Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy

JOHN E. SHORE, MEMBER, IEEE, AND RODNEY W. JOHNSON

Abstract—Jaynes's principle of maximum entropy and Kullback's principle of minimum cross-entropy (minimum directed divergence) are shown to be uniquely correct methods for inductive inference when new information is given in the form of expected values. Previous justifications use intuitive arguments and rely on the properties of entropy and cross-entropy as information measures. The approach here assumes that reasonable methods of inductive inference should lead to consistent results when there are different ways of taking the same information into account (for example, in different coordinate systems). This requirement is formalized as four consistency axioms. These are stated in terms of an abstract information operator and make no reference to information measures. It is proved that the principle of maximum entropy is correct in the following sense: maximizing any function but entropy will lead to inconsistency unless that function and entropy have identical maxima. In other words, given information in the form of constraints on expected values, there is only one distribution satisfying the constraints that can be chosen by a procedure that satisfies the consistency axioms; this unique distribution can be obtained by maximizing entropy. This result is established both directly and as a special case (uniform priors) of an analogous result for the principle of minimum cross-entropy. Results are obtained both for continuous probability densities and for discrete distributions.

I. Introduction

We prove that Jaynes's principle of maximum entropy and Kullback's principle of minimum cross-entropy (minimum directed divergence) are correct methods of inference when given new information in terms of expected values. Our approach does not rely on intuitive arguments or on the properties of entropy and cross-entropy as information measures. Rather, we consider the consequences of requiring that methods of inference be self-consistent.

A. The Maximum Entropy Principle and the Minimum Cross-Entropy Principle

Suppose you know that a system has a set of possible states x_i with unknown probabilities $q^\dagger(x_i)$, and you then learn *constraints* on the distribution $q^\dagger$: either values of certain expectations $\Sigma_i q^\dagger(x_i) f_k(x_i)$ or bounds on these values. Suppose you need to choose a distribution q that is in some sense the best estimate of $q^\dagger$ given what you know. Usually there remains an infinite set of distributions that are not ruled out by the constraints. Which one should you choose?

The principle of maximum entropy states that, of all the distributions q that satisfy the constraints, you should choose the one with the largest entropy $-\Sigma_i q(x_i)\log(q(x_i))$. Entropy maximization was first proposed as a general inference procedure by Jaynes [1], although it has historical roots in physics (e.g., Elasser [67]). It has been applied successfully in a remarkable variety of fields, including statistical mechanics and thermodynamics [1]–[8], statistics [9]–[11, ch. 6], reliability estimation [11, ch. 10], [12], traffic networks [13], queuing theory and computer system modeling [14], [15], system simulation [16], production line decisionmaking [17], [18], computer memory reference patterns [19], system modularity [20], group behavior [21], stock market analysis [22], and general probabilistic problem solving [11], [17], [23]–[25]. There is much current interest in maximum entropy spectral analysis [26]–[29].

The principle of minimum cross-entropy is a generalization that applies in cases when a prior distribution p that estimates $q^\dagger$ is known in addition to the constraints. The principle states that, of the distributions q that satisfy the constraints, you should choose the one with the least cross-entropy $\Sigma_i q(x_i)\log(q(x_i)/p(x_i))$. Minimizing cross-entropy is equivalent to maximizing entropy when the prior is a uniform distribution. Unlike entropy maximization, cross-entropy minimization generalizes correctly for continuous probability densities. One then minimizes the functional

$$H(q,p)=\int dx\, q(x)\log(q(x)/p(x)). \tag{1}$$

The name cross-entropy is due to Good [9]. Other names include expected weight of evidence [30, p. 72], directed divergence [31, p. 7], and relative entropy [32]. First proposed by Kullback [31, p. 37], the principle of minimum cross-entropy has been advocated in various forms by others [9], [33], [34], including Jaynes [3], [25], who obtained (1) with an "invariant measure" playing the role of the prior density. Cross-entropy minimization has been applied primarily to statistics [9], [31], [35], [36], but also to statistical mechanics [8], chemistry [37], pattern recognition [38], [39], computer storage of probability distributions [40], and spectral analysis [41]. For a general discussion and examples of minimizing cross-entropy subject to constraints, see [42, appendix B]. APL computer programs

Manuscript received October 23, 1978; revised March 5, 1979.

The authors are with the Naval Research Laboratory, Washington, DC 20375.

for finding minimum cross-entropy distributions given arbitrary priors and constraints are described in [43]. Both entropy maximization and cross-entropy minimization have roots in Shannon's work [44].

B. Justifying the Principles as General Methods of Inference

Despite its success, the maximum entropy principle remains controversial [32], [45]–[49]. The controversy appears to stem from weaknesses in the foundations of the principle, which is usually justified on the basis of entropy's unique properties as an uncertainty measure. That entropy has such properties is undisputed; one can prove, up to a constant factor, that entropy is the only function satisfying axioms that are accepted as requirements for an uncertainty measure [44, pp. 379–423], [50], and [51]. Intuitively, the maximum entropy principle follows quite naturally from such axiomatic characterizations. Jaynes states that the maximum entropy distribution "is uniquely determined as the one which is maximally noncommittal with regard to missing information" [1, p. 623], and that it "agrees with what is known, but expresses 'maximum uncertainty' with respect to all other matters, and thus leaves a maximum possible freedom for our final decisions to be influenced by the subsequent sample data" [25, p. 231]. Somewhat whimsically, Benes justified his use of entropy maximization as "a reasonable and systematic way of throwing up our hands" [13, p. 234]. Others argue similarly [5]–[9], [11]. Jaynes has further supported entropy maximization by showing that the maximum entropy distribution is equal to the frequency distribution that can be realized in the greatest number of ways [25], an approach that has been studied in more detail by North [52].

Similar justifications can be advanced for cross-entropy minimization. Cross-entropy has properties that are desirable for an information measure [33], [34], [53], and one can argue [54] that it measures the amount of information necessary to change a prior p into the posterior q. Cross-entropy can be characterized axiomatically, both in the discrete case [8], [54]–[56] and in the continuous case [34]. The principle of cross-entropy minimization then follows intuitively much like entropy maximization. In an interesting recent paper [58] Van Campenhout and Cover have shown that the minimum cross-entropy density is the limiting form of the conditional density given average values.

To some, entropy's unique properties make it obvious that entropy maximization is the correct way to account for constraint information. To others, such an informal and intuitive justification yields plausibility but not proof —why maximize entropy; why not some other function?

Such questions are not answered unequivocally by previous justifications because they argue indirectly. Most are based on a formal description of what is required of an information measure; none are based on a formal description of what is required of a method for taking information into account. Since the maximum entropy principle is asserted as a general method of inductive inference, it is reasonable to require that different ways of using it to take the same information into account should lead to consistent results. We formalize this requirement in four consistency axioms. These are stated in terms of an abstract information operator; they make no reference to information measures.

We then prove that the maximum entropy principle is correct in the following sense: maximizing any function but entropy will lead to inconsistencies unless that function and entropy have identical maxima (any monotonic function of entropy will work, for example). Stated differently, we prove that, given new constraint information, there is only one distribution satisfying these constraints that can be chosen by a procedure that satisfies the consistency axioms; this unique distribution can be obtained by maximizing entropy. We establish this result both directly and as a special case of an analogous result for the principle of minimum cross-entropy; we prove that, given a continuous prior density and new constraints, there is only one posterior density satisfying these constraints that can be chosen by a procedure that satisfies the axioms; this unique posterior can be obtained by minimizing cross-entropy.

Informally, our axioms may be phrased as follows.

I. *Uniqueness*: The result should be unique.

II. *Invariance*: The choice of coordinate system should not matter.

III. *System Independence*: It should not matter whether one accounts for independent information about independent systems separately in terms of different densities or together in terms of a joint density.

IV. *Subset Independence*: It should not matter whether one treats an independent subset of system states in terms of a separate conditional density or in terms of the full system density.

These axioms are all based on one fundamental principle: if a problem can be solved in more than one way, the results should be consistent.

Our approach is analogous to work of Cox [59], [60], [11, ch. 1] and similar work of Janossy [61], [62]. From a requirement that probability theory provide a consistent model of inductive inference, they derive functional equations whose solutions include the standard equations of probability theory. Emphasizing invariance, Jeffreys [63] takes the same premise in studying the choice of priors.

C. Outline

The remainder of the paper is organized as follows. In Section II we introduce some definitions and notation. In Section III we motivate and formally state the axioms. Their consequences for continuous densities are explored in Section IV; a series of theorems culminates in our main result justifying the principle of minimum cross-entropy. The discrete case, including the principle of maximum entropy, is discussed in Section V. Section VI contrasts axioms of inference methods with axioms of information measures and contains concluding remarks. A more detailed exposition of our results is contained in [42].

II. Definitions and Notation

To formalize inference about probability densities that satisfy arbitrary expectation constraints, we need a concise notation for such constraints. We also need a notation for the procedure of minimizing some functional to choose a posterior density. We therefore introduce an abstract information operator that yields a posterior density from a prior density and new constraint information. We can then state inference axioms in terms of this operator.

