are taken to be linear in probability—see appendix A).

Proof ("only if"). This is trivial. Whenever $P_1 = P_0(\cdot \mid e)$, $P_0(\cdot \mid e)$ satisfies C_1 and hence satisfies C_0 also. □

Result$_1$ provides, also, for the following:

COROLLARY. *Where C_0 is vacuous and $\{C_i\}(i = 1, \ldots)$ is an increasing sequence of constraint sets, $C_i \subseteq C_{i+1}$, corresponding to a sequence $\{e_i\}$ of mutually consistent observations (measurable) in the initial sample space, then $P_i(\cdot) = P_0(\cdot \mid e_1, \ldots, e_i)$ is the MAXENT probability for constraints C_i.*

Proof. C_i is summarized by the sole constraint: $I_{e_1 \cap \ldots \cap e_i} = 1$. Hence, $C_i = C_{i-1} \cup \{I_{e_i} = 1\}$. Then apply mathematical induction with Result$_1$.[2]

Whenever the constraints arise by observations of events (measurable) in the space X of P_0, the corollary establishes an equivalence of the MAXENT principle and Bayesian conditionalization with a uniform "a priori" probability over X. But before this equivalence is accepted as justification for the fourth or fifth claims (p. 469), two questions must be addressed.

(A) What is the relation between MAXENT and Bayesian solutions that use other than a uniform "a priori" probability over X?

(B) What is the relation between MAXENT and Bayesian solutions when other than indicator-variables appear among the constraints?

I discuss the first of these in section 2.2, following. The significance of the second question is made evident by an example.

Recall that the unconstrained MAXENT solution for the six-sided die, $X = \{1, \ldots, 6\}$, is the uniform probability $p_i = 1/6$ $(i = 1, \ldots, 6)$. As this distribution satisfies the constraint $E[X] = 3.5$, we may take

$$C_0' = \{E[X] = 3.5\}$$

while preserving the uniform probability, $p_i = 1/6$, as the MAXENT solution $P_0'(\cdot) = P_0(\cdot)$. However, if we add the observation, e_1, that an odd-numbered side resulted on the roll, then the MAXENT solution for $C_1' = \{E[X] = 3.5, I_{e_1} = 1\}$ is *not* the uniform distribution over the three out-

[2]Where the support for P_0 is a denumerable set, this argument depends upon σ-additivity to extend it to $C_\infty = \bigcup_{i<\omega} C_i$.

comes {1,3,5}—which is the conditional probability $P_0'(\cdot\,|e_1)$—but instead is the distribution (see appendix)

$$P_1'(i) = \{.21624,\ .31752,\ .46624\} \quad (i = 1,3,5). \tag{5.1}$$

Likewise, had the observation been that the roll yielded an even-numbered side, $I_{e_1} = 0$, the MAXENT solution for the constraint set $C_1'' = \{E[X] = 3.5,\ I_{e_1} = 0\}$ would be

$$P_1''(i) = \{.46624,\ .31752,\ .21624\}\ (i = 2,4,6) \tag{5.2}$$

instead of the conditional probability $P_0'(\cdot\,|\bar{e}_1)$, uniform over {2,4,6}. Bayesian conditionalization requires that $P_{C_1'}(\cdot) = P_{C_0'}(\cdot|e_1)$ and that $P_{C_1''}(\cdot) = P_{C_0'}(\cdot\,|\bar{e}_1)$, both in conflict with (5.1) and (5.2). Expressed in still other words, the MAXENT solutions $P_1'(\cdot)$ and $P_1''(\cdot)$ are not the conditional probabilities $P_0'(\cdot\,|e_1)$ and $P_0'(\cdot\,|\bar{e}_1)$, though the former correspond to an addition of new evidence e_1 or $\bar{e}_1$ to the constraints imposed on P_0'.

Of course, where $P_0(\cdot\,|e)$ fails to satisfy the old constraints, C_0, P_1 must differ from this conditional probability. Unfortunately, whenever the initial constraints C_0 include more than mere 0–1 expectations for indicators (measurable) in the space of P_0, there are events in the algebra of P_0 for which $P_0(\cdot\,|e)$ fails C_0. Hence, without the proviso that $P_0(\cdot\,|e)$ satisfies C_0, Bayesian conditionalization conflicts with shifts according to the MAXENT rule unless *all* constraints (in C_1) are mere 0–1 expectations for indicator variables.

Perhaps there is a way out of this difficulty by extending the algebra so that *all* constraints reduce to 0–1 expectations for indicator variables (measurable) in the extended algebra? This is discussed in sections 3 and 4.

2.2 Aside on Kullback-Information and Its Relation to (Shannon) Uncertainty: There is an important generalization of U_S (2), due to Kullback (1951), essential for a coherent account of "uncertainty" with continuous random variables and useful in widening the scope of the MAXENT principle even for discrete distributions. Let P^0 be an initial ("prior") distribution and P^1 some distribution to be compared with P^0. Define the Kullback-information in a shift from P^0 to P^1 by the formula

$$I_K(P^1, P^0) = \sum_{i=1}^{n} p_i^1 \cdot \log[p_i^1/p_i^0] \tag{6}$$

when P^0 is discrete, and by the analogous integral in densities

$$I_K(P^1, P^0) = \int_X p^1(x) \cdot \log[p^1(x)/p^0(x)]\, dx \tag{7}$$

for continuous distributions.

In the case of discrete distributions, (6) is related to (2) in a straightforward fashion. Whereas U_S purports to measure the residual uncertainty in a distribution, that is, U_S attempts to quantify how far a distribution is from certainty—how far a distribution is from a 0–1 probability—I_K reports the *decrease* in uncertainty in shifting from P^0 to P^1. If we set P^U as the uniform distribution over the finite space X of P^0 (so that P^U is the MAXENT distribution [no constraints] over X), and if we set P^* as a 0–1, point distribution over X (so that P^* depicts a state of certainty with respect to X), then

$$U_S(P^1) = I_K(P^*,P^U) - I_K(P^1,P^U). \tag{8}$$

(See Hobson and Cheng 1973.) Moreover, Hobson (1971) shows that I_K is characterized by five properties (three of which parallel Shannon's conditions for U_S). To wit, (up to a constant) I_K uniquely satisfies

(K_1) I_K is a continuous function of P^0 and P^1.
(K_2) When $P^0 = \{1/n,. . .,1/n\}$ and $P^1 = \{1/m,. . .,1/m,0,. . .,0\}$ ($m \leq n$), then I_K is increasing in n and decreasing in m.
(K_3) I_K is additive over decomposition of the sample space, analogous to (S_3).
(K_4) I_K is invariant over relabeling of the sample space.
(K_5) $I_K = 0$ just in case $P^0 = P^1$.

