codes," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 794–805, Nov. 1972.
[5] C. L. Mallows and N. J. A. Sloane, "An upper bound for self-dual codes," *Inform. Contr.*, vol. 6, pp. 79–94, 1963.
[6] G. Cohen, P. Godlewski, and S. Perrine, "Sur les idempotents des codes," C. R. *Acad. Sci.*, vol. 284, Feb. 28, 1977.
[7] P. Delsarte, "Four fundamental parameters of a code and their combinatorial significance," *Inform. Contr.*, vol. 23, pp. 407–438, 1973.
[8] M. G. Karpovsky, "On the weight distribution of binary linear codes," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 105–109, Jan. 1979.
[9] ——, *Finite Orthogonal Series in the Design of Digital Devices*. New York: Wiley; Jerusalem: IUP, 1976.
[10] M. Deza, "Comparison of arbitrary additive noises," (in Russian), *Problemy Peredatchi Informatsii*, vol. 3, pp. 29–38, 1965.
[11] M. Deza and F. Hoffman, "Some results related to generalized Varshamov–Gilbert bound," *IEEE Trans. Inform. Theory*, pp. 517–518, July 1977.
[12] M. G. Karpovsky and V. D. Milman, "On subspaces contained in subsets of finite homogenous spaces," *Discrete Mathematics*, vol. 22, pp. 273–280, 1978.
[13] A. Tietavainen, "On nonexistence of perfect codes and related topics in combinatorics," (M. Hall, Jr. and J. H. Van Lint, Eds.), *Math. Center Tracks*, vol. 55, pp. 158–178, 1974.
[14] K. C. Andrews and K. L. Caspari, "A generalized technique for spectral analysis," *IEEE Trans. Comput.*, vol. C-19, pp. 16–25, 1970.
[15] M. G. Karpovsky and E. S. Moskalev, "Utilization of autocorrelation functions for the realization of systems of logical functions," *Automat. and Remote Contr.*, vol. 31, N2, pp. 243–250, Feb., 1 (translated from *Automatika i Telemekhanika*, N2, pp. 83–90, 1 Russian).
[16] M. G. Karpovsky and E. A. Trachtenberg, "Linear checking e tions and error-correcting capability for computation chann *Proc. 1977 IFIP Congress*. New York: North–Holland, 1977.
[17] N. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Si Processing*. New York: Springer–Verlag, 1975.
[18] K. A. Post, private communication.
[19] G. Cohen and P. Frankl, "On tilings of the binary vector spa submitted to *Discrete Mathematics*.
[20] J. Vasiliev, "On nongroup closed packed codes," *Probl. Kibern.*, 8, pp. 337–339, 1962.
[21] A. Tietavainen, "Nonexistence of perfect codes," *SIAM J. A Math.*, vol. 24, pp. 88–96, 1973.
[22] O. S. Rothaus, "On Bent Functions," *J. Combinatorial Theory* vol. 20, pp. 300–305, 1976.
[23] H. F. Mattson, Jr., private communication.
[24] N. J. A. Sloane and R. J. Dick, "On the enumeration of cose first-order Reed–Muller codes," *IEEE Int. Conf. on Commun* Montreal, 1971.
[25] H. F. Mattson, Jr. and J. R. Schatz, "Maximum-Leader codes appear.
[26] J. J. Mykkeltveit, "The covering radius of the (128, 8) Reed–M code is 56," *IEEE Trans. Inform. Theory.*, vol. IT-26, pp. 359–May 1980.
[27] G. Cohen, Private Communication.
[28] E. F. Assmus, Jr. and H. F. Mattson, Jr., "Coding and combin ics," *SIAM Rev.*, vol. 16, no. 3, pp. 345–388, July 1974.

Properties of Cross-Entropy Minimization

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

***Abstract*—The principle of minimum cross-entropy (minimum directed divergence, minimum discrimination information) is a general method of inference about an unknown probability density when there exists a prior estimate of the density and new information in the form of constraints on expected values. Various fundamental properties of cross-entropy minimization are proven and collected in one place. Cross-entropy's well-known properties as an information measure are extended and strengthened when one of the densities involved is the result of cross-entropy minimization. The interplay between properties of cross-entropy minimization as an inference procedure and properties of cross-entropy as an information measure is pointed out. Examples are included and general analytic and computational methods of finding minimum cross-entropy probability densities are discussed.**

I. INTRODUCTION

THE PRINCIPLE of minimum cross-entropy provides a general method of inference about an unknown probability density $q^\dagger$ when there exists a prior estimate of $q^\dagger$ and new information about $q^\dagger$ in the form of constrai on expected values. The principle states that, of all densities that satisfy the constraints, one should choose *posterior* q with the least cross-entropy $H[q, p]$ $\int dx\, q(x) \log (q(x)/p(x))$, where p is a *prior* estimate of

Cross-entropy minimization was first introduced Kullback [1], who called it minimum directed diverge and minimum discrimination information. The principle maximum entropy [2], [3] is equivalent to cross-entr minimization in the special case of discrete spaces a uniform priors. Cross-entropy minimization has a l history of applications in a variety of fields (for a list references, see [4]). Recently, the theory has been appl to problems in spectral analysis [5], speech coding [6], a pattern recognition [7].

It is useful and convenient to view cross-entropy mini zation as one implementation of an abstract informat operator $\circ$ that takes two arguments—a prior and n information—and yields a posterior. Thus, we write posterior q as $q = p \circ I$, where I stands for the kno

Manuscript received October 18, 1979; revised March 14, 1980.
The authors are with the Naval Research Laboratory, Code 7591, Washington, DC 20375.

0018-9448/81/0700-0472$00.75

constraints on expected values. Recently we have shown that, if the operator $\circ$ is required to satisfy certain axioms of consistent inference, and if $\circ$ is implemented by means of functional minimization, then the principle of minimum cross-entropy follows necessarily [4].

Cross-entropy minimization satisfies a variety of interesting and useful properties beyond those expressed or implied by the axioms in [4]. It is the purpose of this paper to state and prove these properties. For completeness, we also restate the axioms from [4] (Property 1, and (12), (14), and (16)). Some of the properties of cross-entropy minimization just reflect well-known properties of cross-entropy [1], [8], but there are surprising differences as well. For example, cross-entropy does not generally satisfy a triangle relation involving three arbitrary probability densities. But in certain important cases involving densities that result from cross-entropy minimization, cross-entropy satisfies reverse triangle inequalities and triangle equalities. (See Properties 10, 12, and 13.)

The combined properties of cross-entropy and cross-entropy minimization have recently been shown to be useful in the field of speech processing. In particular, one formulation of the standard linear prediction coding (LPC) equations is based on minimizing a distortion measure introduced by Itakura and Saito [9]. In [10] it is shown that the Itakura–Saito distortion measure is a special case of asymptotic cross-entropy, and in [6] it is shown that the standard LPC equations can be obtained directly by cross-entropy minimization. They newly developed technique of speech coding by vector quantization [11] was also derived in [6] directly by cross-entropy minimization. Furthermore, the original derivation of vector quantization in [11] was carried out by exploiting properties of the Itakura–Saito distortion measure—for example, a triangle equality—that turn out to be special cases of some of the properties presented herein (Properties 12, 14, 15). These properties have since been used in refining Kullback's classification method [1, p. 83], yielding a method that is optimal in a precise information–theoretic sense [7] and computationally efficient.

After introducing necessary definitions and notation in Section II, we first consider properties that are valid for both equality and inequality constraints on expected values (Section III), and then consider properties that are valid only for equality constraints (Section IV). We conclude with a brief discussion in Section V. We also include an Appendix in which we discuss general analytic and computational methods for finding minimum cross-entropy posteriors.

II. Definitions and Notation

In this section, we introduce the same notation as in [4, sec. II]. The discussion here places somewhat greater emphasis on mathematical questions relating to the existence of minimum cross-entropy solutions. (See also the discussion following Property 1.)