We use lowercase boldface roman letters for system states, which may be multidimensional, and uppercase boldface roman letters for sets of system states. We use lowercase roman letters for probability densities and uppercase script letters for sets of probability densities. Thus, let $\boldsymbol{x}$ be a state of some system that has a set $\boldsymbol{D}$ of possible states. Let $\mathcal{D}$ be the set of all probability densities q on $\boldsymbol{D}$ such that $q(\boldsymbol{x}) \geqslant 0$ for $\boldsymbol{x} \in \boldsymbol{D}$ and

$$\int_{\boldsymbol{D}} d\boldsymbol{x}\, q(\boldsymbol{x}) = 1. \tag{2}$$

We use a superscript dagger to distinguish the system's unknown "true" state probability density $q^{\dagger} \in \mathcal{D}$. When $\boldsymbol{S} \subseteq \boldsymbol{D}$ is some set of states, we write $q(\boldsymbol{x} \in \boldsymbol{S})$ for the set of values $q(\boldsymbol{x})$ with $\boldsymbol{x} \in \boldsymbol{S}$.

New information takes the form of linear *equality constraints*

$$\int_{\boldsymbol{D}} d\boldsymbol{x}\, q^{\dagger}(\boldsymbol{x}) a_k(\boldsymbol{x}) = 0 \tag{3}$$

and *inequality constraints*

$$\int_{\boldsymbol{D}} d\boldsymbol{x}\, q^{\dagger}(\boldsymbol{x}) c_k(\boldsymbol{x}) \geqslant 0 \tag{4}$$

for known sets of bounded functions a_k and c_k. The probability densities that satisfy such constraints always comprise a closed convex subset of $\mathcal{D}$. (A set $\mathcal{I} \subseteq \mathcal{D}$ is *convex* if, given $0 \leqslant A \leqslant 1$ and $q, r \in \mathcal{I}$, it contains the weighted average $Aq + (1-A)r$.) Furthermore, any closed convex subset of $\mathcal{D}$ can be defined by equality and inequality constraints, perhaps infinite in number. We express constraints in these terms, using the notation $I = (q^{\dagger} \in \mathcal{I})$, to mean that $q^{\dagger}$ is a member of the closed convex set $\mathcal{I} \subseteq \mathcal{D}$. We refer to I as a *constraint* and to $\mathcal{I}$ as a *constraint set*. We use uppercase roman letters for constraints.

Let $p \in \mathcal{D}$ be some *prior* density that is an estimate of $q^{\dagger}$ obtained, by any means, prior to learning I. We require that priors be strictly positive:

$$p(\boldsymbol{x} \in \boldsymbol{D}) > 0. \tag{5}$$

(This restriction is discussed below.) Given a prior p and new information I, the *posterior* density $q \in \mathcal{I}$ that results from taking I into account is chosen by minimizing a functional $H(q,p)$ in the constraint set $\mathcal{I}$:

$$H(q,p) = \min_{q' \in \mathcal{I}} H(q',p). \tag{6}$$

We introduce an "information operator" $\circ$ that expresses (6) using the notation

$$q = p \circ I. \tag{7}$$

The operator $\circ$ takes two arguments—a prior and new information—and yields a posterior. For some other functional $F(q,p)$, suppose q satisfies (6) if and only if it satisfies

$$F(q,p) = \min_{q' \in \mathcal{I}} F(q',p).$$

Then we say that F and H are *equivalent*. If F and H are equivalent, the operator $\circ$ can be realized using either functional.

If H has the form (1), then (7) expresses the principle of minimum cross-entropy. At this point, however, we assume only that H is some well-behaved functional. In Section III we give consistency axioms for $\circ$ that restrict the possible forms of H. We say that a functional H *satisfies* one of these axioms if the axiom is satisfied by the operator $\circ$ that is realized using H.

In making the restriction (5) we assume that $\boldsymbol{D}$ is the set of states that are possible according to prior information. We do not impose a similar restriction on the posterior $q = p \circ I$ since I may rule out states currently thought to be possible. If this happens, then $\boldsymbol{D}$ must be redefined before q is used as a prior in a further application of $\circ$. The restriction (5) does not significantly restrict our results, but it does help in avoiding certain technical problems that would otherwise result from division by $p(\boldsymbol{x})$. For similar reasons—avoidance of technically troublesome singular cases—we impose on the information I the restriction that there exists at least one density $q \in \mathcal{I}$ with $H(q,p) < \infty$.

For some subset $\boldsymbol{S} \subseteq \boldsymbol{D}$ of states and $\boldsymbol{x} \in \boldsymbol{S}$, let

$$q(\boldsymbol{x} | \boldsymbol{x} \in \boldsymbol{S}) = q(\boldsymbol{x}) \Big/ \int_{\boldsymbol{S}} d\boldsymbol{x}'\, q(\boldsymbol{x}') \tag{8}$$

be the *conditional density*, given $\boldsymbol{x} \in \boldsymbol{S}$, corresponding to any $q \in \mathcal{D}$. We use

$$q(\boldsymbol{x} | \boldsymbol{x} \in \boldsymbol{S}) = q * \boldsymbol{S} \tag{9}$$

as a shorthand notation for (8).

When $\boldsymbol{D}$ is a discrete set of system states, densities are replaced by discrete distributions and integrals by sums in the usual way. We use lowercase boldface roman letters for discrete probability distributions, which we consider to be vectors; for example, $\boldsymbol{q} = q_1, \cdots, q_n$. It will always be clear in context whether, for example, the symbol $\boldsymbol{r}$ refers to a system state or a discrete distribution and whether s_i refers to a probability density or a component of a discrete distribution.

III. The Axioms

We follow the formal statement of each axiom with a justification. We assume, throughout, a system with possible states $\boldsymbol{D}$ and probability density $q^{\dagger} \in \mathcal{D}$.

Axiom I (Uniqueness): The posterior $q = p \circ I$ is unique for any prior $p \in \mathcal{D}$ and new information $I = (q^\dagger \in \mathcal{I})$, where $\mathcal{I} \subseteq \mathcal{D}$.

Justification: If we solve the same problem twice in exactly the same way, we expect the same answer to result both times. Actually, Axiom I is implicit in our notation.

Axiom II (Invariance): Let $\boldsymbol{\Gamma}$ be a coordinate transformation from $\boldsymbol{x} \in \boldsymbol{D}$ to $\boldsymbol{y} \in \boldsymbol{D}'$ with $(\boldsymbol{\Gamma} q)(\boldsymbol{y}) = \boldsymbol{J}^{-1} q(\boldsymbol{x})$, where $\boldsymbol{J}$ is the Jacobian $\boldsymbol{J} = \partial(\boldsymbol{y})/\partial(\boldsymbol{x})$. Let $\boldsymbol{\Gamma}\mathcal{D}$ be the set of densities $\boldsymbol{\Gamma} q$ corresponding to densities $q \in \mathcal{D}$. Let $(\boldsymbol{\Gamma}\mathcal{I}) \subseteq (\boldsymbol{\Gamma}\mathcal{D})$ correspond to $\mathcal{I} \subseteq \mathcal{D}$. Then, for any prior $p \in \mathcal{D}$ and new information $I = (q^\dagger \in \mathcal{I})$,

$$(\boldsymbol{\Gamma} p) \circ (\boldsymbol{\Gamma} I) = \boldsymbol{\Gamma}(p \circ I) \tag{10}$$

holds, where $\boldsymbol{\Gamma} I = ((\boldsymbol{\Gamma} q^\dagger) \in (\boldsymbol{\Gamma}\mathcal{I}))$.

Justification: We expect the same answer when we solve the same problem in two different coordinate systems, in that the posteriors in the two systems should be related by the coordinate transformation.

Suppose there are two systems, with sets $\boldsymbol{D}_1, \boldsymbol{D}_2$ of states and probability densities of states $q_1^\dagger \in \mathcal{D}_1$, $q_2^\dagger \in \mathcal{D}_2$. Then we require the following axiom.

Axiom III (System Independence): Let $p_1 \in \mathcal{D}_1$ and $p_2 \in \mathcal{D}_2$ be prior densities. Let $I_1 = (q_1^\dagger \in \mathcal{I}_1)$ and $I_2 = (q_2^\dagger \in \mathcal{I}_2)$ be new information about the two systems, where $\mathcal{I}_1 \subseteq \mathcal{D}_1$ and $\mathcal{I}_2 \subseteq \mathcal{D}_2$. Then

$$(p_1 p_2) \circ (I_1 \wedge I_2) = (p_1 \circ I_1)(p_2 \circ I_2) \tag{11}$$

holds.

Justification: Instead of $q_1^\dagger$ and $q_2^\dagger$, we could describe the systems using the joint density $q^\dagger \in \mathcal{D}_{12}$. If the two systems were independent, then the joint density would satisfy

$$q^\dagger(x_1, x_2) = q_1^\dagger(x_1) q_2^\dagger(x_2). \tag{12}$$

Now the new information about each system can also be expressed completely in terms of the joint density $q^\dagger$. For example, I_1 can be expressed as $I_1 = (q^\dagger \in \mathcal{I}_1')$, where $\mathcal{I}_1' \subseteq \mathcal{D}_{12}$ is the set of joint densities $q \in \mathcal{D}_{12}$ such that $q_1 \in \mathcal{I}_1$, where

$$q_1(x_1) = \int_{\boldsymbol{D}_2} dx_2 q(x_1, x_2).$$

I_2 can be expressed similarly. Now, since the two priors together define a joint prior $p = p_1 p_2$, it follows that there are two ways to take the new information I_1 and I_2 into account: we can obtain separate posteriors $q_1 = p_1 \circ I_1$ and $q_2 = p_2 \circ I_2$, or we can obtain a joint posterior $q = p \circ (I_1 \wedge I_2)$. Because p_1 and p_2 are independent, and because I_1 and I_2 give no information about any interaction between the two systems, we expect these two ways to be related by $q = q_1 q_2$, whether or not (12) holds.