The remarks (pp. 470–71) about (S_1) − (S_3), and in particular the useful Result$_1$, apply to Kullback-information in parallel with the generalization of Shannon's three conditions by these five. Specifically, (K_3) (analogous to (S_3)) entails a restricted equivalence between Bayesian conditionalization and a minimum Kullback-information shift: where P^0 satisfies a constraint set C_0, and a minimum I_K-shift subject to the extra constraint of an event e_1 yields the revised probability P^1, then $P^1(\cdot) = P^0(\cdot|e_1)$ provided $P^0(\cdot|e_1)$ satisfies C_0.[3]

Just as in Result$_1$, this equivalence is relativized to cases where the conditional probability $P^0(\cdot|e_1)$ satisfies the initial constraints C_0. Where C_0 includes constraints other than the mere observation of events (measurable) in the space of P^0, the important proviso on $P^0(\cdot|e)$ fails for some events. Thus, unless the constraint set is restricted to 0–1 expectations for indicator variables, some (Bayesian) conditionalizations do not agree with the revision from P^0 to P^1 by minimizing the change in Kullback- (or Shannon-) information.

Besides generalizing U_S with discrete distributions, I_K affords a consistent extension of entropy to continuous distributions, unlike the (natural) continuous version of Shannon-uncertainty. That is, where we take

[3]Williams (1980) establishes the special case of this result when C_0 is vacuous.

a continuous version of (2) to be

$$U_S(P) = -\int_X p(x) \cdot \log[p(x)]dx \tag{9}$$

(with p the density for P), then it is well known (see Jaynes 1963) this attempt fails to provide consistent results over smooth transformations of continuous random variables. For example, if $\underset{\sim}{X}$ is confined to the unit interval [0,1], the use of (9) yields a MAXENT distribution uniform on [0,1]. However, if we consider the equivalent random variable $\underset{\sim}{Z}$, defined by $z = x^3$, then $\underset{\sim}{Z}$ (like $\underset{\sim}{X}$) is a continuous variable on [0,1], and (9) generates a MAXENT distribution for $\underset{\sim}{Z}$ uniform on [0,1]—in contradiction with the result for $\underset{\sim}{X}$.

By contrast, if we use I_K to identify minimum information shifts, once P^0 is identified, I_K remains invariant over the class of random variables equivalent to the one chosen for identifying P^0. Of course, in the continuous case the MAXENT program then requires a supplementary principle to fix P^0, where P^0 depicts a "state of ignorance" prior to the introduction of "constraints." Jaynes (1968, 1978, and 1980) (for example), is favorably disposed toward's Jeffreys's (1961) theory of invariants for this component of his MAXENT program. Unfortunately, the policy of using Jeffreys's invariants to fix such "prior" probabilities is inconsistent with basic Bayesian postulates. (See Seidenfeld 1979.) Thus, it remains an open question how to determine an "ignorance" prior for continuous distributions in a fashion consistent with Bayesian theory. Since my discussion in this essay pertains to discrete distributions, we may bypass this problem and use I_K as a generalized account of minimum change in probability.[4]

3. Entropy and Insufficient Reason: Repartitioning the Sample Space. A standard objection to the principle of Insufficient Reason is that it fails to provide consistent answers across simple reformulations of questions of interest. One cannot assign equal probability to disjoint events merely on the grounds that the question posed (together with tacit background assumptions—of fact) fails to include a good deductive reason

[4]For example, I_K provides an account of a minimum shift from a prior probability P^0 which is *not* itself identified as a solution to a MAXENT problem. In his recent (1983) paper, "Highly Informative Prior Probabilities," Jaynes make use of this generalization. If objective Bayesian theory is modulated to admit arbitrary (coherent) "prior" information as part of the "well-posed problem," then the basic dispute with subjectivist Bayesians (such as Savage) is resolved in favor of the latter point of view. That is, even Savage has no objection to a position that makes "objective" a posterior probability constrained by a prior probability and likelihood! Nonetheless, I remain dubious of the claim (v) that Bayesian theory is a special case of Kullback-information theory.

for selecting one answer over another. If you are "ignorant" about the outcome of a roll of a cubical die (with spots from 1 to 6 arranged in conventional order), then you may appeal to Insufficient Reason to assign each of the six outcomes: one-spot uppermost, . . ., six-spot uppermost, equal probability (1/6). Or, you can cite Insufficient Reason to partition the outcomes in two: one-spot uppermost, more than one-spot uppermost, and assign these possibilities equal probability (1/2). On its face, Insufficient Reason does not dictate which of these contrary analyses is appropriate.

Nor will it do to give priority to the more refined partition of possibilities merely on the grounds that added possibilities indicate more information about the circumstances. The added refinement may be both irrelevant and nonsymmetric to the basic question. Consider the standard, cubical die arranged with six numbered spots so that opposite sides sum to seven.[5] A roll of a die typically provides an observer with either two or three visible surfaces. In addition to the single side showing uppermost, the die displays one or two vertical faces as well. Let us partition outcomes as follows: for each of the six sides showing uppermost, characterize the roll also according to whether the *sum* of the visible spots on the side (vertically showing) faces(s) is (a) greater than, (b) equal to, or (c) less than the number of spots showing on the top face.

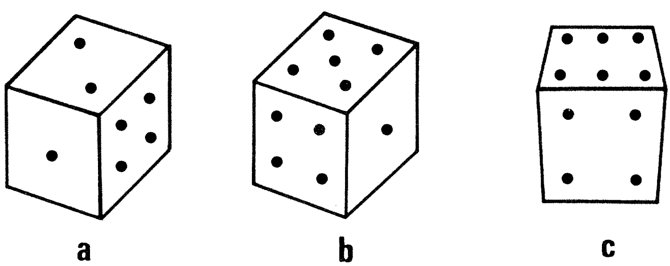

Figure 1. Repartitioning the sample space for a roll of the die.
Outcomes where the sum of visible, side-faces
(a) exceeds
(b) equals
(c) is less than
the top-face.

Instead of six outcomes this partitions the rolls into fourteen different possibilities (as displayed in table 1). These fourteen possibilities constitute a partition of all rolls with a standardly numbered die. Are we to apply Insufficient Reason to this refined partition (of the six familiar events) leading to a probability distribution (.07142, .14285, .21428, .14285,

[5]The arrangement of spots is further constrained so that a pair of dice may sum to seven on each of the six pairs of parallel faces; that is, dice are uniformly oriented.