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 $q(\boldsymbol{x}) \geq 0$ for $\boldsymbol{x} \in \boldsymbol{D}$ and

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

We use a 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*

$$\int_{\boldsymbol{D}} d\boldsymbol{x}\, q^\dagger(\boldsymbol{x}) a_k(\boldsymbol{x}) = \bar{a}_k \tag{2}$$

and *inequality constraints*

$$\int_{\boldsymbol{D}} d\boldsymbol{x}\, q^\dagger(\boldsymbol{x}) c_k(\boldsymbol{x}) \geq \bar{c}_k \tag{3}$$

for known sets of functions a_k, c_k, and known values $\bar{a}_k, \bar{c}_k$. The probability densities that satisfy such constraints always comprise a convex subset $\mathcal{I}$ of $\mathcal{D}$. (A set $\mathcal{I}$ is *convex* if, given $0 \leq A \leq 1$ and $q, r \in \mathcal{I}$, it contains the weighted average $Aq + (1 - A)r$.) We refer to the functions a_k, c_k as *constraint functions* and $\mathcal{I}$ as a *constraint set*. For a given constraint set there may of course be more than one set of constraint functions in terms of which it may be defined. We frequently suppress mention of a particular set of constraint functions, using the notation $I = (q^\dagger \in \mathcal{I})$ to mean that $q^\dagger$ is a member of the constraint set $\mathcal{I} \subseteq \mathcal{D}$ and referring to I as a *constraint*. 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:

$$p(\boldsymbol{x} \in \boldsymbol{D}) > 0. \tag{4}$$

(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 the cross-entropy $H[q, p]$ in the constraint set $\mathcal{I}$:

$$H[q, p] = \min_{q' \in \mathcal{I}} H[q', p], \tag{5}$$

where

$$H[q, p] = \int_{\boldsymbol{D}} d\boldsymbol{x}\, q(\boldsymbol{x}) \log\left(q(\boldsymbol{x})/p(\boldsymbol{x})\right). \tag{6}$$

We introduce an "information operator" $\circ$ that expresses (5) 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 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 $x \in S$, corresponding to any $q \in \mathcal{D}$. We use

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

as a shorthand notation for (8).

In making the restriction (4) we assume that 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 D must be redefined before q is used as a prior in a further application of $\circ$. The restriction (4) does not significantly restrict our results, but it does help in avoiding certain technical problems that would otherwise result from division by $p(x)$. For more discussion, see [8].

When D is a discrete set of system states, densities are replaced by discrete distributions and integrals by sums in the usual way. In a more general setting for the discussion than we have chosen, D would be a measurable space, and p and q would be replaced by prior and posterior probability measures. By continuing to write in terms of probability densities, we would then be implicitly assuming some underlying measure with respect to which the rest were absolutely continuous. Indeed such a measure certainly exists if we demand that no event with zero prior probability can have positive posterior probability, which in the present context we are in effect demanding by assuming (4).

III. Properties Given General Constraints

This section concerns properties that apply in the case of both equality and inequality constraints (2), (3). We follow the formal statement of each property with a brief discussion and then a proof or an appropriate reference. Throughout we assume a system with possible states D, probability density $q^\dagger \in \mathcal{D}$, an arbitrary prior $p \in \mathcal{D}$, and arbitrary new information $I = (q^\dagger \in \mathcal{I})$, where $\mathcal{I} \subseteq \mathcal{D}$ contains at least one density q such that $H(q, p) < \infty$.

Property 1 (Uniqueness): The posterior $q = p \circ I$ is unique.

Discussion: A solution to the cross-entropy minimization problem, if one exists, is unique provided only that $H[q, p]$ is not identically infinite as q ranges over the constraint set $\mathcal{I}$. To guarantee that a solution exists, a little more is required. One condition that suffices for existence is that, in addition to containing a density q with finite cross-entropy, the constraint set $\mathcal{I}$ be *closed*. (We call $\mathcal{I}$ closed if it contains every probability density q that is a limit of densities $q_i \in \mathcal{I}$. Limits are taken in the sense that $q_i \to q$ means $\int |q_i(x) - q(x)|\, dx \to 0$.) For $\mathcal{I}$ to be closed, it suffices in turn that the constraint functions be bounded. (And conversely, any closed convex set of probability densities can be defined by equality and inequality constraints (2), (3) with bounded constraint functions, except that infinitely many may be required.) It is also possible to assert existence of $p \circ I$ under less stringent conditions, which do not imply that $\mathcal{I}$ is closed—see Appendix A in this paper and [12, Theorem 3.3]. This is fortunate, since a number of examples of practical importance involve unbounded constraint functions.

Proof of 1: See [12], [4, sec. IV-E].

Property 2: The posterior satisfies $q = p \circ I = p$ if and only if the prior satisfies $p \in \mathcal{I}$.

Discussion: If one views cross-entropy minimization as an inference procedure, it makes sense that the posterior should be unchanged from the prior if the new information does not contradict the prior in any way. Consider the example of (A10)–(A12). If $a_k = \bar{x}_k$ for $k = 1, \cdots, n$, then $q(x) = p(x)$.

Proof of 2: Property 2 follows directly from the property of cross-entropy that $H[q, p] \geq 0$ with $H[q, p] = 0$ only if $q = p$ ([1, p. 14]).

Property 3 (Idempotence): $(p \circ I) \circ I = p \circ I$.

Discussion: Taking the same information into account twice has the same effect as taking it into account once.

Proof of 3: Since $(p \circ I) \in \mathcal{I}$, idempotence follows from Property 2.

Property 4: Let constraints I_1 and I_2 be given by $I_1 = (q^\dagger \in \mathcal{I}_1)$ and $I_2 = (q^\dagger \in \mathcal{I}_2)$, for constraint sets $\mathcal{I}_1, \mathcal{I}_2 \subseteq \mathcal{D}$. If $(p \circ I_1) \in \mathcal{I}_2$ holds, then

$$p \circ I_1 = (p \circ I_1) \circ (I_1 \wedge I_2) = (p \circ I_1) \circ I_2 = p \circ (I_1 \wedge I_2) \tag{10}$$

also holds.

Discussion: If the result of taking information I_1 into account already satisfies constraints imposed by additional information I_2, taking I_2 into account in various ways has no effect. For example, let I_1 and I_2 be the constraints

$$\int_0^\infty dx\, x q^\dagger(x) = a$$

and

$$\int_0^\infty dx\, x^2 q^\dagger(x) = 2a^2, \tag{11}$$

respectively. For an exponential prior $p(x) = r \exp(-rx)$, the posterior given I_1 is $q = p \circ I_1 = (1/a) \exp(-x/a)$ (see (A10)–(A12)). The second moment of q is just $2a^2$, so that q satisfies $q \in \mathcal{I}_2$, as well as $q = q \circ (I_1 \wedge I_2)$, $q = q \circ I_2$, and $q = p \circ (I_1 \wedge I_2)$. If the right side of (11) were anything but $2a^2$, the result of $p \circ (I_1 \wedge I_2)$ would be a truncated Gaussian or undefined and not an exponential [13, p. 133–140].

Proof of 4: Since $(p \circ I_1) \in \mathcal{I}_1$ holds and, by assumption, $(p \circ I_1) \in \mathcal{I}_2$ also holds, it follows that $(p \circ I_1) \in (\mathcal{I}_1 \cap \mathcal{I}_2)$ holds. The first two equalities of (10) then follow directly from Properties 2 and 3. The last equality of (10) follows from $q = p \circ I_1$ having the smallest cross-entropy $H[q, p]$ of all densities in $\mathcal{I}_1$ and therefore in $\mathcal{I}_1 \cap \mathcal{I}_2$.