Axiom IV (Subset Independence): Let $\boldsymbol{S}_1, \cdots, \boldsymbol{S}_n$ be disjoint sets whose union is $\boldsymbol{D}$, and let $p \in \mathcal{D}$ be any known prior. For each subset $\boldsymbol{S}_i$, let $I_i = (q^\dagger * \boldsymbol{S}_i \in \mathcal{I}_i)$ be new information about the conditional density $q^\dagger * \boldsymbol{S}_i$, where $\mathcal{I}_i \subseteq \mathbb{S}_i$ and $\mathbb{S}_i$ is the set of densities on $\boldsymbol{S}_i$. Let $M = (q^\dagger \in \mathcal{M})$ be new information giving the probability of being in each of the n subsets, where $\mathcal{M}$ is the set of densities q that satisfy

$$\int_{\boldsymbol{S}_i} dx q(x) = m_i \tag{13}$$

for each subset $\boldsymbol{S}_i$, where the m_i are known values. Then

$$(p \circ (I \wedge M)) * \boldsymbol{S}_i = (p * \boldsymbol{S}_i) \circ I_i \tag{14}$$

holds, where $I = I_1 \wedge I_2 \wedge \cdots \wedge I_n$.

Justification: This axiom concerns situations in which the set of states $\boldsymbol{D}$ decomposes naturally into disjoint subsets $\boldsymbol{S}_i$, and new information I_i is obtained about the conditional probability densities $q^\dagger * \boldsymbol{S}_i$ in each subset (see (8) and (9)). One way of accounting for this information is to obtain a conditional posterior $q_i = (p * \boldsymbol{S}_i) \circ I_i$ from each conditional prior $p * \boldsymbol{S}_i$. Another way is to obtain a posterior $q = p \circ I$ for the whole system, where $I = I_1 \wedge \cdots \wedge I_n$. The two results should be related by $q * \boldsymbol{S}_i = q_i$ or

$$(p \circ I) * \boldsymbol{S}_i = (p * \boldsymbol{S}_i) \circ I_i. \tag{15}$$

Moreover, suppose that we also learn the probability of being in each of the n subsets. That is, we learn $M = (q^\dagger \in \mathcal{M})$, where $\mathcal{M}$ is the set of densities q that satisfy (13) for each subset $\boldsymbol{S}_i$. The known numbers m_i are the probabilities that the system is in a state within $\boldsymbol{S}_i$. The m_i satisfy $\sum_i m_i = 1$. Taking M into account should not affect the conditional densities that result from taking I into account. We therefore expect a more general version of (15) to hold, namely (14).

IV. Consequences of the Axioms

A. Summary

Since we require the axioms to hold for both equality and inequality constraints (2) and (3), they must hold for equality constraints alone. We first investigate the axioms' consequences assuming only equality constraints. Later, we show that the resulting restricted form for H also satisfies the axioms in the case of inequality constraints.

We establish our main result in four steps. The first step shows that the subset independence axiom and a special case of the invariance axiom together restrict H to functionals that are equivalent to the form

$$F(q,p) = \int_{\boldsymbol{D}} dx f(q(x), p(x)) \tag{16}$$

for some function f. We call this the "sum form." In the axiomatic characterizations in [34], [55], and [56], the sum form was assumed rather than derived. Our next step shows that the general case of the invariance axiom restricts H to functionals that are equivalent to the form

$$F(q,p) = \int_{\boldsymbol{D}} dx q(x) h(q(x)/p(x)) \tag{17}$$

for some function h. Our third step applies the system independence axiom and shows that if H is a functional that satisfies all four axioms, then H is equivalent to cross-entropy (1). Since it could still be imagined that no functional satisfies the axioms, our final step is to show that cross-entropy does. We do this in the general case of equality and inequality constraints.

B. Deriving the Sum Form

We derive the sum form in several steps. First, we show that when the assumptions of the subset independence axiom hold, the posterior values within any subspace are independent of the values in the other subspaces. Next, we move formally to the discrete case and show that invariance implies that H is equivalent to a symmetric function. We then apply the subset independence axiom and prove that H is equivalent to functions of the form $F(\boldsymbol{q},\boldsymbol{p})=\Sigma_j f(q_j,p_j)$, where $\boldsymbol{p}$ and $\boldsymbol{q}$ are discrete prior and posterior distributions, respectively, and we return to the continuous case yielding (16).

We begin with the following lemma concerning subset independence.

Lemma I: Let the assumptions of Axiom IV hold, and let $q=p\circ(I\wedge M)$ be the posterior for the whole system $(q\in\mathfrak{D})$. Then $q(\mathbf{x}\in\boldsymbol{S}_i)$ is functionally independent of $q(\boldsymbol{x}\notin\boldsymbol{S}_i)$, of the prior $p(\boldsymbol{x}\notin\boldsymbol{S}_i)$, and of n.

Proof: Let

$$q_i=(p*\boldsymbol{S}_i)\circ I_i \tag{18}$$

be the conditional posterior density in the ith subspace $(q_i\in\mathbb{S}_i)$. Since $p*\boldsymbol{S}_i$ depends on p only in terms of $p(\boldsymbol{x}\in\boldsymbol{S}_i)$ (see (8) and (9)), so does q_i. Furthermore, since q_i is the solution (18) to a problem in which $\boldsymbol{x}\in\boldsymbol{S}_i$ only, q_i cannot depend on $q(\boldsymbol{x}\notin\boldsymbol{S}_i)$. Now, (14) states that $q(\boldsymbol{x})=m_iq_i(\boldsymbol{x})$ for $\boldsymbol{x}\in\boldsymbol{S}_i$, where we have used (8) and (13). Since the m_i are fixed, it follows that $q(\boldsymbol{x}\in\boldsymbol{S}_i)$ is independent of $q(\boldsymbol{x}\notin\boldsymbol{S}_i)$ and $p(\boldsymbol{x}\notin\boldsymbol{S}_i)$, proving Lemma I.

Our next step is to transform to the discrete case.

Lemma II: Let $\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_n$ be disjoint sets whose union is $\boldsymbol{D}$. For a prior p and a posterior $q=p\circ I$ let

$$p_j=\int_{\boldsymbol{S}_j}dxp(x),\quad\text{and}\quad q_j=\int_{\boldsymbol{S}_j}dxq(x).$$

Suppose that $p(\boldsymbol{x}\in\boldsymbol{S}_j)$ is constant for each subset $\boldsymbol{S}_j$, and let the new information I be provided by constraints (3) and (4) in which the functions a_k and c_k are also constant in each subset. Then the posterior $q=p\circ I$ is also constant in each subset, and H is equivalent to a symmetric function of the n pairs of variables (q_j,p_j) (We refer to this situation as the *discrete case.*)

Proof: Since the a_k and c_k are constant in each subset, the constraints have the form

$$\sum_j q_j^\dagger a_{kj}=0 \tag{19}$$

or

$$\sum_j q_j^\dagger c_{kj}\geqslant 0, \tag{20}$$

where $a_{kj}=a_k(\boldsymbol{x}\in\boldsymbol{S}_j)$, $c_{kj}=c_k(\boldsymbol{x}\in\boldsymbol{S}_j)$, and

$$q_j^\dagger=\int_{\boldsymbol{S}_j}dxq^\dagger(x).$$

Now, let $\boldsymbol{\Gamma}$ be a measure-preserving transformation that scrambles the $\boldsymbol{x}$ within each subset $\boldsymbol{S}_i$. This leaves the prior and the constraints (19) and (20) unchanged. It follows from invariance (10) that $\boldsymbol{\Gamma}$ also leaves q unchanged, which will only be the case if q is constant in each $\boldsymbol{S}_i$. In the discrete case, H becomes a function $H(\boldsymbol{q},\boldsymbol{p})$ of $2n$ variables $q_1,\cdots,q_n$ and $p_1,\cdots,p_n$. To show that H is equivalent to a symmetric function let π be any permutation. By invariance, the minima of H and πH coincide, where

$$\pi H(\boldsymbol{q},\boldsymbol{p})=H(q_{\pi(1)},\cdots,q_{\pi(n)},p_{\pi(1)},\cdots,p_{\pi(n)}).$$

Therefore the minima of H and F coincide, where F is the mean of the πH for all permutations π, and H is equivalent to the symmetric function F. This completes the proof of Lemma II.

We now prove that H is equivalent to functions with the discrete sum form.

Theorem I: In the discrete case let $H(\boldsymbol{q},\boldsymbol{p})$ satisfy uniqueness, invariance, and subset independence. Then H is equivalent to a function of the form

$$F(\boldsymbol{q},\boldsymbol{p})=\sum_j f(q_j,p_j) \tag{21}$$

for some function f.

Theorem I is proved in the Appendix. The proof rests primarily on the subset independence property (Lemma I).

We return to the continuous case by taking the limit of a large number of small subspaces $\boldsymbol{S}_i$. The discrete sum form (21) then becomes (16).

C. Consequence of General Invariance in the Continuous Case

Although invariance was invoked for the special case of discrete permutations in deriving (21), the continuous sum form (16) does not satisfy the invariance axiom for arbitrary continuous transformations and arbitrary functions f. The invariance axiom restricts the possible forms of f as follows.

Theorem II: Let the functional $H(q,p)$ satisfy uniqueness, invariance, and subset independence. Then H is equivalent to a functional of the form

$$F(q,p)=\int_{\boldsymbol{D}}dxq(x)h(q(x)/p(x)) \tag{22}$$

for some function h.