TABLE 1

# Spots Showing on Upper Face of Die	Sum of Spots Visible on Side Face(s)		
i	$<i$	$=i$	$>i$
1	No	No	Yes
2	Yes	No	Yes
3	Yes	Yes	Yes
4	Yes	No	Yes
5	Yes	Yes	Yes
6	Yes	Yes	Yes

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

.21428, .21428) over the basic six outcomes of the roll (how the die landed)? If we believe that added refinement of possibilities reflects more information, then the fourteen-fold partition of states has priority in the application of Insufficient Reason. Of course, what is lacking is a judgment of relevance of the refinement introduced by considering the (nuisance) factor: sum of spots showing on side face(s).

Does the MAXENT program offer new guidance in this old problem? We noted (in discussion of Shannon's condition (S_2), p. 471) the well-known result that the uniform distribution $p_i = 1/n$ ($i = 1,\ldots,n$) maximizes entropy over all discrete distributions with $\Sigma_{i=1}^{n} p_i = 1$. Thus, MAXENT faces the same sensitivity to repartitions of the sample space as does the simpler principle of Insufficient Reason. Perhaps, in the absence of any constraints other than the number of possibilities, advocates of MAXENT can argue that refinement of possibilities by an observable (as with the modified sample space for the die, see table 1) does constitute new, relevant information. Unfortunately, the problem is not restricted to "a priori" MAXENT probability assignments. That is, the question of which partition is appropriate for application of MAXENT arises even when "constraints" are imposed.

As in Jaynes's example from his Brandeis Lectures (1963), restated in greater detail fifteen years later (1978), let us impose the "constraint"

$$E[\text{number of spots showing}] = {}^{55}/_{14} \simeq 3.9285. \tag{10}$$

If we apply MAXENT to the partition of outcomes by number of spots showing (up), that is, in the familiar six-fold partition, the distribution which maximizes entropy subject to (10) is (to five places—see the appendix)

$$(.11122, .12908, .14981, .17387, .20180, .23422) \tag{11}$$

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 fourteen states that maximizes entropy, subject to (12), is *not* one that yields a (marginal) distribution for the number of spots showing corresponding to (11). Instead MAXENT, applied to the refined partition, subject to the constraint (12) yields the (marginal) distribution for the number of spots showing:

$$(.07142, .14285, .21428, .14285, .21428, .21428).^{6} \tag{13}$$

The difference between these solutions can be conceptualized in the following terms. When the empirical "constraints" all involve a quantity (parameter) of interest, the MAXENT distribution for the parameter of interest is sensitive to which (refined) algebra of possibilities the investigator uses to solve the problem. Even though the investigator professes "ignorance" about the nuisance factor, and bases the MAXENT solution on the empirical constraints (all of which involve the parameter of interest alone), the MAXENT solution (like the principle of Insufficient Reason) changes with the addition of a refined partition of possibilities.

A sufficient condition for ensuring that the refinement does *not* affect the MAXENT solution is to make the refined algebra a product space in which the new factor (constituting the refinement) is probabilistically independent of the parameter of interest.[7] Where the nuisance factor is (for other reasons) required to be probabilistically dependent with the parameter of interest, this sufficient condition can be simulated by imposing a degenerate 0–1 marginal distribution on the nuisance factor. Then the nuisance factor is, in effect, a constant and constants are (vacuously) probabilistically independent of other variables. In section 4, where

[6]That is, the MAXENT solution to this problem corresponds to the uniform distribution over the fourteen states. That (13) is the MAXENT solution follows directly from the fact that $(1/n)\ \Sigma_{i=1}^{14} f(\text{state}_j) = {}^{55}/_{14}$. The uniform distribution over the fourteen states in table 1 satisfies condition (12). Recall that the uniform distribution maximizes entropy over all discrete distributions.

[7]In their application of MAXENT to estimating frequencies in contingency tables, subject to constraints of lower dimensional contingency tables, Denzau, Gibbons, and Greenberg (1984) note this independence is necessary for a coherent solution.

MAXENT is contrasted with Bayesian inference, the device of using a degenerate 0–1 distribution with nuisance factors is key to understanding an important objection raised by Friedman and Shimony (1971).

Summary. The question addressed in this section is prompted by claim (iv), that the MAXENT program provides a satisfactory account of the Laplacian principle of Insufficient Reason. That principle, in its simplest version, succumbs to inconsistencies when the space of possibilities is repartitioned and Insufficient Reason is applied to both algebras of possibilities. The same inconsistencies can arise with the MAXENT principle: (i) in the absence and (ii) in the presence of empirical constraints on the quantities of interest.

What is lacking is an account of how nuisance factors are judged for their relevance. Left to the MAXENT rule, the verdict is loaded in favor of relevance of the nuisance factor (since mere repartitioning is enough to affect MAXENT, as demonstrated above).[8] The problem is encapsulated in

Result$_2$. Given that constraints $C = \{c_i\}$ are a function of θ the parameter of interest alone, the MAXENT (marginal) distribution for θ may differ from the θ-marginalization of the (joint) MAXENT solution. That is, maximizing entropy in a marginal (average) distribution does not agree with marginalizing (averaging) the overall maximum entropy unless independence obtains between the parameter of interest and the nuisance parameter. The MAXENT solution is consistent with respect to marginalization only if the joint MAXENT distribution is a product of marginal MAXENT distributions.

Proof. By the construction above. However, Result$_2$ does not apply to cross-entropy (Kullback-information) shifts, as shown in the lemma (appendix B).

4. Entropy and Bayesian Theories. Claim (v) (p. 469) asserts the thesis that MAXENT inference subsumes Bayesian theory as a special case.

[8]This policy, to presume that changes that result from refinement of the algebra of possibilities reflect added *relevant* information in the refinement, seems to underlie Jaynes's (1980) analysis of the "marginalization paradoxes" (from Dawid, Stone, and Zidek (1973)). As Dawid, Stone and Zidek use their anomalies to question this policy (Does it work consistently for Bayesians using "improper" prior distributions?), it comes as no surprise to me that the involved parties accuse each other of missing the point (see the discussion and rebuttal to Jaynes's 1980).

For an alternative account of these "paradoxes," based on an interpretation of improper distributions as finite but not countably additive probabilities, see Sudderth (1980) and Kadane, Schervish, and Seidenfeld (1986).