Property 5 (Invariance): Let Γ be a coordinate transformation from $x \in D$ to $y \in D'$ with $(\Gamma q)(y) = J^{-1} q(x)$, where J is the Jacobian $J = \partial(y)/\partial(x)$. Let $\Gamma\mathcal{D}$ be the set of densities Γq corresponding to densities $q \in \mathcal{D}$. Let $(\Gamma\mathcal{I})$

$\subseteq (\mathbf{\Gamma}\mathcal{D})$ correspond to $\mathcal{I} \subseteq \mathcal{D}$. Then

$$(\mathbf{\Gamma}p) \circ (\mathbf{\Gamma}I) = \mathbf{\Gamma}(p \circ I) \tag{12}$$

and

$$H[\mathbf{\Gamma}(p \circ I), \mathbf{\Gamma}p] = H[p \circ I, p] \tag{13}$$

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

Discussion: Equation (12) states that the same answer is obtained when one solves the inference problem in two different coordinate systems, in that the posteriors in the two systems are related by the coordinate transformation. Moreover, the cross-entropy between the posteriors and the priors has the same value in both coordinate systems.

As an example, let y_1 and y_2 be the real and imaginary parts of a complex sinusoidal signal; let x_1 be the total power $x_1 = y_1^2 + y_2^2$, and let x_2 be the phase, so that

$$(y_1, y_2) = \mathbf{\Gamma}(x_1, x_2) = \left(x_1^{1/2}\cos(x_2), x_1^{1/2}\sin(x_2)\right).$$

Then the Jacobian is constant:

$$\boldsymbol{J} = \det\begin{bmatrix} \frac{1}{2}x_1^{-1/2}\cos(x_2) & -x_1^{1/2}\sin(x_2) \\ \frac{1}{2}x_1^{-1/2}\sin(x_2) & x_1^{1/2}\cos(x_2) \end{bmatrix} = 1/2.$$

Therefore, if the prior density $p(\boldsymbol{x})$ is uniform in some region in the $\boldsymbol{x}$ coordinate space, the transformed prior $(\mathbf{\Gamma}p)(\boldsymbol{y})$ will be uniform on a corresponding region in the $\boldsymbol{y}$ coordinate space. In particular, suppose

$$p(\boldsymbol{x}) = \begin{cases} 1/2\pi R^2, & \left(0 \le x_1 \le R^2, \quad -\pi < x_2 \le \pi\right) \\ 0, & \text{otherwise,} \end{cases}$$

which makes p uniform in a certain rectangle. Then we find that

$$(\mathbf{\Gamma}p)(\boldsymbol{y}) = \begin{cases} 1/\pi R^2, & \left(y_1^2 + y_2^2 \le R^2\right) \\ 0, & \text{otherwise,} \end{cases}$$

which makes $\mathbf{\Gamma}p$ uniform on a certain disk. (Notice $1/\pi R^2 = \boldsymbol{J}^{-1}(1/2\pi R^2)$.) Let new information I specify the expected power

$$\int_0^\infty dx_1 \int_{-\pi}^{\pi} dx_2\, x_1 q^\dagger(\boldsymbol{x}) = P.$$

The resulting posterior $q = p \circ I$ is exponential with respect to x_1:

$$q(\boldsymbol{x}) = \begin{cases} A\exp[-\lambda x_1], & \left(0 \le x_1 \le R^2, \quad -\pi < x_2 < \pi\right) \\ 0, & \text{otherwise,} \end{cases}$$

for certain constants A and λ. The new information in the transformed coordinates, $\mathbf{\Gamma}I$, is

$$\int dy_1 \int dy_2 \left(y_1^2 + y_2^2\right) q'^\dagger(\boldsymbol{y}) = P,$$

and the resulting posterior $q' = (\mathbf{\Gamma}p) \circ (\mathbf{\Gamma}I)$ has the form of a bivariate Gaussian inside the disk:

$$q'(\boldsymbol{y}) = \begin{cases} 2A\exp\left[-\lambda\left(y_1^2 + y_2^2\right)\right], & \left(y_1^2 + y_2^2 \le R^2\right) \\ 0, & \text{otherwise} \end{cases}$$

The two posteriors q and q' are related by $q'(\boldsymbol{y}) = (\mathbf{\Gamma}q)(\boldsymbol{y})$, as stated in (12).

Proof of 5: See [4, sec. IV-E]. The proof of (12) follows directly from the fact that cross-entropy is transformation invariant. Equation (13) is just a special case of this invariance.

Property 6 (System Independence): Let there be two systems, with sets $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ of states and probability densities of states $q_1^\dagger \in \mathcal{D}_1$ and $q_2^\dagger \in \mathcal{D}_2$. 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 $\mathcal{I}_2 \subset \mathcal{D}_2$. Then

$$(p_1 p_2) \circ (I_1 \wedge I_2) = (p_1 \circ I_1)(p_2 \circ I_2) \tag{14}$$

and

$$H[q_1 q_2, p_1 p_2] = H[q_1 p_1] + H[q_2, p_2], \tag{15}$$

hold, where $q_1 = p \circ I_1$ and $q_2 = p \circ I_2$.

Discussion: Property 6 states that it does not matter whether one accounts for independent information about two systems separately or together in terms of a joint density. Whether the two systems are in fact independent is irrelevant; the property applies as long as there are independent priors and independent new information. Examples can be easily generated from the multivariate exponential and multivariate Gaussian examples in the Appendix.

Proof of 6: See [4, sec. IV-E].

Property 7 (Subset Independence): Let $\boldsymbol{S}_1, \cdots, \boldsymbol{S}_n$ be disjoint sets whose union is $\boldsymbol{D}$. Let the new information I comprise information about the conditional densities $q^\dagger * \boldsymbol{S}_i$. Thus, $I = I_1 \wedge I_2 \wedge \cdots \wedge I_n$, and $I_i = (q^\dagger * \boldsymbol{S}_i \in \mathcal{I}_i)$, where $\mathcal{I}_i \subseteq \mathcal{S}_i$ and $\mathcal{S}_i$ 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

$$\int_{\boldsymbol{S}_i} d\boldsymbol{x}\, q(\boldsymbol{x}) = m_i$$

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

$$(p \circ (I \wedge M)) * \boldsymbol{S}_i = (p * \boldsymbol{S}_i) \circ I_i \tag{16}$$

and

$$H[p \circ (I \wedge M), p] = \sum_i m_i H[q_i, p_i] + \sum_i m_i \log\left(\frac{m_i}{s_i}\right) \tag{17}$$

hold, where $p_i = p * \boldsymbol{S}_i$, $q_i = p_i \circ I_i$, and the s_i are the prior probabilities of being in each subset,

$$s_i = \int_{\boldsymbol{S}_i} d\boldsymbol{x}\, p(\boldsymbol{x}). \tag{18}$$

Discussion: This property concerns situations in which the set of states $\boldsymbol{D}$ decomposes naturally into disjoint subsets $\boldsymbol{S}_i$, in which the new information $I = I_1 \wedge I_2 \wedge \cdots \wedge I_n$ comprises disjoint information about the conditional probability densities $q^\dagger * \boldsymbol{S}_i$ in each subset, and in which there is also new information M giving the total probability m_i of being in each subset $\boldsymbol{S}_i$. Given this information, there are two ways to obtain posterior conditional densities for each subset. One way is to obtain a

conditional posterior $(p * S_i) \circ I_i$ from each conditional prior $p * S_i$. Another way is to obtain a posterior $q = p \circ (I \wedge M)$ for the whole system and then to compute a conditional posterior $q * S_i$. Property 7 states that the results are the same in both cases; it does 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.

To illustrate Property 7, suppose that a six-sided die was rolled a large number of times. The frequencies with which the different die faces turned up were not recorded individually, but the mean number of spots showing was determined separately for the odd results and for the even results. There is no prior reason to expect any face of the die to turn up more often than any other. Indeed, the probability for an odd number of spots showing was found to be 0.5. However, the mean number of spots showing, given that the number is odd, was found to be four; the mean number of spots showing, given that the number is even, also was found to be four. Given this information, we are asked to estimate the probability for each face of the die to turn up, as well as the conditional probability given whether the face is odd or even. Let $S_1 = \{1, 3, 5\}$ and $S_2 = \{2, 4, 6\}$. We will first solve the problem on S_1 and S_2 separately and then solve it on $S_1 \cup S_2$.