Proof: From previous results we may assume H to have the form (16). Consider new information I consisting of a single equality constraint

$$\int_{\boldsymbol{D}}dxq^\dagger(x)a(x)=0. \tag{23}$$

Then, by standard techniques from the calculus of variations, it follows that the posterior $q=p\circ I$ satisfies

$$\lambda+\alpha a(x)+g(q(x),p(x))=0, \tag{24}$$

where λ and α are Lagrangian multipliers corresponding to the constraints (2) and (23) and where the function g is defined as

$$g(b,c)=\frac{\partial}{\partial b}f(b,c). \tag{25}$$

Now let $\boldsymbol{\Gamma}$ be a coordinate transformation from $\boldsymbol{x}$ to $\boldsymbol{y}$ in the notation of Axiom II. Then the transformed prior is $p'(\boldsymbol{y})=\boldsymbol{J}^{-1}p(\boldsymbol{x})$ and the transformed constraint function is $a'(\boldsymbol{y})=\boldsymbol{\Gamma}a=a(\boldsymbol{x})$. The posterior $q'=p'\circ(\boldsymbol{\Gamma}I)$ satisfies

$$\lambda'+\alpha'a'(\boldsymbol{y})+g(q'(\boldsymbol{y}),p'(\boldsymbol{y}))=0, \tag{26}$$

where λ' and α' are Lagrangian multipliers. Invariance (10) requires that $q'(\boldsymbol{y})=\boldsymbol{J}^{-1}q(\boldsymbol{x})$ holds, so (26) becomes

$$\lambda'+\alpha'a(\boldsymbol{x})+g(\boldsymbol{J}^{-1}q(\boldsymbol{x}),\boldsymbol{J}^{-1}p(\boldsymbol{x}))=0. \tag{27}$$

Combining (24) and (27) yields

$$g(\boldsymbol{J}^{-1}q(\boldsymbol{x}),\boldsymbol{J}^{-1}p(\boldsymbol{x}))=g(q(\boldsymbol{x}),p(\boldsymbol{x}))+(\alpha-\alpha')a(\boldsymbol{x})+\lambda-\lambda'. \tag{28}$$

Now let $\boldsymbol{S}_1,\cdots,\boldsymbol{S}_n$ be disjoint subsets whose union is $\boldsymbol{D}$ and let the prior p be constant within each $\boldsymbol{S}_j$. It follows from Lemma II that q is also constant within each $\boldsymbol{S}_j$, which in turn results in the right side of (28) being constant within each $\boldsymbol{S}_j$. (The primed Lagrangian multipliers may depend on the transformation $\boldsymbol{\Gamma}$, but they are constants.) On the left side, however, the Jacobian $\boldsymbol{J}(\boldsymbol{x})$ may take on arbitrary values since $\boldsymbol{\Gamma}$ is an arbitrary transformation. It follows that g can only depend on the ratio of its arguments, i.e., $g(b,c)=g(b/c)$. Equation (25), therefore, has the general solution $f(a,b)=ah(a/b)+v(b)$, for some functions h and v. Substitution of this solution into (16) yields

$$F(q,p)=\int_{\boldsymbol{D}}dxq(x)h(q(x)/p(x))+\int_{\boldsymbol{D}}dxv(p(x)).$$

Since the second term is a function only of the fixed prior, it cannot affect the minimization of F and may be dropped. This completes the proof of Theorem II.

D. Consequence of System Independence

Our results so far have not depended on Axiom III. We now show that system independence restricts the function h in (22) to a single equivalent form.

Theorem III: Let the functional $H(q,p)$ satisfy uniqueness, invariance, subset independence, and system independence. Then H is equivalent to cross-entropy (1).

Proof: With $i=1,2$, consider two systems with states $\boldsymbol{x}_i\in\boldsymbol{D}_i$, unknown densities $q_i^\dagger\in\mathcal{D}_i$, prior densities $p_i\in\mathcal{D}_i$, and new information I_i in the form of single equality constraints

$$\int_{\boldsymbol{D}_i}dx_iq_i^\dagger(x_i)a_i(x_i)=0. \tag{29}$$

From Theorem II, we may assume that H has the form (22). It follows that the posteriors $q_i=p_i\circ I_i$ satisfy

$$\lambda_i+\alpha_ia_i(x_i)+u(r_i(x_i))=0, \tag{30}$$

where λ_i and α_i are Lagrangian multipliers corresponding to the constraints (2) and (29), where $r_i(x_i)=q_i(x_i)/p_i(x_i)$, and where

$$u(r)=h(r)+r\frac{d}{dr}h(r). \tag{31}$$

The two systems can also be described in terms of a joint probability density $q^\dagger\in\mathcal{D}_{12}$, a joint prior $p=p_1p_2$, and new information I in the form of the three constraints

$$\int_{\boldsymbol{D}_1}\int_{\boldsymbol{D}_2}dx_1dx_2q^\dagger(x_1,x_2)=1, \tag{32}$$

$$\int_{\boldsymbol{D}_1}\int_{\boldsymbol{D}_2}dx_1dx_2q^\dagger(x_1,x_2)a_i(x_i)=0 \qquad (i=1,2). \tag{33}$$

The posterior $q=p\circ I$ satisfies

$$\lambda'+\alpha'_1a_1(x_1)+\alpha_2'a_2(x_2)+u(r(x_1,x_2))=0, \tag{34}$$

where the multipliers λ', α_1', and α_2' correspond to (32) and (33), and $r=q/p$.

Now, system independence (11) requires $q=q_1q_2$, from which follows $r=r_1r_2$. Combining (30) and (34) therefore yields

$$u(r_1r_2)-u(r_1)-u(r_2)=(\alpha_1-\alpha_1')a_1+(\alpha_2-\alpha_2')a_2+\lambda_1+\lambda_2-\lambda'. \tag{35}$$

Consider the case when $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are both the real line. Then, differentiating this equation with respect to x_1 and differentiating the result with respect to x_2 yields

$$u''(r_1r_2)r_1r_2+u'(r_1r_2)=0. \tag{36}$$

By suitable choices for the priors and the constraints, r_1r_2 can be made to take on any arbitrary positive value s. It follows from (36) that the function u satisfies the differential equation $u'(s)+su''(s)=0$, which has the general solution $u(s)=A\log(s)+B$, for arbitrary constants A and B. Combining this solution with (31) yields

$$h(r)+r\frac{d}{dr}h(r)=A\log(r)+B,$$

which in turn has the general solution

$$h(r)=A\log(r)+C/r+B-A. \tag{37}$$

Substitution of (37) into (22) yields

$$F(q,p)=A\int_{\boldsymbol{D}}dxq(x)\log(q(x)/p(x))+(C+B-A), \tag{38}$$

since p integrates to one. Since the constants A, B, and C cannot affect the minimization of (38), provided $A>0$, this completes the proof of Theorem III.

E. Cross-Entropy Satisfies the Axioms

So far we have shown that if $H(q,p)$ satisfies the axioms, then H is equivalent to cross-entropy (1). This still leaves open the possibility that no functional H satisfies the axioms for arbitrary constraints. By showing that cross-entropy satisfies the axioms for arbitrary constraints, we complete the proof of our main result.

Theorem IV: Cross-entropy (1) satisfies uniqueness, invariance, system independence, and subset independence.

Every other functional that satisfies the axioms is equivalent to cross-entropy.

Proof: We need only show that cross-entropy satisfies the axioms.

Uniqueness: Let $\mathcal{I}$ be any closed convex set $\mathcal{I} \subseteq \mathcal{D}$, and let densities $q,r \in \mathcal{I}$ have the same cross entropy $H(q,p)=H(r,p)$ for some prior $p \in \mathcal{D}$. We define $g(u)=u\log(u)$, with $g(0)=0$, so that H can be written as

$$H(q,p)=\int_D dx\,p(x)g(q(x)/p(x)).$$

Now since $g''(u)=1/u>0$, g is strictly convex. It follows that

$$\alpha g(u)+(1-\alpha)g(v)>g(\alpha u+(1-\alpha)v),$$

for $0<\alpha<1$ and $u \neq v$. We set $q(x)/p(x)$ for u and $r(x)/p(x)$ for v, multiply both sides by $p(x)$, and integrate, obtaining

$$\begin{aligned} H(q,p)&=H(r,p)\\ &=\alpha H(q,p)+(1-\alpha)H(r,p)\\ &\geqslant H(\alpha q+(1-\alpha)r,p). \end{aligned}$$

The inequality is strict unless $q=r$. (We write $q=r$ when $q(x)=r(x)$ for almost all x, since in this case q and r define the same probability distribution.) Thus, if $q \neq r$ and $H(q,p)=H(r,p)$ both hold, there is a density $\alpha q+(1-\alpha)r$ that belongs to $\mathcal{I}$ (since $\mathcal{I}$ is convex) and has cross-entropy smaller than $H(q,p)$. Therefore, there cannot be two distinct densities $q,r \in \mathcal{I}$ having the minimum cross-entropy in $\mathcal{I}$. For the existence of one such density see Csiszár [66, theorem 2.1]. This proves that cross-entropy satisfies Axiom I.

Invariance: Let Γ be a coordinate transformation from x to y in the notation of Axiom II. A change of variables in (1) shows that cross-entropy is transformation invariant:

$$H(q,p)=H(\Gamma q,\Gamma p).$$

The minimum in $\Gamma\mathcal{I}$ therefore corresponds to the minimum in $\mathcal{I}$, which proves that cross-entropy satisfies Axiom II.