To assess the claim we need guidelines for what counts as Bayesian inference. The point is not moot. (See Good 1971 for some 4.7×10^4 varieties.) I rest content here with some core postulates for Bayesian theory:

(B_1) An agent's belief state is represented by a coherent, finitely additive 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's 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, that is, under the hypothesis that BK is enlarged by addition of e (and its new consequences given BK)—conditionalization.[9]

We have Result$_1$ (p. 471—from Shannon's property (S_3)), establishing a restricted equivalence between revising probabilities through MAXENT and through conditionalization.[10] The restriction in Result$_1$ is that the "old" conditional MAXENT probability satisfy *all* the constraints and not merely the final constraint, newly imposed, which prompts the revision. Of course, if the restriction is satisfied then the constraint set is mutually consistent. Otherwise, not only is conditionalization at odds with a revised MAXENT solution, but where (in a sequence of shifts) constraints imposed at earlier stages are not retained in subsequent stages, the net Shannon or Kullback shift depends upon the order in which constraints are introduced and replaced. (See Tribus and Rossi 1973.) Hence, if a net Shannon or Kullback change is to be path-invariant over the order in which constraints are added, they must be mutually consistent—or else, some constraints must be dropped before others are introduced. (See Shore and Johnson 1981, property 13, and related discussion.)[11]

[9]This brief statement of the core postulates rides roughshod over several important subtleties in a proper formulation. In particular, I have not attended to temporal versus atemporal interpretations of (B_3)-conditionalization. See Levi (1981), and references cited there, for a careful discussion of such matters.

[10]Recall, too, this restricted equivalence extends to the Kullback-information approach—see section 2.2, p. 474.

[11]This problem is exacerbated by the unpleasant fact that I_K induces a semi-metric only—it does not satisfy the triangle inequality in general. (See Burbea and Rao 1982 for additional results.) In his (1968, §III) example of the distribution of impurities in a crystal lattice, Jaynes constructs a "prior" MAXENT solution from a constraint that is *not* satisfied by the "posterior" he obtains through data from a subsequent (neutron reflection) experiment. Thus, the question raised in this note has a basis in the current application of the MAXENT program. (I thank Professor E. Greenberg [Economics, Washington University] for the last reference.)

We may satisfy the restriction in Result_1 by limiting all constraints to 0–1 expectations for indicator variables (measurable) in the initial algebra of P_0. Then, by the corollary to Result_1, MAXENT reduces to Bayesian theory with a uniform "a priori" probability. If we use the Kullback-information generalization, the parallel result and corollary equates minimum-information shifts with Bayesian conditionalization from an arbitrary (a priori) probability.

Canonical illustrations of the MAXENT program, for instance Jaynes's Brandeis dice problem, use constraints that do *not* reduce to 0–1 expectations for indicator variables in the measure space of P_0. If, as a Bayesian, one hopes to understand a constraint as "evidence," then it is reasonable to ask whether, by extending the algebra, the constraint can be interpreted as an "event" in the larger algebra of possibilities. The question also holds out the hope that, in the extended algebra, there will be a Bayesian model for MAXENT formalism by application of Result_1 in the larger space of possibilities. Hence, if we are to consider the more interesting version of MAXENT theory (with an enriched language of constraints), the thesis that 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 interesting results of Friedman and Shimony (1971) (and Shimony's generalization of 1973). Let me rehearse their analysis in some detail.[12] Suppose we require a MAXENT probability distribution for a discrete space $X = \{x_1, \ldots, x_n\}$ of n-states ($n \geq 3$) based on "a priori" considerations, that is, in the absence of additional information about X. As noted (above), the uniform distribution $P(x_i) = p_i = 1/n$ ($i = 1, \ldots, n$) is the MAXENT solution. This result is not altered by adding the structure of distinct numerical magnitudes $f(x_i) = a_i$ to X, $a_i \neq a_j$ if $i \neq j$, so long as we profess ignorance

[12] I bother with the particulars since Jaynes ([1978] 1983, pp. 249–51) finds the F-S argument unacceptable. His complaint is that they use ill-defined constraints. In an otherwise patient review of several objections to the MAXENT program, he writes (following brief but general remarks about the difference between testable constraints and conditioning events),

> Of course, it is as true in probability theory as in carpentry that introduction of more powerful tools brings with it the obligation to exercise a higher level of understanding and judgment in using them. If you give a carpenter a fancy new power tool, he *may* use it to turn out more precise work in greater quantity; or he may just cut off his thumb with it. It depends upon the carpenter.
>
> The FS article led to considerably more discussion . . . in which severed thumbs proliferated like hydras; but the level of confusion about the points already noted is such that it would be futile to attempt any analysis of the FS arguments. (1983, pp. 250–51)

about, for example, an expected value $E[f]$.[13]

Next, suppose we imagine acquiring information about X, reported by a constraint $E[f] = r$ (where, of course, r lies between the minimum and maximum of the n-values $\{a_1,\ldots,a_n\}$). We may apply MAXENT to determine a new distribution $P_r(X)$, subject to the constraint $E[f] = r$. Friedman and Shimony ask, in effect, for necessary and sufficient conditions that $P_r(X)$ can be a conditional probability $P(X|$"$E[f] = r$") obtained by extending P to a field that includes the constraint as an event. Their findings are remarkable:

THEOREM [F-S]. *Subject to the conditions above, P can be so extended just in case* $P($"$E[f] = (1/n) \cdot \Sigma_{i=1}^{n} a_i$"$) = 1$. In words, P can be extended if and only if the extension makes the constraint, $E(f) = r =$ average of the a_i's, practically certain.[14]

A simple example brings home the point. Following Frieden (1984), let us simplify the "dice" problem by collapsing the space of outcomes to the three-sided die with 1, 2, and 3 spots (respectively) on each face. (Just identify a roll of a usual six-sided die by the *minimum* of the horizontal faces.) Then the MAXENT distribution for a roll of the die, based on "a priori" information, is the uniform ($^1/_3$, $^1/_3$, $^1/_3$) for each face. Suppose we quantify outcomes by identifying a state with the number of spots, $f(i\text{-spots}) = i$ $(i = 1,2,3)$. As in the dice example of section 1 (pp. 468–69), we can calculate a MAXENT solution for a constraint $E[f] = r$ $(1 \leq r \leq 3)$, denoted by $P_r(i)$. Since the average $(^1/_3) \cdot \Sigma_{i=1}^{3} i = 2$, the F-S Theorem dictates that, in extending P to make P_r a conditional probability, it is practically certain that $r = 2$.

If the constraint can be interpreted as fixing the center of gravity of the die as belonging to a region that makes $E[f] = r$ a correct statement of "chance," then the F-S result shows the MAXENT solution requires an a priori assignment of probability 1 to the empirical claim that the die is loaded so that $r = 2$. (See Shimony 1973.) I doubt this is the intended interpretation Jaynes wants for the "constraint" in his Brandeis Dice prob-

[13]In the 1971 version of this argument, there is the added premise that for one state, say the mth, $f(x_m) = (1/n) \cdot \Sigma_{i=1}^{n} f(x_i)$. That is, in the 1971 formulation, it is supposed there is one state whose magnitude a_m equals the average of the n magnitudes $a_m = (1/n) \cdot \Sigma_{i=1}^{n} a_i$. This condition is relaxed in Shimony's 1973 generalization. For the example which follows involving the trinomial "die," the 1971 applies. At the expense of complicating the calculations, Shimony's (1973) version is applicable to Jaynes's Brandeis Dice example as presented in Jaynes's 1983, pp. 243–45.