In all cases, the prior is uniform. The prior p_1 on S_1 is $p_1(1) = p_1(3) = p_1(5) = 1/3$. The information I_1 giving the expected value for an odd number of spots is

$$\sum_{n \in S_1} n q_1^\dagger(n) = 4;$$

therefore, we compute a posterior $q_1 = p_1 \circ I_1$ on S_1 by minimizing $H[q_1, p_1]$ subject to $q_1(1) + 3q_1(3) + 5q_1(5) = 4$. The result is

$$q_1(1) = 0.1162, \quad q_1(3) = 0.2676, \quad q_1(5) = 0.6162. \tag{19}$$

Similarly, the prior p_2 on S_2 is $p_2(2) = p_2(4) = p_2(6) = 1/3$, the posterior q_2 is subject to the constraint I_2, $2q_2(2) + 4q_2(4) + 6q_2(6) = 4$, and the result of minimizing $H[q_2, p_2]$ is

$$q_2(2) = 1/3, \quad q_2(4) = 1/3, \quad q_2(6) = 1/3. \tag{20}$$

On $S_1 \cup S_2$, the prior p is $p(1) = p(2) = \cdots = p(6) = 1/6$. The information I_1, which concerns $q^\dagger * S_1$, may be expressed as $q^\dagger(1) + 3q^\dagger(3) + 5q^\dagger(5) = 4(q^\dagger(1) + q^\dagger(3) + q^\dagger(5))$. We therefore subject the posterior q to the constraint

$$-3q(1) - q(3) + q(5) = 0. \tag{21}$$

Similarly, because of I_2, we have the constraint

$$-2q(2) + 2q(6) = 0. \tag{22}$$

Finally, because of the information M, we subject q to the constraint

$$q(1) - q(2) + q(3) - q(4) + q(5) - q(6) = 0, \tag{23}$$

since this is equivalent to $q(1) + q(3) + q(5) = 0.5 = q(2) + q(4) + q(6)$. Upon minimizing $H[q, p]$ subject to the constraints (21–23), we find that $q = p \circ (I_1 \wedge I_2 \wedge M)$ is given by

$$\begin{aligned} q(1) &= 0.0581, & q(2) &= 1/6, \\ q(3) &= 0.1338, & q(4) &= 1/6, \\ q(5) &= 0.3081, & q(6) &= 1/6. \end{aligned} \tag{24}$$

To find the conditional probabilities $q * S_1$ and $q * S_2$, we divide both columns in this result by 0.5; the results agree with q_1 and q_2 as computed above ((19), (20)), and as stated in (16).

Proof of 7: See [4, sec. IV-E].

Property 8 (Weak Subset Independence): For the same definitions and notation as Property 7,

$$(p \circ I) * S_i = (p * S_i) \circ I_i \tag{25}$$

and

$$H[p \circ I, p] = \sum_i r_i H[q_i, p_i] + \sum_i r_i \log\left(\frac{r_i}{s_i}\right) \tag{26}$$

hold, where $p_i = p * S_i$, $q_i = p_i \circ I_i$, the s_i are the prior probabilities of being in each subset (18), and the r_i are the posterior probabilities of being in each subset,

$$r_i = \int_{S_i} dx\, q(x), \tag{27}$$

for $q = p \circ I$.

Discussion: This property states that the two ways of obtaining the posterior conditional densities also lead to the same result in the case when one does not have information giving the total probability in each subset. Results for the full system posterior, however, will not in general be the same for the cases covered by Properties 7 and 8. That is, $q \circ I$ and $q \circ (I \wedge M)$ will not generally be equal.

To illustrate Property 8, we solve the example problem from Property 7, omitting the information M that the probability of an odd (or of an even) number of spots is 0.5. The separate solutions on S_1 and S_2 proceed exactly as before and yield the same posteriors q_1 and q_2. The solution on $S_1 \cup S_2$ differs from the previous one only in that we minimize $H[q, p]$ subject to the constraints (21) and (22), but not subject to (23). The result, $q' = p \circ (I_1 \wedge I_2)$, is given by

$$\begin{aligned} q'(1) &= 0.0524, & q'(2) &= 0.1831, \\ q'(3) &= 0.1206, & q'(4) &= 0.1831, \\ q'(5) &= 0.2778, & q'(6) &= 0.1831, \end{aligned}$$

and differs from the previous result (24). Moreover, the subset probabilities r_1 and r_2 do not satisfy M: summing the two columns gives $r_1 = 0.4508$ and $r_2 = 0.5492$. Dividing the two columns respectively by r_1 and r_2, however, gives the same conditional probabilities as before: $q' * S_1 = q_1$ and $q' * S_2 = q_2$ (see (19), (20)).

Proof of 8: For $q = p \circ I$, let r_i be given by (27). Then let R be information $R = q^\dagger \in \mathcal{R}$, where $\mathcal{R}$ is the set of densities satisfying (27). It follows from Property 4 that

$p \circ I = p \circ (I \wedge R)$ holds; (25) and (26) then follow from Property 7.

Property 9 (Subset Aggregation): Let $S_1, S_2, \cdots, S_n$ be disjoint sets whose union is D. Let ψ be a transformation such that, for any $q \in \mathcal{D}$, $q' = \psi q$ is a discrete distribution with

$$q'(x_i) = \int_{S_i} dx\, q(x),$$

where x_i is a discrete state corresponding to $x \in S_i$. Thus the transformation ψ aggregates the states in each subset S_i. Suppose new information $I' = ((\psi q^\dagger) \in \mathcal{I}')$ is obtained about the aggregate distribution $\psi q^\dagger$, where $\mathcal{I}'$ is a convex set of discrete distributions. Then for any prior $p \in \mathcal{D}$,

$$p * S_i = (p \circ I) * S_i, \tag{28}$$

$$(\psi p) \circ I' = \psi(p \circ I), \tag{29}$$

and

$$H[\psi(p \circ I), \psi p] = H[p \circ I, p] \tag{30}$$

all hold, where $I = \psi^{-1} I'$ is the information I' expressed in terms of $q^\dagger$ instead of in terms of $\psi q^\dagger$. (That is, $I = (q^\dagger \in (\psi^{-1}\mathcal{I}'))$, where $(\psi^{-1}\mathcal{I}') \subseteq \mathcal{D}$ are the densities q such that $(\psi q) \in \mathcal{I}'$.)

Discussion: Note that (29) and (30), in which ψ is a many-to-one mapping, have the same form as the invariance property, which holds for one-to-one coordinate transformations Γ (see (12), (13)). Indeed, both invariance and subset aggregation can be viewed as special cases of a more general, measure-theoretic invariance. In mathematical terms, the operator $\circ$ is functorial.

Proof of 9: Let the information I' be a set of known expectations $\sum_i g_{ki} q^{\dagger\prime}(x_i)$, for $k = 1, \cdots, m$, or bounds on these expectations, where $q^{\dagger\prime} = \psi q^\dagger$. In terms of $q^\dagger$, this becomes a set of known or bounded expectations

$$\int_D dx\, q^\dagger(x) f_k(x),$$

where $f_k(x \in S_i) = g_{ki}$ is constant in each subset S_i. The posterior $q = p \circ I$ has the form

$$q(x) = p(x) \exp\left(-\lambda_0 - \sum_{k=1}^{m} \lambda_k f_k(x)\right), \tag{31}$$

where some of the terms in the summation over k may be omitted in the case of inequality constraints (see (A4)). Since f_k is constant on each subset, (31) has the form $q(x \in S_i) = A_i p(x \in S_i)$, where A_i is a subset dependent constant. This proves (28). In general, for any $q, p \in \mathcal{D}$, the cross-entropy $H[q, p]$ can be expressed [4] as

$$H[q, p] = \sum_i r_i H[q_i, p_i] + \sum_i r_i \log\left(\frac{r_i}{s_i}\right), \tag{32}$$

where $p_i = p * S_i$, $q_i = q * S_i$,

$$s_i = \int_{S_i} dx\, p(x), \quad \text{and} \quad r_i = \int_{S_i} dx\, q(x).$$

In the present case we have $q_i = p_i$ from (28). Since $H[q_i, q_i] = 0$, (32) reduces to

$$H[q, p] = \sum_i r_i \log\left(\frac{r_i}{s_i}\right) = H[\psi q, \psi p].$$

Minimizing the left side subject to I, yielding $q = p \circ I$, is equivalent to minimizing the right side subject to I'. This proves (29) and (30).