System Independence: We use the notation in Axiom III. Consider densities $q_1,p_1 \in \mathcal{D}_1$ and $q_2,p_2 \in \mathcal{D}_2$. Let $q \in \mathcal{D}_{12}$ satisfy $q \neq q_1q_2$,

$$\int_{D_1} dx_1 q(x_1,x_2)=q_2, \quad \text{and} \quad \int_{D_2} dx_2 q(x_1,x_2)=q_1;$$

i.e., q and q_1q_2 are different densities with the same marginal densities. A straightforward computation of the cross-entropy difference between q and q_1q_2 for the same prior p_1p_2 yields

$$H(q,p_1p_2)-H(q_1q_2,p_1p_2)=H(q,q_1q_2).$$

Now, cross-entropy has the property that $H(q,p)\geqslant 0$ with $H(q,p)=0$ only if $q=p$ (for example, see [31, p. 14]). It follows that

$$H(q,p_1p_2)>H(q_1q_2,p_1p_2) \tag{39}$$

holds, since $q \neq q_1q_2$ by assumption. This means that of all the densities $q \in \mathcal{D}_{12}$ with given marginal densities q_1 and q_2, the one with the least cross-entropy is q_1q_2. Since I_1 and I_2 restrict only the marginal densities of q in $q=(p_1p_2)\circ(I_1 \wedge I_2)$—see Axiom III and its justification—the density q with the least cross-entropy in the constraint set is of product form q_1q_2. But the cross-entropy of a density of this form satisfies

$$H(q_1q_2,p_1p_2)=H(q_1,p_1)+H(q_2,p_2) \tag{40}$$

and so assumes its minimum when the two terms on the right assume their individual minima—the first subject to I_1 and the second to I_2. Thus we have $q=(p_1p_2)\circ(I_1\wedge I_2)$ $=q_1q_2=(p_1\circ I_1)(p_2\circ I_2)$, and we have proved that cross-entropy satisfies Axiom III.

Subset Independence: We use the notation in Axiom IV. We also define $q=p\circ(I\wedge M)$, $q_i=q*S_i$, and $p_i=p*S_i$. (Equation (14) then becomes $q_i=p_i\circ I_i$.) The cross-entropy of q with respect to p may be written

$$\begin{aligned} H(q,p)&=\sum_i \int_{S_i} dx\, m_i q_i(x)\log\left(\frac{m_i q_i(x)}{s_i p_i(x)}\right)\\ &=\sum_i m_i H(q_i,p_i)+\sum_i m_i \log\left(\frac{m_i}{s_i}\right), \end{aligned} \tag{41}$$

where the s_i are the prior probabilities of being in each subset,

$$s_i=\int_{S_i} dx\,p(x).$$

The second sum on the right of (41) is a constant and has no effect on minimization. Minimizing the left side of (41) subject to $(I\wedge M)$ is equivalent to minimizing each term of $\sum_i m_i H(q_i,p_i)$ individually subject to I_i. This proves that cross-entropy satisfies subset independence and completes the proof of Theorem IV.

V. The Discrete Case

A. Principle of Minimum Cross-Entropy for Discrete Systems

Theorem IV states that if one wishes to select a posterior $q=p\circ I$ in a manner that satisfies Axioms I—IV, the unique result can be obtained by minimizing the cross-entropy (1). Although the equivalent result for the discrete case can be obtained in the usual informal way by replacing integrals with sums and densities with distributions, it can also be obtained formally as follows.

Suppose a system has a finite set of n states with probabilities $\boldsymbol{q}^\dagger$. Let $\boldsymbol{p}$ be a prior estimate of $\boldsymbol{q}^\dagger$ and let new information I be provided in the form

$$\sum_i q_i^\dagger a_{ki}=0 \tag{42}$$

or

$$\sum_i q_i^\dagger c_{ki}\geqslant 0, \tag{43}$$

for known numbers a_{ki} and c_{ki}. Then it is clear that there exist problems with continuous states and densities for which the foregoing finite problem is the discrete case as defined in Lemma II. It follows from Lemma II and Theorem IV that the cross-entropy functional becomes a function of $2n$ variables and that the posterior $\boldsymbol{q}=\boldsymbol{p}\circ I$ can be obtained by minimizing the function $H(\boldsymbol{q},\boldsymbol{p})=\Sigma_i q_i \log(q_i/p_i)$, subject to the constraints (42) and (43).

B. The Maximum Entropy Principle

Using transformation group arguments, Jaynes [25] has shown that a uniform prior $p_i = n^{-1}$ is appropriate when we know only that each of the n system states is possible (as distinct from "complete ignorance" when we do not even know this much). It follows that, given a finite state space and constraints of the form (42) and (43), the posterior is obtained by minimizing the function

$$H(\boldsymbol{q})=\sum_i q_i \log(q_i)-\log(n).$$

This is equivalent to maximizing the entropy $-\Sigma_i q_i \log(q_i)$. Thus, entropy maximization is a special case of cross-entropy minimization.

It is also possible to obtain the maximum entropy principle formally and directly. We show how in the following although we omit some of the formal details. The first step is to rewrite the axioms so that they refer to the discrete case in which no prior is available. In this case, given new information I in the form of constraints (42) and (43), the unary operator $\circ$ selects a posterior distribution $\boldsymbol{q}=(\circ I)$ from all distributions that satisfy the constraints. The operator is realized by minimizing some function $H(\boldsymbol{q})$. The axioms become (see Section III) the following.

I. *Uniqueness*: The posterior $q=(\circ I)$ is unique.

II. *Permutation Invariance*: $\circ(\Gamma I)=\Gamma(\circ I)$ for any permutation Γ.

III. *System Independence*: $(\circ(I_1 \wedge I_2))=(\circ I_1)(\circ I_2)$.

IV. *Subset Independence*: $(\circ(I\wedge M)) * S_i=(\circ I_i)$. (44)

Theorem I goes through in a straightforward way with the prior deleted. This shows that, if $H(\boldsymbol{q})$ satisfies uniqueness, permutation invariance, and subset independence, it is equivalent to a function of the form

$$H(\boldsymbol{q})=\sum_i f(q_i). \tag{45}$$

Next we assume this form and apply system independence in a manner analogous to the proof of Theorem III. Consider a system with n states and an unknown distribution $\boldsymbol{q}^{\dagger}$, and another system with m states and an unknown distribution $\boldsymbol{r}^{\dagger}$. New information is provided in terms of single constraints:

$$\sum_{i=1}^{n} q_i^{\dagger} a_i=\sum_{k=1}^{m} r_k^{\dagger} b_k=0.$$

The posteriors $\boldsymbol{q}$ and $\boldsymbol{r}$ satisfy

$$u(q_i r_k)=u(q_i)+u(r_k)+(\alpha-\alpha')a_i+(\beta-\beta')b_k+\lambda_1+\lambda_2-\lambda',$$

where $u(x)=f'(x)$ and α, α', β, β', λ_1, λ_2, and λ' are Lagrangian multipliers. This is the discrete analog of (35). It leads to

$$u(q_i r_k)-u(q_i r_v)=u(q_u r_k)-u(q_u r_v)=G(r_k,r_v) \tag{46}$$

for some function G. Since the right side of (46) does not depend on q_i, we pick an arbitrary value for q_i on the left side. This shows that G satisfies

$$G(x,y)=s(x)-s(y) \tag{47}$$

for some function s. (We note that G satisfies Sincov's functional equation $G(x,y)=G(x,z)+G(z,y)$ which has the general solution (47) [64, p. 223].) Some manipulation of (46) and (47) yields

$$u(xy)-s(x)-s(y)=u(wz)-s(w)-s(z).$$

Since the two sides are independent of each other, they must be equal to some constant. Thus, u satisfies $u(xy)=g(x)+g(y)$, for some function g. Using standard techniques of functional equations [64, pp. 34, 302], we obtain the general solution for u, namely $u(x)=A\log(x)+B$, where A and B are constants. Combining this with $u(x)=f'(x)$ and integrating yields the solution for f in (45), $f(x)=Ax\log(x)+Bx-A$, which in turn yields

$$H(\boldsymbol{q})=A\sum_i q_i \log(q_i)-nA+B. \tag{48}$$

This function has a unique minimum provided that A is positive.

Minimizing the function H in (48) is equivalent to maximizing the entropy $-\Sigma_i q_i \log(q_i)$. This proves that if one wishes to select a discrete posterior distribution $\boldsymbol{q}=(\circ I)$ in a manner that satisfies the axioms (44), the unique result can be obtained by maximizing entropy.

VI. Concluding Remarks

Our approach has been to axiomatize desired properties of inference methods rather than to axiomatize desired properties of information measures. Yet it might seem that the axioms in Section III are no more than a thinly disguised characterization of cross-entropy. In this view Axioms I and II might correspond to axioms requiring that H have unique minima and be transformation invariant, and Axioms III and IV might correspond to axioms requiring that H be "additive" [34] and satisfy something like the "branching property" [65]. These correspondences are meaningful and not surprising—after all, inference methods should relate to information measures—but it is important to realize that there are significant differences as well. For example, if we knew that H itself must be transformation invariant, the deduction of (22) from (16) would be direct (Theorem II). But Axiom II implies only that the minima of H must be transformation invariant, so the proof of Theorem II reasons in terms of invariance at the minima.

As another example, consider the following axiom.