[14]The Friedman-Shimony proof uses disintegrability of P in the partition by r, the constraint. This assumption is not guaranteed for a general, finitely additive probability. But in the application to the Dice problem (and its generalizations), where r is the "sample average" in the first N rolls, this problem does not arise since, given N, r has a finite sample space.

lem.[15] So, instead, let us examine Jaynes's own interpretation of the constraint.

In his 1978 article (reprinted in his 1983) he writes,

> When a die is tossed, the number of spots up can have any value i in $1 \leq i \leq 6$. Suppose a die has been tossed N times and we are told only that the average number of spots up was not 3.5 as we might expect from an "honest" die but 4.5. Given this information, *and nothing else,* what probability should we assign to i spots in the next toss? (p. 244)

And in the discussion which follows, Jaynes uses the MAXENT distribution given the constraint to determine a predictive (subjective) distribution for the $N + 1$st roll. Thus, we can apply the F-S result to the problem of the three-sided die, in accord with Jaynes's proposal for interpreting the constraint. Let r be the "sample average" of the first N rolls. Then, $P_r(i)$ is a conditional probability for the $N + 1$st roll, in the extended product field X^{N+1} ($X = \{1,2,3\}$), just in case the a priori probability is 1 that $r = 2$.

A connection with the problem of section 3 is obvious. The question posed by Friedman and Shimony addresses the coherence of the MAXENT program by providing necessary and sufficient conditions for interpreting the MAXENT solution as a conditional marginal distribution in a refined (product) algebra that includes the "constraint" as a conditioning event. Not surprisingly, since the constraint is a relevant bit of information for fixing the (marginal) distribution of the $N + 1$st roll, coherence is achieved by converting the nuisance parameter (r) into a constant, almost surely. (See the discussion on pp. 475–76.) That is, the problem of repartitioning the algebra to permit the same MAXENT distribution for the $N + 1$st roll—both in the minimal field of X *and* in the product X^{N+1}—admits only degenerate solutions for the nuisance parameter, r, defined on the subfield X^N.

We can press the investigation further into the realm of Bayesian models. What if we allow the agent to hold an *exchangeable* probability for rolls of the die. (The probability P is exchangeable if, for any sub-sequence of n trials, P is invariant under permutations of the order of outcomes.) Then even the F-S solution is barred. That is:

Result$_3$. If $E(f) = r$ is a constraint imposed on the distribution for the $N + 1$st roll of an n-sided die ($n \geq 3$), based on the "sample average" from N (different) rolls, and P_r is this MAXENT solution, then there is

[15]Though his response to Rowlinson's (1970) question concerning Wolf dice data suggests the geometric interpretation above. (See Jaynes [1978] 1983, pp. 258–68.)

no exchangeable Bayesian model that makes P_r a conditional probability with $P(i \mid \text{"sample average"} = r) = P_r(i)$ $(i = 1,\ldots,n)$.

Proof (outline). According to de Finetti's representation theorem, such an exchangeable P is a mixture of i.i.d. multinomial distributions (each on a sample space of n outcomes) for some "mixing" prior π on the multinomial parameter. Recall, where $r = (n + 1)/2$, that is, when the "sample average" equals the average of the number of spots showing on the n faces of the die, then $P_r(i) = p_i = 1/n$, the uniform distribution. In other words, when the constraint satisfies $r = (n + 1)/2$, the MAXENT distribution for the $N + 1$st roll is the uniform probability, independent of the sample size N. Let Π be the class of "mixing" priors that satisfy this restriction, that is, where the conditional probability for the $N + 1$st outcome is uniform given that the "sample average" of the first N outcomes equals $(n + 1)/2$, for each N. Then π^+, the (degenerate) "mixing" prior that assigns probability 1 to the multinomial parameter $(1/n,\ldots,1/n)$, belongs to this class Π. Given N, verify that among $\pi \in \Pi$, π^+ maximizes the probability of the event $\{r = (n + 1)/2\}$. But, for this "mixing" prior (π^+), hence for all "mixing" prior in Π, the event $\{r = (n + 1)/2\}$ has probability less than 1. This contradicts the Friedman-Shimony theorem, establishing Result$_3$.[16] □

Summary (Section 4). In this section we investigate the coherence of MAXENT theory when the constraint set includes more than 0–1 expec-

[16]Frieden (1984) considers the case of a trinomial die with a uniform "prior" distribution for the multinomial parameter. He shows the interesting result that $P(i|r = m = 2) = 1/4$ for $i = 1, 3$ and $P(i|r = m = 2) = 1/2$ for $i = 2$, for all $N \geq 3$. The uniform prior corresponds to a Carnapian confirmation function c^* ($\lambda = 3$ in his continuum of inductive methods). Thus it is enlightening to compare Frieden's analysis with the Dias-Shimony result (1981), appendix B (B.5a) and (B.5b), for this case. Their results are in agreement, of course.

For contrast, I note that with an "improper" prior (whose density is $1/w_1 \cdot w_2 \cdot w_3$ for the multinomial parameter (w_1,w_2,w_3)), the predictive $P(i|r = m = 2)$ is likewise "improper," with all its mass concentrated at the extreme point (0,1,0), $i = 1, 2, 3$.

Note also, Result$_3$ (like the F-S theorem) depends upon the assumption $n \geq 3$. For $n = 2$, the "sample average" is a sufficient statistic with an exchangeable P (unlike the case with $n > 2$). Then the predictive probability $P_r(i) = P(i|r)$ has a Bayesian model with the "improper" prior (whose density is $1/w_1 \cdot w_2$), corresponding to the "straight-rule" in Carnap's (1951) continuum of inductive methods ($\lambda = 0$). But for $n = 2$, $P_r(i)$ then is determined *without* appeal to entropy considerations, since the class of distributions satisfying the constraint $E[i] = r$ is a unit set!

Lastly, Dias-Shimony (1981) prove a restricted agreement between MAXENT and Bayesian methods for the case of the trinomial die. Their theorem, §IV (4.10) shows that the extreme Carnapian method ($\lambda = \infty$), $c^\dagger$, is in asymptotic agreement (for increasing population sizes) with MAXENT solutions to select problems of direct inference. Result$_3$ demonstrates this agreement cannot be extended to simple problems of predictive inference. (Recall, $\lambda = \infty$ corresponds to the point-probability 1 for the multinomial parameter $(1/n,\ldots,1/n)$ in de Finetti's representation of Carnapian methods.)