Property 10 (Triangle Relations): For any $r \in \mathcal{I}$,

$$H[r, p] \geq H[r, q] + H[q, p], \tag{33}$$

where $q = p \circ I$. When I is determined by a finite set of equality constraints only, equality holds in (33).

Proof of 10: We have

$$H[q, p] = \min_{q' \in \mathcal{I}} H[q', p].$$

The densities $q' = (1 - t)q + tr$ belong to $\mathcal{I}$ for all $t \in [0, 1]$ since $q \in \mathcal{I}$, $r \in \mathcal{I}$, and $\mathcal{I}$ is convex. For all such t we therefore have

$$H[(1 - t)q + tr, p] \geq H[q, p], \tag{34}$$

or $F(t) \geq F(0)$, where we have written $F(t)$ for the left side of (34). It follows that $F'(0) \geq 0$ (provided F is differentiable at zero). We therefore set

$$\frac{d}{dt}\left(\int dx\, [(1 - t)q(x) + tr(x)] \cdot \log\frac{(1 - t)q(x) + tr(x)}{p(x)}\right)\Bigg|_{t=0} \geq 0$$

and differentiate under the integral sign. (For justification of this step and the existence of $F'(0)$, see Csiszár [12], who gives the proof in a more general measure-theoretic setting.) The result is

$$\int dx\, [r(x) - q(x)]\left[1 + \log\frac{q(x)}{p(x)}\right] \geq 0,$$

which implies

$$\int dx\, r(x) \log\frac{q(x)}{p(x)} \geq \int dx\, q(x) \log\frac{q(x)}{p(x)}$$

and therefore $H[r, p] \geq H[r, q] + H[q, p]$.

Assume I is determined by finitely many equality constraints. Since $q = p \circ I$, $\log(q(x)/p(x))$ assumes the form

$$\log\frac{q(x)}{p(x)} = -\lambda_0 - \sum_{k=1}^{m} \lambda_k f_k(x)$$

(cf. (A4)). But then

$$\int dx\, r(x) \log\frac{q(x)}{p(x)} = -\lambda_0 - \sum_{k=1}^{m} \lambda_k \bar{f}_k = \int dx\, q(x) \log\frac{q(x)}{p(x)} = H[q, p],$$

Since r and q both satisfy the equality constraints. The equality

$$\int dx\, r(x) \log \frac{r(x)}{p(x)} = \int dx\, r(x) \log \frac{r(x)}{q(x)} + \int dx\, r(x) \log \frac{q(x)}{p(x)}$$

then implies $H[r, p] = H[r, q] + H[q, p]$.

Property 11:

$$H[q^\dagger, p \circ I] \leq H[q^\dagger, p] \tag{35}$$

holds with equality if and only if $p \circ I = p$.

Discussion: This property states that the posterior $q = p \circ I$ is always closer to $q^\dagger$, in the cross-entropy sense, than is the prior p.

Proof of 11: Since $q^\dagger \in \mathcal{I}$ holds, (35) follows directly from (33) with $r = q^\dagger$.

IV. Properties Given Equality Constraints

This section concerns properties that apply when some of the new information is in the form of equality constraints (2) only. Throughout we assume a system with possible states $\boldsymbol{D}$ and an arbitrary prior $p \in \mathcal{P}$.

Property 12: Let the system have a probability density $q^\dagger \in \mathcal{P}$, and let there be information $I = (q^\dagger \in \mathcal{I})$ that is determined by a finite set of equality constraints only. Then

$$H[q^\dagger, p] = H[q^\dagger, q] + H[q, p] \tag{36}$$

holds, where $q = p \circ I$.

Discussion: This triangle equality is important for applications in which cross-entropy minimization is used for purposes of pattern classification and cluster analysis [7]. Since the difference $H[q^\dagger, p] - H[q^\dagger, q]$ is just $H[q, p]$, and since $H[q, p]$ is a measure [1] of the information divergence between q and p, Property 12 shows that $H[p \circ I, p]$ can be interpreted as the amount of information provided by I that is not inherent in p. Stated differently, $H[p \circ I, p]$ is the amount of information-theoretic distortion introduced if p is used instead of $p \circ I$. Since for any prior p and any density $r \in \mathcal{P}$ with $H(r, p) < \infty$, there exists a finite set of equality constraints I_r such that $r = p \circ I_r$ (see Appendix B), $H[r, p]$ is generally the amount of information needed to determine r when given p, or the amount of information-theoretic distortion introduced if p is used instead of r.

Proof of 12: Equation (36) follows directly from (33), since $q^\dagger \in \mathcal{I}$ holds.

Property 13: Let the system have a probability density $q^\dagger \in \mathcal{P}$, and let there be information $I_1 = (q^\dagger \in \mathcal{I}_1)$ and information $I_2 = (q^\dagger \in \mathcal{I}_2)$, where $\mathcal{I}_1, \mathcal{I}_2 \in \mathcal{P}$ are constraint sets with a nonempty intersection. Suppose that $\mathcal{I}_1$ is determined by a set of equality constraints (2) only. Then

$$(p \circ I_1) \circ (I_1 \wedge I_2) = p \circ (I_1 \wedge I_2) \tag{37}$$

and

$$H[q, p] = H[q, q_1] + H[q_1, p] \tag{38}$$

hold, where $q = p \circ (I_1 \wedge I_2)$ and $q_1 = p \circ I_1$.

Discussion: When I_1 is determined by equality constraints, (37) holds whether $(p \circ I_1) \in \mathcal{I}_2$ (compare with Property 4). Property 13 is important for applications in which constraint information arrives piecemeal, and states that intermediate posteriors can be used as priors in computing final posteriors without affecting the results. Thinking in terms of inference procedures, one might think of (37) as obvious and wonder why it does not hold for general constraints. But $p \circ I_1 \neq p$ unless $p \in \mathcal{I}_1$, so that some information about p can generally be lost on the left side of (37). From this point of view, it is somewhat surprising that (37) holds at all.

As an example of Property 13, we consider minimum cross-entropy spectral analysis [5]. If one describes a stochastic band-limited discrete-spectrum signal in terms of a probability density $q^\dagger(x) = q^\dagger(x_1, \cdots, x_n)$, where x_k is the energy at frequency f_k, known values of the autocorrelation function can be expressed as expectations of $q^\dagger$, namely,

$$R_r = \int dx \left(\sum_k 2x_k \cos(2\pi t_r f_k) \right) q^\dagger(x),$$

where R_r is the autocorrelation value at lag t_r. Let I_1 be a limited set of autocorrelations $R_1, \cdots, R_m$. Then, for a prior p_W with a flat (white) power spectrum $P_k = \int dx\, x_k p_W(x) = P$, the power spectrum of the posterior $q_{\mathrm{LPC}} = p_W \circ I_1$ is just the mth order maximum-entropy or linear predictive coding (LPC) spectrum [5]. Let I_2 be the set of autocorrelation samples $R_{m+1}, R_{m+2}, \cdots$ that together with I_1 fully determine the power spectrum of $q^\dagger$. Then (37) yields $q_F = p_W \circ (I_1 \wedge I_2) = q_{\mathrm{LPC}} \circ (I_1 \wedge I_2)$.