Additivity:

$$H(q_1q_2,p_1p_2)=H(q_1,p_1)+H(q_2,p_2)$$
$$\text{for all } q_1,p_1\in\mathcal{D}_1 \text{ and } q_2,p_2\in\mathcal{D}_2. \quad (49)$$

This can be used [34] in characterizing the directed divergences. In Section IV we showed that if H has the sum form (22) and satisfies system independence, then H is equivalent to cross-entropy (Theorem III). When we proved, as part of Theorem IV, that cross-entropy itself satisfies system independence, we used the fact that cross-entropy satisfies additivity (49) (see (41)). It might seem that any functional that satisfies additivity also satisfies system independence. But Johnson [34] proved that the information measures $H(q,p)$ of the form (22) that satisfy additivity (49) are those of the form

$$H(q,p)=A\int_D dxq(x)\log(q(x)/p(x)) + B\int_D dxp(x)\log(p(x)/q(x)), \quad (50)$$

for some constants $A,B\geqslant 0$, not both zero. That is, (22) and additivity (49) of H yields the linear combination of both directed divergences, whereas (24) and system independence of $\circ$ yields only one of the directed divergences, cross-entropy. The key to the difference is the property expressed by (39)—for all densities $q\in\mathcal{D}_{12}$ with given marginal densities q_1 and $q_2,H(q,p_1p_2)$ has its minimum at $q=q_1q_2$. This property is necessary if H is to satisfy system independence; it is satisfied by the first term in (50) but not by the second, even though the second term satisfies additivity.

In summary, we have proved that, in a well-defined sense, Jaynes's principle of maximum entropy and Kullback's principle of minimum cross-entropy (minimum directed divergence) provide correct general methods of inductive inference when given new information in the form of expected values. When Jaynes first advocated the maximum entropy principle more than 20 years ago, he did not ignore such questions as "why maximize entropy, why not some other function?" We have established the sense in which the following conjecture [1, p. 623] is correct: "deductions made from any other information measure, if carried far enough, will eventually lead to contradictions."

Acknowledgment

The authors would like to thank A. Ephremides, W. S. Ament, and J. Aczél for their reviews of an earlier version of this paper.

Appendix
Proof of Theorem I

After showing that $\partial H/\partial q_i$ has the form

$$\frac{\partial H}{\partial q_i}=z(\boldsymbol{q},\boldsymbol{p})+g(q_i,p_i)s(\boldsymbol{q},\boldsymbol{p}), \quad (\text{A1})$$

we show that (A1) results in H being functionally dependent on $F(\boldsymbol{q},\boldsymbol{p})=\sum_i f(q_i,p_i)$, where f satisfies $g=\partial f(b,c)/\partial b$. We then show that the functional dependence is monotonic so that H and F are equivalent.

In realizing the operator $\circ$, the only relevant values of $H(\boldsymbol{q},\boldsymbol{p})$ are at points $\boldsymbol{q}$ that satisfy the discrete form of (1):

$$\sum_{j=1}^{n} q_j=1. \quad (\text{A2})$$

We refer to the hyperplane of such points $\boldsymbol{q}$ as the *normalization subspace*. In selecting posteriors by minimizing H, we are further restricted to the *positive region* in which $q_i\geqslant 0$ for $i=1,\cdots,n$. On the normalization subspace (A2), $H(\boldsymbol{q},\boldsymbol{p})$ is a function of only $n-1$ independent variables q_i (the prior $\boldsymbol{p}$ is assumed fixed). For convenience, however, we consider H to be extended off the normalization subspace to a well-behaved function of n independent variables that is symmetric under identical permutations of $\boldsymbol{q}$ and $\boldsymbol{p}$ (see Lemma II). This enables us to express the gradient ∇H as

$$\nabla H=\sum_{i=1}^{n}\frac{\partial H}{\partial q_i}\hat{e}_i,$$

where $\{\hat{e}_1,\cdots,\hat{e}_n\}$ is a standard orthonormal basis. The operator $\circ$ can be realized by minimizing the extended H in the positive region provided that (A2) is always imposed as a constraint. In the continuous case we have assumed that the functional $H(q,p)$ is well-behaved. We take this to mean, in particular, that the function $H(\boldsymbol{q},\boldsymbol{p})$ is continuously differentiable in the interior of the positive region of the normalization subspace and that the projection of $\boldsymbol{\nabla} H$ into the normalization subspace is zero only at minima of H.

Now let N be the set $\{1,\cdots,n\}$, let $M\subset N$ be a set of m integers from N, and let $M-N$ be the set that remains after deleting M. Let $\boldsymbol{q}_M$ comprise the components q_i with $i\in M$ and let $\boldsymbol{q}_{N-M}$ comprise the rest. We refer to points $\boldsymbol{q}_M$ as points in the M-subspace. We assume both $n\geqslant 6$ and $m\geqslant 4$. Suppose new information comprises a set of constraints (19) that satisfy $a_{kj}=0$ either for all $j\in M$ or for all $j\in N-M$, including the constraint

$$\sum_{j\in M} q_j^\dagger=r. \quad (\text{A3})$$

Any constraint satisfying $a_{kj}=0$ for $j\in M$ can be written as a constraint

$$\sum_{j\in M} a_{kj}q_j^\dagger=\sum_{j\in M} a_{kj}(q_j^\dagger/r)=0$$

on the conditional distribution given $j\in M$: $(\boldsymbol{q}_M/r)$. Similarly, constraints that satisfy $a_{kj}=0$ for $j\in N-M$ can be written as constraints on the conditional distribution $\boldsymbol{q}_{N-M}/(1-r)$. Therefore, the system decomposes into two subsets (M and $N-M$) with new information that satisfies the assumptions of Axiom IV (subset independence). It follows from Lemma I that, when $H(\boldsymbol{q},\boldsymbol{p})$ is minimized over the constraint set, the resulting $\boldsymbol{q}_M$ are independent of the $\boldsymbol{q}_{N-M}$, of the $\boldsymbol{p}_{N-M}$, and of n.

Now, the constraint (A3) requires that the solution $\boldsymbol{q}_M$ be found on the $m-1$ dimensional hyperplane defined by (A3). Therefore, finding this solution depends not on the projection of ∇H into the M-subspace,

$$(\nabla H)_M=\sum_{j\in M}\frac{\partial H}{\partial q_i}\hat{e}_j,$$

but on its projection onto the $(m-1)$ dimensional hyperplane defined by (A3). This projection is given by $\boldsymbol{B}_M=(\boldsymbol{\nabla} H)_M-(\hat{\boldsymbol{n}}\cdot(\boldsymbol{\nabla} H)_M)\hat{\boldsymbol{n}}$, where $\hat{\boldsymbol{n}}$ is a unit vector normal to the hyperplane. $\boldsymbol{B}_M$

has components

$$B_{Mi} = \frac{\partial H}{\partial q_i} - \frac{1}{m} \sum_{j \in M} \frac{\partial H}{\partial q_j} \tag{A4}$$

for $i \in M$. Now, since H is symmetric (Lemma II),

$$\frac{\partial H}{\partial q_i} = h(q_i, \boldsymbol{q}_{N-i}, p_i, \boldsymbol{p}_{N-i}) \equiv h_i$$

holds for some function h, where $\boldsymbol{q}_{N-i}$ is any permutation of $\boldsymbol{q}$ with q_i deleted and $\boldsymbol{p}_{N-i}$ is the same permutation of $\boldsymbol{p}$ with p_i deleted. Hence, (A4) becomes

$$B_{Mi} = B(q_i, \boldsymbol{q}_{N-i}, p_i, \boldsymbol{p}_{N-i}),$$

for some function B.

To find the solution for $\boldsymbol{q}_M$, one moves on the constraint hyperplane opposite the direction of maximum change in H—i.e., opposite the direction of $\boldsymbol{B}_M$—until no further movement is possible within the constraint set (19). Since the solution cannot depend on $\boldsymbol{q}_{N-M}$ or $\boldsymbol{p}_{N-M}$, neither can the direction of $\boldsymbol{B}_M$. This direction is also independent of n, since the subspace solution $\boldsymbol{q}_M$ is independent of n (Lemma I). If $\boldsymbol{U}_M$ is a unit vector in the direction of $\boldsymbol{B}_M$, with components U_{Mi}, it follows that

$$U_{Mi} \equiv \frac{B_{Mi}}{|\boldsymbol{B}_M|} = U(q_i, \boldsymbol{q}_{M-i}, p_i, \boldsymbol{p}_{M-i}), \tag{A5}$$

for some function U, where $\boldsymbol{q}_{M-i}$ is any permutation of $\boldsymbol{q}_M$ with q_i deleted, etc. The function U is well-defined everywhere on the constraint hyperplane except at a point at which H is minimized subject only to (A3). Such a point is characterized equivalently by $\boldsymbol{B}_M = 0$ and by $h_i = h_j$ for all $i,j \in M$. By uniqueness, there is at most one such point. For if there were more, H would reach its minimum value at more than one point or would have local minima in addition to an absolute minimum. In either case, one could define convex constraint sets in which the minimum of H would occur at more than one point, thereby violating uniqueness.

The point at which (A5) is ill-defined is also characterized by the equality of the ratios $(q_i/p_i) = (q_j/p_j)$ for all $i,j \in M$. To see this, we apply the subset independence axiom. Minimizing H subject only to (A3) means that (14) applies without the additional information I. Then, given

$$b = \sum_{j \in M} p_j,$$

(14) becomes $(q_j/r) = (p_j/b)$ so that q_j/p_j is a constant independent of j for $j \in M$. In the case of $n = m$, the constraint hyperplane becomes the entire positive region of the normalization subspace; (A3) becomes equivalent to (A2) and $r = b = 1$ holds. This shows that there is only one point at which all of the h_i are equal, namely the point $\boldsymbol{q} = \boldsymbol{p}$. Similarly, by taking $m = 2$ and $M = \{i,j\}$, one can show that the condition $h_i = h_j$ is equivalent to the condition $(q_i/p_i) = (q_j/p_j)$.