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

tations for indicator variables (measurable) in the initial space of possibilities. The question asked is whether, by extending the algebra, MAXENT solutions have Bayesian models. The Friedman-Shimony result (1971) shows that where we attend to MAXENT solutions with even a single constraint (not a 0–1 expectaiton for an indicator variable), only degenerate Bayesian models exist. The degenerate Bayesian model is one in which the "constraint" is a nuisance parameter having, a priori, a 0–1 distribution. This agrees with the findings from section 3, dealing with repartitioning the sample space, where such degenerate solutions avoid the conflict reported in Result_2. Last, using Jaynes's recent (1978) presentation of his (1963) Brandeis Dice problem, we show there is no exchangeable (Bayesian) probability that preserves his recommended interpretation of the constraints—Result_3.

Whereas the MAXENT principle is sensitive to the choice of measure space (Result_2), that is not the case with cross-entropy (Kullback-information) shifts—see appendix B. However, the phenomenon pointed out in the Friedman-Shimony theorem (to wit: there are only degenerate Bayesian models that make "constraints" into events and make the MAXENT distributions into conditional probabilities given the constraints), does generalize to cross-entropy. This is shown in appendix B, Corollary_2 to Result_4. This finding uses a generalization of an observation from van Fraassen (1981).

5. Comments on the Concentration Theorem (Jaynes 1979, and see his 1963, pp. 51–52).

THEOREM (JAYNES). *Consider N repetitions of an experiment with n possible outcomes on a given trial. Let* f_i $(1 \leq i \leq n)$ *be the observed relative frequency of the i-th outcome in these N trials. Then the class of sequences of possible outcomes (from the N trials), satisfying a set of m constraints linear in these frequencies, is asymptotically (with* $N \to \infty$*) concentrated as* $\chi^2/2N$ *(with n-m-1 degrees of freedom) about the MAXENT distribution for the* f_i*'s. Here the "metric" across possible sequences of outcomes is given by the difference in the entropy of the corresponding* f_i*'s.*

I have two remarks to make about this interesting result.

First, unless there is some connection drawn between the long-run and short-run properties of the principle under question, mere asymptotics are insufficient for justification. To cite two (well-known) cases where asymptotic concerns prove inadequate because they lack relevance for the short run: neither the limiting-frequency definition of probability, nor the criterion of asymptotic consistency of point-estimates is well received. (See Fisher 1973, pp. 34–35 and pp. 148–49 for discussion of these two examples.) So, at least, the asymptotic argument needs to be supplemented with analysis of the rate at which concentration about the MAXENT distribution occurs. But then it is hard to understand how Jaynes will use the concentration theorem to defend application of MAXENT in statistical mechanics since he will need to show how to resolve the very problem he used in 1957 to undercut the grounding of statistical mechanics on ergodic theory. That is, to apply the concentration theorem in statistical mechanics, Jaynes needs to show, for example, what are the appropriate time intervals to use to achieve "concentration" about the MAXENT distribution.

A second objection to the argument that seeks justification of the MAXENT rule by appeal to claim (ii) is based on consideration of how, in Jaynes's (1979) result, limiting frequencies from repeated trials are contrasted with a subjective, MAXENT probability for a single trial. The concentration theorem establishes that, relativized to the given constraints (interpreted with limiting frequencies as probabilities), the class of limiting frequencies concentrate (in the sense of having entropy) close to the MAXENT distribution for a single trial. Apart from the important question why "constraints" on a MAXENT probability for a single trial translate into parallel conditions on limiting frequencies from repeated trials (see also note 16 in section 4), there is the following difficulty with the attempted justification.

After relativizing the class of possible limiting frequencies to those satisfying the given constraints, we are directed to count each *logically* possible sequence of repeated trials as a separate state. Then the concentration about 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 a problem because, if the concentration about the MAXENT solution demonstrates how highly *probable* the MAXENT solution is, then as Jaynes's points out (1979, p. 322), the argument equates possibility with probability. In other words, if the concentration theorem is to show how probabilistically atypical "low" entropy distributions are (in repeated trials), logically distinct sequences must be judged equally probable.

An assignment of equal probability to distinct states characterizes an extreme Carnapian method, $\lambda = \infty$, whose Bayesian description is of an i.i.d. process with uniform ($p = 1/n$) probability for each of the n outcomes of a single trial. Recall that the "a priori" MAXENT probability over n outcomes is the uniform distribution, $p_i = 1/n$ ($i = 1, \ldots, n$). By the strong law of large numbers, we know that in an i.i.d. process, with probability 1 the limiting relative frequencies concentrate about this "a priori" distribution. If we restrict the limiting relative frequencies, so they satisfy the constraints imposed on the MAXENT solution, then the concentration, given the constraints, is at the MAXENT distribution. But, even here the argument depends upon the uniform "prior" probability, corresponding to $\lambda = \infty$. (See Dias and Shimony 1981, pp. 192–93, for related discussion.)

The point of the objection is that, were the argument modified by choosing a different "prior" in place of the uniform one, the law of large numbers would continue to hold—in an i.i.d. process there still would be a concentration of limiting frequencies about the "a priori" probability and a related, conditional concentration given frequency constraints. Of course, with the change in "prior," the concentration of frequencies would not be determined simply by the *proportion* of sequences close to the "prior," but by some weighted proportion in which sequences were assigned unequal probability as dictated by the "prior." In short, the concentration theorem singles out MAXENT whenever distinct sequences are counted equally; however, by tailoring the weights on sequences to the "prior" chosen, we can defend any "prior" by a concentration-of-frequencies result. How does claim (ii) distinguish MAXENT from rival (Bayesian) methods?

APPENDIX A: ON THE MAXENT FORMALISM.

Here we review some of the mathematics for calculating MAXENT solutions. Following Shore and Johnson (1980), a constraint is an expectation (linear in probability) for a bounded function of the state variables. (We use only linear, equality constraints, $E[f] = c$, instead of the more general class including inequalities too.) Hence, the class of distributions satisfying a (finite) set of constraints is convex. Thus $c_j = \Sigma_{i=1}^{n} p_i f_j(x_i)$ is the jth constraint.