Proof of 13: The density q_1 has the form (A4),

$$q_1(x) = p(x) \exp\left(-\lambda_0 - \sum_{k=1}^{m} \lambda_k a_k(x) \right).$$

For an arbitrary density $q \in \mathcal{P}$, the cross-entropy with respect to q_1 satisfies

$$H[q, q_1] = \int dx\, q(x) \log \frac{q(x) \exp\left[\lambda_0 + \sum_k \lambda_k a_k(x) \right]}{p(x)} = H[q, p] + \lambda_0 + \int dx\, q(x) \sum_k \lambda_k a_k(x).$$

If q satisfies $q \in \mathcal{I}_1$, this becomes

$$H[q, q_1] = H[q, p] + \lambda_0 + \sum_k \lambda_k \bar{a}_k, \tag{39}$$

where λ_0, λ_k, and $\bar{a}_k$ are constants. Since $H[q, q_1]$ and $H[q, p]$ differ by a constant on $\mathcal{I}_1$, it follows that they have the same minima on any subset of $\mathcal{I}_1$. Since $(\mathcal{I}_1 \cap \mathcal{I}_2) \subseteq \mathcal{I}_1$ holds, this proves (37). Moreover, (39) and (A5) yield (38), which is also a special case of (33).

Property 14: Suppose there are two underlying probability densities $q_1^\dagger$ and $q_2^\dagger$. Let I_1 and I_2, respectively, stand

for the sets of equality constraints

$$\int dx\, f_i(x) q_1^\dagger(x) = F_i^{(1)}, \qquad i = 1,\cdots,m, \tag{40}$$

and

$$\int dx\, f_i(x) q_2^\dagger(x) = F_i^{(2)}, \qquad i = 1,\cdots,s, \tag{41}$$

where $s \geq m$. Then

$$(p \circ I_1) \circ (I_2) = p \circ I_2 \tag{42}$$

holds. Moreover, if $\lambda_k^{(1)}$, $\lambda_k^{(12)}$, and $\lambda_k^{(2)}$ are the Lagrangian multipliers associated with $q_1 = p \circ I_1$, $q_{12} = q_1 \circ I_2$, and $q_2 = p \circ I_2$, respectively, then

$$\lambda_k^{(2)} = \lambda_k^{(1)} + \lambda_k^{(12)}, \qquad k = 0, 1,\cdots,m, \tag{43}$$

$$\lambda_k^{(2)} = \lambda_k^{(12)}, \qquad k = m+1,\cdots,s, \tag{44}$$

and

$$H[q_2, p] = H[q_2, q_1] + H[q_1, p] + \sum_{r=1}^{m} \lambda_r^{(1)}\left(F_r^{(1)} - F_r^{(2)}\right) \tag{45}$$

also hold.

Discussion: Property 10 can apply to situations in which $q_1^\dagger$ and $q_2^\dagger$ are system probability densities at different times and in which $q_1^\dagger$ or estimates of $q_1^\dagger$ are considered to be good estimates of $q_2^\dagger$. If I_2 is determined in part by expectations of the same functions as I_1, but with different expected values, then the results of taking I_1 into account are completely wiped out by subsequently taking I_2 into account. As an example, consider frame-by-frame minimum cross-entropy spectral analysis in which I_i is determined by autocorrelation samples in frame i at a fixed set of lags ($s = m$). Equation (42) shows that the results for frame i are the same whether the assumed prior is an original prior p, the posterior from frame $i-1$, or some intermediate estimate. (However, there may be computational or bandwidth-reduction advantages to using $p \circ I_{i-1}$ as a prior in frame i.) Note that if $s \geq m$ and $F_r^{(1)} = F_r^{(2)}$ for $r = 1,\cdots,m$, Property 14 reduces to Property 13.

Proof of 14: From (A4) we have

$$q_1(x) = p(x) \exp\left(-\lambda_0^{(1)} - \sum_{k=1}^{m} \lambda_k^{(1)} a_k(x)\right),$$

where the $\lambda_k^{(1)}$ are chosen to satisfy the constraints (40). Similarly,

$$q_{12}(x) = q_1(x) \exp\left(-\lambda_0^{(12)} - \sum_{k=1}^{s} \lambda_k^{(12)} a_k(x)\right),$$

holds, This is of the form $p(x)\exp[-\lambda_0^{(2)} - \Sigma_k \lambda_k^{(2)} a_k(x)]$, with $\lambda_k^{(2)} = \lambda_k^{(1)} + \lambda_k^{(12)}$ ($k = 0,\cdots,m$) and $\lambda_k^{(2)} = \lambda_k^{(12)}$ ($k = m+1,\cdots,s$), and it is a probability density satisfying the constraints (41); it is therefore equal to $p \circ I_2 = q_2$, which proves (43), (44). Equation (45) follows from straightforward applications of (A5).

Property 15 (Expected Value Matching): Let I be the constraints

$$\int_D dx\, q^\dagger(x) f_k(x) = \bar{f}_k, \qquad k = 1,\cdots,m \tag{46}$$

for a fixed set of functions f_k, and let $q = p \circ I$ be the result of taking this information into account. Then, for an arbitrary fixed density $q^* \in \mathcal{D}$, the cross entropy $H[q^*, q] = H[q^*, p \circ I]$ has a minimum value, as the $\bar{f}_k$ vary, when the constraints (46) satisfy

$$\bar{f}_k = \bar{f}_k^* = \int_D dx\, q^*(x) f_k(x).$$

Discussion: This property states that for a density q of the general form (A4), $H[q^*, q]$ is smallest when the expectations of q match those of q^*. In particular, note that $q = p \circ I$ is not only the density that minimizes $H[q, p]$, but also is the density of the form (A4) that minimizes $H[q^\dagger, q]$! Property 15 is a generalization of a property of orthogonal polynomials [14, p. 12] that in the case of speech analysis [15, ch. 2] is called the "correlation matching property" [10].

Proof of 15: The cross-entropy $H[q^*, q]$ is given by

$$\begin{aligned} H[q^*, q] &= \int dx\, q^*(x) \log\left(q^*(x)/p(x)\right) \\ &\quad + \int dx\, q^*(x)\left(\lambda_0 + \sum_k \lambda_k f_k(x)\right) \\ &= \int dx\, q^*(x) \log\left(q^*(x)/p(x)\right) + \lambda_0 + \sum_k \lambda_k \bar{f}_k^*, \end{aligned} \tag{47}$$

where we have used (A4). Since the multipliers λ_k are functions of the expected values $\bar{f}_k$, variations in the expected values are equivalent to variations in the multipliers. Hence, to find the minimum of $H[q^*, q]$, we solve

$$\frac{\partial}{\partial \lambda_k} H[q^*, q] = 0 = \frac{\partial \lambda_0}{\partial \lambda_k} + \bar{f}_k^*,$$

where we have used (47). It follows from (A9) that the minimum occurs when $\bar{f}_k = \bar{f}_k^*$.

V. General Discussion

Property 1 and (12), (14), and (16) are the inference axioms on which the derivation in [4] is based. It is important to recognize that it is these inference properties, and not the corresponding cross-entropy properties ((13), (15), and (17)) that characterize cross-entropy minimization. For more information on this distinction, see [4, sec. VI] and [8].

An interesting aspect of the results presented in this paper is the interplay between properties of cross-entropy minimization as an inference procedure and properties of cross-entropy as an information measure. The well-known [1] and unique [8] properties of cross-entropy as an information measure in the case of arbitrary probability densi-

ties are extended and strengthened when one of the densities involved is the result of cross-entropy minimization, showing that cross-entropy minimization is optimal in a sense that has not been appreciated previously. In particular, (35) shows that $p \circ I$ is at least as close to $q^\dagger$ as is p; in the case of equality constraints, (36) shows that $H[p \circ I, p]$ is the amount of information provided by I that is not inherent in p, and Property 15 shows that $p \circ I$ is not only closer to $q^\dagger$ than is p, but it is the closest possible density of the form (A4). Indeed, the combination of these properties has led to an information-theoretic method of pattern analysis and classification [11] that is a refinement of a method due to Kullback [1, p. 83].

Acknowledgment

We thank Jacob Feldman for suggesting Property 9, and we thank Bernard O. Koopman and a referee for drawing our attention to technical issues concerning the existence of minimum cross-entropy posteriors.