From (A5) we obtain

$$\frac{B_{Mi} - B_{Mj}}{B_{Mk} - B_{Mj}} = \frac{U_{Mi} - U_{Mj}}{U_{Mk} - U_{Mj}} \tag{A6}$$

for $i,j,k \in M$. But

$$\frac{B_{Mi} - B_{Mj}}{B_{Mk} - B_{Mj}} = \frac{h_i - h_j}{h_k - h_j} \tag{A7}$$

follows from (A4). Since the right-hand side of (A7) cannot depend on the definition of M, neither can the right-hand side of (A6). It follows that

$$\frac{h_i - h_j}{h_k - h_j} = W(q_i, q_j, q_k, p_i, p_j, p_k) \equiv W_{ijk} \tag{A8}$$

holds for some function W. By this construction W is well-defined when $q_i + q_j + q_k < 1$ and $h_k \neq h_j$; however,

$$\frac{h_i - h_j}{h_k - h_j} = \frac{W_{iju}}{W_{kju}} = W_{ijk}$$

holds, and further manipulation yields

$$\frac{W_{iru} - W_{jru}}{W_{kru} - W_{jru}} = W_{ijk}. \tag{A9}$$

Since (A9) is independent of q_r, q_u, p_r, and p_u, we may take arbitrary values of these variables and use (A9) to extend the definition of W. By the discussion following (A8), the numerator and denominator on the left of (A9) are defined as long as $(q_r/p_r) \neq (q_u/p_u)$ holds and then the fraction is well-defined whenever $(q_k/p_k) \neq (q_j/p_j)$ and $0 < q_\varphi < 1 - q_u - q_r$ hold, where $\varphi = i,j,k$. But we can make $1 - q_u - q_r$ arbitrarily close to 1 so that we may extend the domain of W_{ijk} to include all arguments such that q_i, q_j, and q_k are between 0 and 1 and $(q_k/p_k) \neq (q_j/p_j)$ holds. Moreover, (A9) continues to hold on this extended domain.

Now we may write $g(q_i,p_i) \equiv g_i$ for W_{iru} with some particular fixed values of q_r, q_u and obtain

$$\frac{h_i - h_j}{h_k - h_j} = \frac{g(q_i,p_i) - g(q_j,p_j)}{g(q_k,p_k) - g(q_j,p_j)} \tag{A10}$$

for some function g. It follows that $h_i = \partial H/\partial q_i$ has the form (A1) for some functions z, s, and g. This implies that s is given by

$$s = \frac{h_i - h_j}{g_i - g_j}. \tag{A11}$$

For a particular point $\boldsymbol{q}$, the right-hand side may be ill-defined for certain values of i and j. Since s is independent of i and j, however, s is well-defined unless $g_i = g_j$ for all i,j. But from the construction of g, the condition $g_i = g_j$ is equivalent to $h_i = h_j$ and therefore to $(q_i/p_i) = (q_j/p_j)$. It follows that s is well-defined everywhere in the positive region of the normalization subspace except perhaps at the single point where $\boldsymbol{q} = \boldsymbol{p}$ holds. The function z is likewise well-defined except perhaps at this point.

Furthermore, s and g are continuous except perhaps at $\boldsymbol{q} = \boldsymbol{p}$. Since H is continuously differentiable, the derivatives h_i are continuous and finite everywhere in the positive region of the normalization subspace (except possibly on the boundary, at points that satisfy $q_i = 0$ for some i). It follows that each of the functions $(\nabla H)_M$, $\boldsymbol{B}_M$, U, W, s, and g is continuous except perhaps at certain "obvious" points where it is ill-defined because of a vanishing denominator in the construction.

Let t parameterize some curve $\boldsymbol{q}(t)$ in the positive region of the normalization subspace. It follows from (A1) that

$$\frac{d}{dt} H(\boldsymbol{q}(t), \boldsymbol{p}) = s \sum_i \dot{q}_i g_i + z \sum_i \dot{q}_i$$

holds, where $\dot{q}_i = dq_i/dt$. But (A2) implies $\sum_i \dot{q}_i = 0$, and

$$\frac{d}{dt} H(\boldsymbol{q}(t), \boldsymbol{p}) = s(\boldsymbol{q}(t), \boldsymbol{p}) \frac{d}{dt} F(\boldsymbol{q}(t), \boldsymbol{p}), \tag{A12}$$

where

$$F(\boldsymbol{q}, \boldsymbol{p}) = \sum_{i=1}^{n} f(q_i, p_i) \tag{A13}$$

for some function f related to g by $g(q_i,p_i)=\partial f(q_i,p_i)/\partial q_i$. Suppose the curve $\boldsymbol{q}(t)$ lies in a level surface of H. Then $dH/dt=0$ and (A12) shows that F is also constant on any such curve, unless perhaps s is zero. However, (A11) shows that, in the interior of the normalization subspace, s is not zero unless $h_i=h_j$ for all i,j, which is true only at the point $\boldsymbol{q}=\boldsymbol{p}$. It follows that F is constant on connected components of level surfaces of H and that F and H are functionally dependent—locally, F can be written as a function of H, with $dF/dH=1/s(\boldsymbol{q},\boldsymbol{p})$. Next we show that the functional dependence is monotonic. If it were not, then dF/dH would change sign at a point $\boldsymbol{q}$ and, therefore, in some neighborhood of $\boldsymbol{q}$ along a level surface of H, but we have seen that s is continuous and nonzero in the interior of the normalization subspace except perhaps at one point ($\boldsymbol{q}=\boldsymbol{p}$); it follows that s is of constant sign. Hence, the functional dependence of F on H is monotonic. The function F in (A13) is therefore equivalent to H, as stated in Theorem I.

References

[1] E. T. Jaynes, "Information theory and statistical mechanics I," *Phys. Rev.*, vol. 106, pp. 620–630, 1957.

[2] ——, "Information theory and statistical mechanics II," *Phys. Rev.*, vol. 108, pp. 171–190, 1957.

[3] ——, "Information theory and statistical mechanics," in *Statistical Physics*, vol. 3, Brandeis Lectures, K. W. Ford, Ed. New York: Benjamin, 1963, pp. 182–218.

[4] ——, "Foundations of probability theory and statistical mechanics," in *Delaware Seminar in the Foundations of Science, Vol. I*, M. Bunge, Ed. New York: Springer—Verlag, 1967, pp. 77–101.

[5] O. C. de Beauregard and M. Tribus, "Information theory and thermodynamics," *Helvetica Physica Acta*, vol. 47, pp. 238–247, 1974.

[6] M. Tribus, *Thermostatics and Thermodynamics*. Princeton, NJ: Van Nostrand, 1961.

[7] A. Katz, *Principles of Statistical Mechanics—The Information Theory Approach*. New York: Freeman, 1967.

[8] A. Hobson, *Concepts in Statistical Mechanics*. New York: Gordon and Breach, 1971.

[9] I. J. Good, "Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables," *Annals Math. Stat.*, vol. 34, pp. 911–934, 1963.

[10] J. P. Noonan, N. S. Tzannes, and T. Costello, "On the inverse problem of entropy maximizations," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 120–123, Jan. 1976.

[11] M. Tribus, *Rational Descriptions, Decisions, and Designs*. New York: Pergamon, 1969.

[12] ——, "The use of the maximum entropy estimate in the estimation of reliability," in *Recent Developments in Information and Decision Processes*, R. E. Machol and D. Gray, Eds. New York: Macmillan, 1962, pp. 102–139.

[13] V. E. Benes, *Mathematical Theory of Connecting Networks and Telephone Traffic*. New York: Academic, 1965.

[14] A. E. Ferdinand, "A statistical mechanics approach to systems analysis," *IBM J. Res. Develop.*, vol. 14, pp. 539–547, 1970.

[15] J. E. Shore, "Derivation of equilibrium and time-dependent solutions to $M/M/\infty//N$ and $M/M/\infty$ queueing systems using entropy maximization," in *1978 Nat. Computer Conf. AFIPS Conf. Proc.*, pp. 483–487, 1978.

[16] M. Chan, "System simulation and maximum entropy," *Operations Research*, vol. 19, pp. 1751–1753, 1971.

[17] E. T. Jaynes, "New engineering applications of the information theory," in *Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability*, J. L. Bogdanoff, Ed. New York: Wiley, 1963, pp. 163–203.

[18] M. Tribus and G. Fitts, "The widget problem revisited," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, pp. 241–248, 1968.

[19] B. Ramakrishna Rau, "The exact analysis of models of program reference strings," Stanford Elec. Lab., Stanford Univ., Stanford, CA, Tech. Rep. 124 (SU-SEL-77-003), Dec. 1976.

[20] A. E. Ferdinand, "A theory of general complexity," *Int. J. General Syst.*, vol. 1, pp. 19–33, 1974.

[21] M. Takatsuji, "An information theoretical approach to a system of interacting elements," *Biol. Cybern.*, vol. 17, pp. 207–210, 1975.

[22] J. M. Cozzolino and M. J. Zahner, "The maximum-entropy distribution of the future market price of a stock," *Operations Research*, vol. 21, pp. 1200–1211, 1973.

[23] E. T. Jaynes, "Probability theory in science and engineering," Field Res. Lab., Socony Mobil Oil Co., Inc., Colloquium Lectures in Pure and Applied Science No. 4, 1958.

[24] ——, *Probability Theory in Science and Engineering*, unpublished lecture notes (available from Physics Dept., Washington Univ., St. Louis, MO), 1972.