With k constraints, $c_1, \ldots, c_k$ the matter of choosing a distribution that satisfies these constraints and maximizes entropy is a variational problem (familiar in physics), solved by the device of Lagrange multipliers. (See Courant and Hilbert 1963, pp. 164–74.) The formal solution obeys:

$$p_i = P(x_i) = [Z(\lambda_1, \ldots, \lambda_k)]^{-1} \cdot \exp[-\lambda_1 f_1(x_i) - \ldots - \lambda_k f_k(x_i)] \quad \text{(A1)}$$

where

$$Z(\lambda_1, \ldots, \lambda_k) = \sum_{i=1}^{n} \exp[-\lambda_1 f_1(x_i) - \ldots - \lambda_k f_k(x_i)], \quad \text{(A2)}$$

and the λ's are the Lagrange multipliers, chosen to satisfy the k constraints, that is,

$$c_j = -\frac{\partial}{\partial \lambda_j} \log Z. \quad \text{(A3)}$$

In the case of Jaynes's Brandeis Dice problem, there is one constraint arising from the expectation for the function $f(i\text{-spots}) = i$ $(i = 1, \ldots, 6)$, so that

$$c_1 = \sum_{i=1}^{6} i \cdot p(i\text{-spots}). \quad \text{(A4)}$$

As Jaynes shows (see his [1978] 1980, p. 244)

$$Z(\lambda_1) = \sum_{i=1}^{6} e^{-\lambda_1 i} = x(1 - x^6)/(1 - x) \quad \text{(A5)}$$

where $x = e^{-\lambda 1}$. (The r.h.s. of (A5) is by the usual rule for geometric series.) Then, by (A3) and (A4)

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

In the problem discussed on p. 477, (10) sets the constraint: $c_1 = {}^{55}/_{14}$. Solving (A6) for this value yields:

$$x \simeq 1.160601 \quad Z \simeq 10.43509 \tag{A7}$$

(as obtained on my TI 58C calculator). This results in the MAXENT distribution (11) on p. 477, in accord with (A1). The MAXENT distributions (5.1) and (5.2) are calculated in the identical manner.

It is interesting to note, as reported in Denzau, Gibbons and Greenberg (1984), the MAXENT solution (A1) is associated with a LOGIT model by a simple reidentification of parameters. (See the interesting papers in Manski and McFadden (1981) for a very helpful discussion of the role played by LOGIT models in econometric models of composite data from individual decision problems.)

APPENDIX B: ON MINIMUM INFORMATION SHIFTS ARISING FROM THE SPECIFICATION OF NEW CONDITIONAL PROBABILITIES

Recall, the entropy in a distribution P is given by

$$-\Sigma_i p(x_i) \cdot \log p(x_i),$$

and the cross-entropy (or Kullback-information) in a shift from P^0 to P^1 is given by

$$\Sigma_i p^1(x_i) \cdot \log [p^1(x_i)/p^0(x_i)].$$

Result$_1$. Let $X = (x_+, \ldots, x_n)$ with $x_i \cap x_j = \emptyset$ for $i \neq j$ and $n \geq 3$. Let $E_1, E_2 \subset X$ with $E_1 \cap E_2 = \emptyset$ and $X - (E_1 \cup E_2) = E_3 \neq \emptyset$. Let $N = (1, \ldots, n)$, and choose $I_1, I_2 \subset N$, with $I_1 \cap I_2 = \emptyset$ so that $E_i = \cup_{i \in I_i} x_i$ $(i = 1,2)$. Assume $|E_1| = k$ and $|E_2| = m$, so $k + m < n$. Specify a constraint c: $p(E_1)/p(E_2) = (1 - \alpha)/\alpha$. If P is the MAXENT solution subject to constraint, c, then either $P(E_3) > [n - (k + m)]/n$, or else $\alpha = m/(k + m)$ when P is the uniform distribution and $P(E_3) = [n - (k + m)]/n$.

COROLLARY$_1$. *Let P^0 be a probability on X. Let E_1 and E_2 be as above. Let P^1 be a minimum Kullback-information (cross-entropy) shift from P^0 subject to the constraint c, as above. Then $P^1(E_3) > P^0(E_3)$, unless $P^1 = P^0$. (Note: Van Fraassen 1981 gives a direct argument for this corollary in the special case $P^0(E_1) = P^0(E_2) = .25$. His analysis makes tacit use of the lemma* [*below*].)

Proof of Corollary$_1$. The corollary follows from Result$_4$ and a simple lemma about cross-entropy.

LEMMA. *Let P^0 be a distribution on X, and let P^1 be a minimum cross-entropy shift from P^0 subject to the set of constraints* C. *Let Y be a refinement of X, that is, $\forall x \in X$ ($x \subset Y$). Let $P^0{}_Y$ be a distribution on Y that agrees with P^0 on X, and let C_Y be the reformulation of* C *in the measure space generated by Y. If $P^1{}_Y$ is the minimum cross-entropy shift from $P^0{}_Y$ subject to C_Y, then $P^1{}_Y$ agrees with P^1 on X.* (I thank Ben Wise, of Carnegie-Mellon University, for raising the question of this lemma.)

Proof of the Lemma. This is immediate from the additive decomposition of a cross-entropy shift from P^0 to P^1 into a sum of a marginal cross-entropy shift and an expected P^1 shift in conditional probability. □

(*Proof of Corollary*$_1$—continued.) Without loss of generality, assume $P^0{}_X$ is rational-valued. (Otherwise, consider a sequence $\langle P_i^0{}_X \rangle$ of rational-valued probabilities converging to $P^0{}_X$) Refine X to Y so that $P^0{}_Y$ is uniform on 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,Y), subject to the constraint c. Then apply Result$_4$ to show that $P^1{}_Y$ has the desired property on E_3. To wit: $P^1{}_Y(E_3) > P^0{}_Y(E_3)$, unless $P^1{}_Y = P^0{}_Y$. □

Now, for the *proof of Result$_4$*.

Let X, E_1, E_2, and E_3 be as stated. Introduce the constraint of new conditional odds via "called-off" bets. That is, define

$$\begin{aligned} f(x_i) &= -\alpha && \text{if } x_i \in E_1 \\ &= (1-\alpha) && \text{if } x_i \in E_2 \\ &= 0 && \text{if } x_i \in E_3. \end{aligned}$$

The constraint c, then, is formulated by: $E[f] = 0$. Distributions satisfying this constraint also satisfy $P(E_1)/P(E_2) = (1 - \alpha)/\alpha$. The MAXENT distribution subject to c, denoted by P, is determined through the equation

$$P(x_i) = e^{-\lambda f(x_i)} \cdot Z^{-1} \tag{1}$$

where $Z(\lambda) = ky^{-\alpha} + my^{(1-\alpha)} + (n - k - m)$ (2)

for $y = e^{-\lambda}$

and $c = 0 = -\dfrac{d \log Z(\lambda)}{d\lambda}$. (3)

Then $y = (\alpha/1 - \alpha) \cdot (k/m)$. (4)

Substituting (4) into (1), we arrive at

$$P(E_1 \cup E_2) \cdot Z = k[m(1-\alpha)/k\alpha)]^\alpha + m[k\alpha/(m(1-\alpha))]^{1-\alpha} \tag{5}$$

$$\leq k + m. \tag{6}$$

The inequality in (6) is strict unless $\propto = m/(k + m)$, when P is the uniform distribution, P^u, on X. □

COROLLARY$_2$. *With X, E_1, E_2, $P^0{}_X$ and c as above, the only coherent probability that makes $P^1{}_X$ into a conditional probability $P_W(\cdot|$"c") in some measure space* (W,W) *which extends* (X,X), *is where $P_W((1 - \alpha)/\alpha = P^0{}_X(E_1)/P^0{}_X(E_2)) = 1$.*

That is, for coherence, with probability 1 the constraint c is irrelevant to P^0. This corollary augments the Friedman-Shimony (1971) and Shimony (1973) theorems by generalizing their observation to cross-entropy shifts from arbitrary probability distributions.