Appendix A
Mathematics of Cross-Entropy Minimization

We derive the general solution for cross-entropy minimization given arbitrary constraints, and we illustrate the result with the important cases of exponential and Gaussian densities. In general, however, it is difficult or impossible to obtain a closed-form analytic solution expressed directly in terms of the known expected values rather than in terms of the Lagrangian multipliers. We therefore discuss a numerical technique for obtaining the solution, namely the Newton–Raphson method. This method is the basis for a computer program that solves for the minimum cross-entropy posterior given an arbitrary prior and arbitrary expected value constraints.

Given a positive prior density p and a finite set of equality constraints

$$\int q(x)\,dx = 1, \tag{A1}$$

$$\int f_k(x) q(x)\,dx = \bar{f}_k, \qquad k = 1,\cdots,m, \tag{A2}$$

we wish to find a density q that minimizes

$$H[q, p] = \int q(x) \log \frac{q(x)}{p(x)}\,dx,$$

subject to the constraints. For conditions that imply the existence of a unique minimum, see the discussion of Property 1 (uniqueness). One standard method for seeking the minimum is to introduce Lagrangian multipliers β and λ_k $(k = 1,\cdots,m)$ corresponding to the constraints, forming the expression

$$\int q(x) \log \frac{q(x)}{p(x)}\,dx + \beta \int q(x)\,dx + \sum_{k=1}^{m} \lambda_k \int f_k(x) q(x)\,dx,$$

and to equate the variation, with respect to q, of this quantity to zero:

$$\log \frac{q(x)}{p(x)} + 1 + \beta + \sum_{k=1}^{m} \lambda_k f_k(x) = 0. \tag{A3}$$

Solving for q leads to

$$q(x) = p(x) \exp\left(-\lambda_0 - \sum_{k=1}^{m} \lambda_k f_k(x)\right), \tag{A4}$$

where we have introduced $\lambda_0 = \beta + 1$.

In fact, the q, if it exists, that minimizes $H[q, p]$ has this form with the possible exception of a set $\boldsymbol{S}$ of points on which the constraints imply that q vanishes. (Such a situation would arise, for instance, if we had a constraint $\int q(x) f(x)\,dx = 0$, where $f(x) > 0$ when $x \in \boldsymbol{S}$ and $f(x) = 0$ when $x \notin \boldsymbol{S}$.) Informally, we could then imagine some of the Lagrangian multipliers becoming infinite in such a way that the argument of exp in (A4) becomes $-\infty$ when $x \in \boldsymbol{S}$.) Conversely, if a density q is found that is of this form and satisfies the constraints, then the minimum cross-entropy density exists and equals q [12], [1]. For simplicity in the following, we assume the set $\boldsymbol{S}$ is empty.

The cross-entropy at the minimum can be expressed in terms of the λ_k and the $\bar{f}_k$ by multiplying (A3) by $q(x)$ and integrating. The result is

$$H[q, p] = -\lambda_0 - \sum_{k=1}^{m} \lambda_k \bar{f}_k. \tag{A5}$$

It is necessary to choose λ_0 and the λ_k so that the constraints are satisfied. In the presence of the constraint (A1) we may rewrite the remaining constraints in the form

$$\int \left(f_k(x) - \bar{f}_k\right) q(x)\,dx = 0. \tag{A6}$$

If we find values for the λ_k such that

$$\int \left(f_i(x) - \bar{f}_i\right) p(x) \exp\left(-\sum_{k=1}^{m} \lambda_k f_k(x)\right) dx = 0, \qquad i = 1,\cdots,m, \tag{A7}$$

we are assured of satisfying (A6); and we can then satisfy (A1) by setting

$$\lambda_0 = \log \int p(x) \exp\left(-\sum_{k=1}^{m} \lambda_k f_k(x)\right) dx. \tag{A8}$$

If the integral in (A8) can be performed, one can sometimes find values for the λ_k from the relations

$$-\frac{\partial}{\partial \lambda_k} \lambda_0 = \bar{f}_k. \tag{A9}$$

The situation for inequality constraints is only slightly more complicated. Suppose we replace all the equal signs in (A2) by $\leq$. (We lose no generality thereby; we can change inequalities with $\geq$ into inequalities with $\leq$ by changing the signs of the corresponding f_k and $\bar{f}_k$, and any equality constraint is equivalent to a pair of inequality constraints.) The q that minimizes $H(q, p)$ subject to the resulting constraints will in general satisfy equality for certain values of k in the modified (A2), while strict inequality will hold for the rest. We can still use the solution (A4), subjecting the Lagrange multipliers to the conditions $\lambda_k \leq 0$ for k such that equality holds in the constraint, and $\lambda_k = 0$ for k such that strict inequality holds in the constraint.

It unfortunately is usually impossible to solve (A7) or (A9) for the λ_k explicitly, in closed form; however, it is possible in certain important special cases. For example, consider the case in which the prior $p(x)$ is a multivariate exponential,

$$p(x) = \prod_{k=1}^{n} (1/a_k) \exp[-x_k/a_k], \tag{A10}$$

where $x = (x_1, \cdots, x_n)$ and the x_k each range over the positive real line, and in which the constraints are

$$\int dx\, x_k q(x) = \bar{x}_k, \tag{A11}$$

$k = 1, \cdots, n$. Solving (A9) in order to express the minimum cross-entropy posterior directly in terms of the known expected values $\bar{x}_k$ yields

$$q(x) = \prod_k (1/\bar{x}_k) \exp[-x_k/\bar{x}_k]. \tag{A12}$$

Thus, the density remains multivariate exponential, with the prior mean values a_k being replaced by the newly learned values $\bar{x}_k$.

Now consider the case in which the x_k range over the entire real line, and in which the prior density is Gaussian,

$$p(x) = \prod_k (2\pi b_k)^{-1/2} \exp\left[-(x_k - a_k)^2/2b_k\right].$$

Suppose that the constraints are (A11) and

$$\int dx\, (x_k - \bar{x}_k)^2 q(x) = v_k.$$

In this case the minimum cross-entropy posterior is

$$q(x) = \prod_k (2\pi v_k)^{-1/2} \exp\left[-(x_k - \bar{x}_k)^2/2v_k\right].$$

Thus, the density remains multivariate Gaussian, with the prior means and variances being replaced by the newly learned values.

Here is an example of a simple problem for which the solution of (A7) cannot be expressed in closed form. Consider a discrete system with n states x_j and prior probabilities $p(x_j) = p_j$ ($j = 1, \cdots, n$). The discrete form of (A1) is

$$\sum_{j=1}^{n} q_j = 1, \tag{A13}$$

where $q_j = q(x_j)$. Suppose the only other constraint is that the mean m of the indices j is prescribed: $f(x_j) = j$, and

$$\sum_{j=1}^{n} j q_j = m. \tag{A14}$$

Then (A4) becomes $q_j = p_j \exp[-\lambda_0 - \lambda j]$, which we write as $q_j = a p_j z^j$ by introducing the abbreviations $a = \exp[-\lambda_0]$ and $z = \exp[-\lambda]$. From (A16) and (A17) we then obtain

$$a = \left(\sum_{j=1}^{n} p_j z^j \right)^{-1}$$

and

$$\sum_{j=1}^{n} (j - m) p_j z^j = 0. \tag{A15}$$

The problem then reduces to finding a positive root of the polynomial in (A15). As in the continuous case, there are special forms for the prior that lead to important particular solutions. But when $n > 5$, the roots of the polynomial (other than zero) cannot in general be written as explicit closed-form expressions in the coefficients for arbitrary priors. Numerical methods of solution therefore become important. Our obtaining a polynomial equation in the present example was an accidental consequence of the fact that the values of the constraint function f formed a subset of an arithmetic progression ($j = 1, 2, \cdots$). Thus, for more general types of problems, numerical methods are even more important.