[25] ——, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, pp. 227–241, 1968.

[26] J. Burg, "Maximum entropy spectral analysis," Ph.D. dissertation, Stanford Univ., Stanford, CA, Univ. Microfilms No. 75-25,499, 1975.

[27] J. G. Ables, "Maximum entropy spectral analysis," *Astron. Astrophys. Suppl.*, vol. 15, pp. 383–393, 1974.

[28] T. J. Ulrych and T. N. Bishop, "Maximum entropy spectral analysis and autoregressive decomposition," *Rev. Geophysics and Space Physics*, vol. 43, no. 1, pp. 183–200, 1975.

[29] S. J. Wenecke and L. R. D'Addario, "Maximum entropy image reconstruction," *IEEE Trans. Comput.*, vol. C-26, pp. 351–364, 1977.

[30] I. J. Good, *Probability and the Weighing of Evidence*. London: Griffin, 1950.

[31] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[32] A. Wehrl, "Entropy," *Rev. Mod. Phys.*, vol. 50, no. 2, pp. 221–260, 1978.

[33] A. Hobson and B. Cheng, "A comparison of the Shannon and Kullback information measures," *J. Stat. Phys.*, vol. 7, no. 4, pp. 301–310, 1973.

[34] R. W. Johnson, "Axiomatic characterization of the directed divergences and their linear combinations," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 6, pp. 709–716, Nov. 1979.

[35] R. S. Ingarden and A. Kossakowski, "The Poisson probability distribution and information therodynamics," *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys.*, vol. 9, no. 1, pp. 83–85, 1971.

[36] D. V. Gokhale and S. Kullback, *The Information in Contingency Tables*. New York: Marcel Dekker, 1978.

[37] G. S. Arnold and J. L. Kinsey, "Information theory for marginal distributions: Application to energy disposal in an exothermic reaction," *J. Chem. Phys.*, vol. 67, no. 8, pp. 3530–3532, 1977.

[38] R. L. Kashyap, "Prior probability and uncertainty," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 641–650, Nov. 1971.

[39] R. W. Johnson, "Comments on 'Prior probability and uncertainty,'" *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 129–132, 1979.

[40] P. M. Lewis, II, "Approximating probability distributions to reduce storage requirements," *Inform. Contr.*, vol. 2, pp. 214—225, 1959.

[41] J. E. Shore, "Minimum cross-entropy spectral analysis," Naval Res. Lab., Washington, DC 20375, NRL Memo. Rep. 3921, Jan. 1979.

[42] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," Naval Res. Lab., Washington DC 20375, NRL Memo. Rep. 3898, Dec. 1978.

[43] R. W. Johnson, "Determining probability distributions by maximum entropy and minimum cross-entropy," in *APL 1979 Conf. Proc.*, pp. 24–29, 1979.

[44] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[45] J. S. Rowlinson, "Probability, information, and entropy," *Nature*, vol. 225, pp. 1196–1198, Mar. 28, 1970.

[46] J. M. Jauch and J. G. Baron, "Entropy, information, and Szilard's paradox," *Helvetica Physica Acta*, vol. 45, p. 220, 1972.

[47] K. Friedman and A. Shimony, "Jaynes' maximum entropy prescription and probability theory," *J. Stat. Phys.*, vol. 3, no. 4, pp. 381–384, 1971.

[48] M. Tribus and H. Motroni, "Comments on the paper Jaynes'

maximum entropy prescription and probability theory," *J. Stat. Phys.*, vol. 4, no. 2/3, pp. 227–228, 1972.

[49] A. Hobson, "The interpretation of inductive probabilities," *J. Stat. Phys.*, vol. 6, no. 2/3, pp. 189–193, 1972.

[50] A. I. Khinchin, *Mathematical Foundations of Information Theory*. New York: Dover, 1957.

[51] D. K. Faddeyev, "On the concept of entropy of a finite probabilistic scheme," *Uspekhi Mat. Nauk*, vol. 11, no. 1(67), pp. 227–231, 1956.

[52] D. W. North, "The invariance approach to the probabilistic encoding of information," Ph.D. dissertation, Stanford Univ., Stanford, CA, Univ. Microfilms No. 70-22,189, 1970.

[53] I. M. Gel'fand and A. M. Yaglom, "Calculation of the amount of information about a random function contained in another such function," *Uspekhi Mat. Nauk*, vol. 12, no. 1(73) (Trans: *American Mathematical Society Translation*, Series 2, vol. 12. Providence, RI: Amer. Math. Soc., 1959, pp. 199–246).

[54] A. Hobson, "A new theorem of information theory," *J. Stat. Phys.*, vol. 1, no. 3, pp. 383–391, 1969.

[55] Pl. Kannappan, "On Shannon's entropy, directed divergence, and inaccuracy," *Z. Wahrscheinlichkeitstheorie verw. Geb.*, vol. 22, pp. 95–100, 1972.

[56] ——, "On directed divergence and inaccuracy," *Z. Wahrscheinlichkeitstheorie verw. Geb.*, vol. 25, pp. 49–55, 1972.

[57] J. Aczél and Z. Daróczy, *On Measures of Information and Their Characterizations.* New York: Academic, 1975.

[58] J. Van Campenhout and T. M. Cover, "Maximum entropy and conditional probability," Dept. Stat., Stanford Univ., Stanford, CA, Tech. Rep. No. 32, July 1978.

[59] R. T. Cox, "Probability, frequency, and reasonable expectation," *Am. J. Phys.*, vol. 14, pp. 1–13, 1946.

[60] ——, *The Algebra of Probable Inference.* Baltimore, MD: Hopkins, 1961.

[61] L. Janossy, "Remarks on the foundation of probability calculus," *Acta Phys. Acad. Hungar.*, vol. 4, pp. 333–349, 1955.

[62] J. Aczél, "A solution of some problems of K. Borsuk and L. Janossy," *Acta Phys. Acad. Hungar.*, vol. 4, pp. 351–362, 1955.

[63] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. Roy. Soc.*, vol. A186, pp. 453–461, 1946.

[64] J. Aczél, *Lectures on Functional Equations and their Applications*. New York: Academic, 1966.

[65] C. T. Ng, "Representation for measures of information with the branching property," *Inform. Contr.*, vol. 25, pp. 45–56, 1974.

[66] I. Csiszár, "*I*-divergence geometry of probability distributions and minimization problems," *Ann. Prob.*, vol. 3, pp. 146–158, 1975.

[67] W. M. Elsasser, "On quantum measurements and the role of the uncertainty relations in statistical mechanics," *Phys. Rev.*, vol. 52, pp. 987–999, 1937.

Lower Bounds for Constant Weight Codes

R. L. GRAHAM AND N. J. A. SLOANE, FELLOW, IEEE

***Abstract*—Let $A(n,2\delta,w)$ denote the maximum number of codewords in any binary code of length n, constant weight w, and Hamming distance 2δ. Several lower bounds for $A(n,2\delta,w)$ are given. For w and δ fixed, $A(n,2\delta,w) \gtrsim n^{w-\delta+1}/w!$ and $A(n,4,w) \sim n^{w-1}/w!$ as $n \to \infty$. In most cases these are better than the "Gilbert bound." Revised tables of $A(n,2\delta,w)$ are given in the range $n \leqslant 24$ and $\delta \leqslant 5$.**

I. Lower Bounds for $A(n,4,w)$

Theorem 1:

$$A(n,4,w) \geqslant \frac{1}{n}\binom{n}{w}.$$

Proof: Let $\mathbb{F}_w^n$ denote the set of $\binom{n}{w}$ binary vectors of length n and weight w, and let $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$ denote the residue classes modulo n. Consider the map

$$T\colon \mathbb{F}_w^n \to \mathbb{Z}_n$$

whose value at $\boldsymbol{a} = (a_0, \cdots, a_{n-1}) \in \mathbb{F}_w^n$ is

$$\begin{aligned} T(\boldsymbol{a}) &= \sum_{a_i=1} i \qquad (\operatorname{mod} n) \\ &= \sum_{i=0}^{n-1} i a_i \qquad (\operatorname{mod} n). \end{aligned} \tag{1}$$

For $0 \leqslant i \leqslant n-1$ let C_i be the constant weight code $T^{-1}(i)$. We claim that the Hamming distance between any two distinct codewords of C_i, say $\boldsymbol{a}$ and $\boldsymbol{b}$, is at least four. For suppose it is two. Since $\boldsymbol{a}$ and $\boldsymbol{b}$ have weight w this means that $\boldsymbol{a}$ and $\boldsymbol{b}$ agree everywhere except for two positions, one (say the rth) where $\boldsymbol{a}$ is one and $\boldsymbol{b}$ is zero and another (say the sth) where $\boldsymbol{a}$ is zero and $\boldsymbol{b}$ is one. But $T(\boldsymbol{a}) = T(\boldsymbol{b}) = i$, so from (1)

$$\begin{aligned} T(\boldsymbol{a}) &= x + r = i \qquad (\operatorname{mod} n), \\ T(\boldsymbol{b}) &= x + s = i \qquad (\operatorname{mod} n) \end{aligned}$$

for some $x \in \mathbb{Z}_n$. This implies $r \equiv s (\operatorname{mod} n)$, which is impossible. Thus C_i has a Hamming distance of at least four

Manuscript received February 20, 1979; revised August 22, 1979.
The authors are with Bell Laboratories, Inc., Murray Hill, NJ 07974.

0018-9448/80/0100-0037$00.75 ©1980 IEEE