Proof. Note that for any value of α other than the one irrelevant to P^0 the P^1-probability of the event E_3 increases. Thus, no probability mixture of the conditional probabilities $P_W(E_3|$"c") can equal the unconditional probability $P_W(E_3)(= P^0{}_X(E_3))$ unless "c" is irrelevant to $P_W(E_3)$ almost surely. □

REFERENCES

Burbea, J., and Rao, C. R. (1982), "On the Convexity of Some Divergence Measures Based on Entropy Functions", *IEEE Transactions on Information Theory IT-28* (3): 489–95.

Carnap, R. (1952), *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Courant, R., and Hilbert, D. (1963), *Methods of Mathematical Physics*. Vol. 1. New York: Interscience Publishers.

Dawid, A. P.; Stone, M.; and Zidek, J. V. (1973), "Marginalization Paradoxes in Bayesian and Structural Inference", *Journal of the Royal Statistical Society Series B35*: 189–233; with discussion.

Denzau, A. T.; Gibbons, P. C.; and Greenberg, E. (1984), "Bayesian Estimation of Proportions with an Entropy Prior", Department of Economics, Washington University, St. Louis.

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

Fisher, R. A. (1973), *Statistical Methods and Scientific Inference*. Third edition. New York: Hafner.

Frieden, B. R. (1972), "Restoring with Maximum Likelihood and Maximum Entropy", *Journal of the Optical Society of America 62*: 511–18.

——— (1984), "Dice, Entropy and Likelihood", Optical Sciences Center, University of Arizona, Tucson, Arizona.

Friedman, K., and Shimony, A. (1971), "Jaynes's Maximum Entropy Prescription and Probability Theory", *Journal of Statistical Physics 3*: 381–84.

Good, I. J. (1971), "46656 Varieties of Bayesians", *American Statistician 25*: 62–63. (Reprinted in *Good Thinking*, Minneapolis: University of Minnesota Press, 1983.)

Hobson, A. (1971), *Concepts in Statistical Mechanics*. New York: Gordon and Breach.

Hobson, A., and Cheng, Bin-Kang (1973), "A Comparison of the Shannon and Kullback Information Measures", *Journal of Statistical Physics 7*: 301–10.

Jaynes, E. T. (1957), "Information Theory and Statistical Mechanics", I and II, *Physical Review 106*: 620–30; *108*: 171–90. (Reprinted in Jaynes 1983.)

——— (1963), "Information Theory and Statistical Mechanics", *1962 Brandeis Summer Institute in Theoretical Physics*, K. Ford (ed.). New York: Benjamin. (Reprinted in Jaynes 1983.)

——— (1978), "Where Do We Stand on Maximum Entropy?", in *The Maximum Entropy Formalism*, R. D. Levine and M. Tribus (eds.). Cambridge: The MIT Press. (Reprinted in Jaynes 1983.)

——— (1979), "Concentration of Distributions at Entropy Maxima", in Jaynes 1983.

——— (1980), "Marginalization and Prior Probabilities", in *Bayesian Analysis in Econometrics and Statistics*, A. Zellner (ed.). Amsterdam: North-Holland. (Reprinted in Jaynes 1983.)

——— (1981), "What Is the Question?", in *Bayesian Statistics*, J. M. Bernardo et al. (eds.). Valencia Spain: University of Valencia Press. (Reprinted in Jaynes 1983.)

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

——— (1983), "Highly Informative Priors", in *Proceedings of the Second Valencia International Meeting on Bayesian Statistics*, J. M. Bernardo et al. (eds.).

Jeffreys, H. (1961), *Theory of Probability*. Third edition. Oxford: Oxford University Press.

Kadane, J.; Schervish, M.; and Seidenfeld, T. (1986), "Statistical Implications of Finitely Additive Probability", in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, P. K. Goel and A. Zellner (eds.). Amsterdam: North-Holland.

Kullback, S. (1951), *Information Theory and Statistics*. New York: Wiley.

Levi, I. (1981), "Direct Inference and Confirmational Conditionalization", *Philosophy of Science 48*: 532–52.

Manski, C. F., and McFadden, D. (eds.). (1981), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: The MIT Press.

Rosenkrantz, R. (1977), *Inference, Method and Decision*. Dordrecht: D. Reidel Publishing.

Rowlinson, J. (1970), "Probability, Information, and Entropy", *Nature 255*: 1196–98.

Seidenfeld, T. (1979), "Why I Am Not an Objective Bayesian", *Theory and Decision 11*: 413–40.

Shannon, C. (1948), "A Mathematical Theory of Communication", *Bell System Technical Journal 27*: 379–423, 623–56.

Shimony, A. (1973), "Comment on the Interpretation of Inductive Probabilities", *Journal of Statistical Physics 9*: 187–91.

Shore, J., and Johnson, R. (1980), "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy", *IEEE Transactions on Information Theory IT-26* (1): 26–37.

——— (1981), "Properties of Cross-Entropy Minimization", *IEEE Transactions on Information Theory IT-27* (4): 472–82.

Sudderth, W. (1980), "Finitely Additive Priors, Coherence and the Marginalization Paradox", *Journal of the Royal Statistical Society B 42*: 339–41.

Tribus, M., and Rossi, R. (1973), "On the Kullback Information Measure as a Basis for Information Theory: Comments on a Proposal by Hobson and Chang", *Journal of Statistical Physics 9*: 331–38.

van Fraassen, B. (1981), "A Problem for Relative Information Minimizers in Probability Kinematics", *British Journal for Philosophy of Science 34*: 375–79.

Wiener, N. (1948), *Cybernetics*. New York: Wiley.

Williams, P. M. (1980), "Bayesian Conditionalisation and the Principle of Minimum Information", *British Journal for Philosophy of Science 31*: 131–44.