One such method is the Newton–Raphson method, which is for finding solutions for systems of equations that, like (A7), are of the form

$$F_i(\lambda_1, \cdots, \lambda_m) = 0, \qquad i = 1, \cdots, m. \tag{A16}$$

The method starts with an initial guess at the solution, $\boldsymbol{\lambda}^{(1)} = (\lambda_1^{(1)}, \cdots, \lambda_m^{(1)})$, and produces further approximate solutions $\boldsymbol{\lambda}^{(2)}, \boldsymbol{\lambda}^{(3)}, \cdots$ in succession. If the initial guess $\boldsymbol{\lambda}^{(1)}$ is close enough to a solution of (A16), if the F_i are continuously differentiable, and if the Jacobian $[\partial F_i/\partial \lambda_j]$ is nonsingular, then the $\boldsymbol{\lambda}^{(r)}$ will converge to the solution in the limit as $r \to \infty$.

The method is based on the fact that, for small changes $\Delta\boldsymbol{\lambda}^{(r)}$ in the arguments $\boldsymbol{\lambda}^{(r)}$, we have the approximate equality

$$F_i(\boldsymbol{\lambda}^{(r)} + \Delta\boldsymbol{\lambda}^{(r)}) \cong F_i(\boldsymbol{\lambda}^{(r)}) + \sum_{k=1}^{m} \frac{\partial F_i(\boldsymbol{\lambda}^{(r)})}{\partial \lambda_k^{(r)}} \Delta\lambda_k^{(r)}$$

up to a term of order $o(\Delta\boldsymbol{\lambda}^{(r)})$. We therefore take $\Delta\boldsymbol{\lambda}^{(r)}$ to be a solution of the linear equation

$$\sum_{k=1}^{m} \frac{\partial F_i(\boldsymbol{\lambda}^{(r)})}{\partial \lambda_k^{(r)}} \Delta\lambda_k^{(r)} = -F_i(\boldsymbol{\lambda}^{(r)}) \tag{A17}$$

and set $\boldsymbol{\lambda}^{(r+1)} = \boldsymbol{\lambda}^{(r)} + \Delta\boldsymbol{\lambda}^{(r)}$. In applying the Newton–Raphson method to cross-entropy minimization, we let $F_i(\boldsymbol{\lambda})$ be proportional to the discrete form of the left side of (A7); we set

$$F_i(\boldsymbol{\lambda}^{(r)}) = \sum_{j=1}^{n} f_{ij} p_j \exp\left(- \sum_{u=1}^{m} \lambda_u^{(r)} f_{uj} \right), \tag{A18}$$

$$\frac{\partial F_i(\boldsymbol{\lambda}^{(r)})}{\partial \lambda_k} = - \sum_{j=1}^{n} f_{ij} f_{kj} p_j \exp\left(- \sum_{u=1}^{m} \lambda_u^{(r)} f_{uj} \right), \tag{A19}$$

where $f_{ij} = f_i(x_j) - \bar{f}_i$, and we have removed a factor of $\exp[-\sum_u \lambda_u^{(r)} f_u]$. With the abbreviation

$$g_j = p_j^{1/2} \exp\left(-\frac{1}{2} \sum_{u=1}^{m} \lambda_u^{(r)} f_{uj} \right),$$

we express the right sides of (A18) and (A19) in matrix notation as $[\boldsymbol{f}\,\mathrm{diag}(\boldsymbol{g})\boldsymbol{g}]_i$ and $[\boldsymbol{f}\,\mathrm{diag}(\boldsymbol{g})^2\boldsymbol{f}^t]_{ik}$, respectively, where $\mathrm{diag}(\boldsymbol{g})$ is the diagonal matrix whose diagonal elements are the g_j, and $\boldsymbol{f}^t$ is the transpose of $\boldsymbol{f}$. The solution of (A17) is then given by

$$\Delta\boldsymbol{\lambda}^{(r)} = \left[\left(\boldsymbol{f}\,\mathrm{diag}(\boldsymbol{g})^2 \boldsymbol{f}^t \right)^{-1} \boldsymbol{f}\,\mathrm{diag}(\boldsymbol{g}) \right] \boldsymbol{g}.$$

We remark that the quantity in brackets is the Moore–Penrose generalized inverse [16] of the matrix $\mathrm{diag}(\boldsymbol{g})\boldsymbol{f}^t$. The approach just described has been made the basis for a computer program [17], written in APL, for solving cross-entropy minimization problems with arbitrary positive discrete priors p and equality constraints specified by matrices $\boldsymbol{f}$. The approach is particularly convenient for programming in APL since the generalized inverse is a built-in APL primitive function [18]. To solve a minimum cross-entropy problem with 500 states and 10 constraints, the program typically requires 15 seconds of central processing unit (CPU) time when running under the APL SF interpreter on a DEC-10 system with a KI central processor.

Gokhale and Kullback [19] describe a somewhat different algorithm, also based on the Newton–Raphson method, that has been implemented in *PL/I*. Agmon, Alhassid, and Levine [20], [21] describe yet another cross-entropy minimization algorithm and a Fortran implementation. Tribus [13] presents programs in Basic that compute singly and doubly truncated Gaussian distributions as maximum entropy distributions with prescribed means and variances.

APPENDIX B
REMARK ON THE DISCUSSION OF PROPERTY 12

In the discussion of Property 12, it was stated that for any prior p and any density $r \in \mathcal{D}$ with $H(r, p) < \infty$, there exists a finite set of equality constraints I_r such that $r = p \circ I_r$. In fact, at most two are needed. Let

$$f_1(x) = \begin{cases} 0, & r(x) \neq 0 \\ 1, & r(x) = 0, \end{cases}$$

$$\bar{f}_1 = 0,$$

$$f_2(x) = \begin{cases} \log\left(p(x)/r(x)\right), & r(x) \neq 0 \\ 0, & r(x) = 0, \end{cases}$$

$$\bar{f}_2 = -H(r, p),$$

and impose constraints

$$\int q(x) f_1(x)\, dx = \bar{f}_1, \tag{B1}$$

$$\int q(x) f_2(x)\, dx = \bar{f}_2. \tag{B2}$$

The first constraint implies $(p \circ I)(x) = 0$ where $r(x) = 0$. On the complementary set, where $r(x) \neq 0$, define $q(x)$ by (A4) will all $\lambda_j = 0$ except $\lambda_2 = 1$; this gives a function q that satisfies the second constraint as well as the first and also agrees with r. Hence $r = q$ is the result of minimizing $H(q, p)$ with respect to (B1) and (B2).

REFERENCES

[1] S. Kullback, *Information Theory and Statistics*. New York; Wiley, 1959.

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

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

[4] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 26–37, Jan. 1980.

[5] J. E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acous. Speech Signal Processing*, vol. ASSP-29, pp. 230–237, Apr. 1981.

[6] R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory*, to be published.

[7] J. E. Shore and R. M. Gray, "Minimum cross-entropy pattern classification and cluster analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* to be published.

[8] 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.

[9] F. Itakura and S. Saito, "Analysis synthesis telephone based upon maximum likelihood method," *Reports of the 6th Int. Cong. Acoustics*, Y. Yonasi, ed. Tokyo, 1968.

[10] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-28, pp. 367–376, 1980.

[11] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-28, pp. 562–574, Oct. 1980.

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

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

[14] L. Geronimus, *Orthogonal Polynomials*. New York: Consultants Bureau, 1961.

[15] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[16] A. E. Albert, *Regression and the Moore–Penrose Pseudoinverse*. New York: Academic, 1972.

[17] R. W. Johnson, "Determining probability distributions by maximum entropy and minimum cross-entropy," *Proc. APL79*, 24–29.

[18] M. A. Jenkins, "The solution of linear systems of equations and linear least squares problems in APL," Scientific Center Tech. Rep. No. 320-2989, IBM, NY, June 1970.

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

[20] Y. Alhassid, N. Agmon, and R. D. Levine, "An upper bound for the entropy and its applications to the maximal entropy problem," *Chem. Phys. Lett.*, vol. 53, no. 1, pp. 22–26, 1978.

[21] N. Agmon, Y. Alhassid, and R. D. Levine, "An algorithm for finding the distribution of maximal entropy," *J. Comput. Phys.*, vol. 30, pp. 250–258, 1979.