Comments on and Correction to "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy"

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

Abstract—**An error in the subject paper is pointed out: when the axioms given there are restricted to the discrete case, they do not imply the discrete case of the principle of minimum cross-entropy. The principle is shown to follow, however, from the adoption of an additional axiom: if new information is consistent with a prior estimate of a probability distribution, then the posterior estimate equals the prior. Minor other improvements and corrections to the arguments in the paper are made.**

I. Introduction

In Section IV of the above paper,[1] we showed that cross-entropy minimization for the case of continuous probability densities follows from the four axioms given in Section III and summarized here.

I. *Uniqueness*:

$$q = p \circ I \text{ is unique.}$$

II. *Invariance*:

$$(\Gamma p) \circ (\Gamma I) = \Gamma(p \circ I).$$

III. *System Independence*:

$$(p_1 p_2)(I_1 \wedge I_2) = (p_1 \circ I_1)(p_2 \circ I_2).$$

IV. *Subset Independence*:

$$(p \circ (I \wedge M)) * S_i = (p * S_i) \circ I_i.$$

In Section V, we considered the discrete case. In Section V-A we argued that the derivation for the continuous case also applied to the discrete case. In Section V-B, we showed that entropy maximization follows in the discrete case from a version of the axioms that does not include a prior (eq. (44)).

The argument in Section V-A is wrong—cross-entropy minimization does not follow from the four axioms in the discrete case. In particular, for discrete probability distributions, the invariance axiom reduces to the special case of permutation invariance, which is insufficient for proving Theorem II (that H is equivalent to the form $\int dx\, q(x) g(q(x)/p(x))$, where g depends only on the ratio q/p). Indeed, there is nothing in the axioms that forces the functional H to depend on the prior p in the discrete case. To see this another way, note that entropy maximization satisfies the discrete form of the axioms—the prior is just ignored. Entropy maximization does not, however, satisfy Property 2 of [1],

$$p \circ I = p \qquad \text{if and only if } p \in \mathscr{I}, \tag{2}$$

which just states that the posterior estimate of $q^\dagger$ should be the same as the prior estimate p if the new information does not contradict the prior in any way. This property, which is one way of expressing the sense in which we use the term *prior*, is clearly desirable for an inference procedure. Here we adopt a weak form of (2) as an axiom for the discrete case. We use the notation and definitions of the paper.[1]

[1] J. E. Shore and R. W. Johnson, *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 26–37, Jan. 1980.

Manuscript received June 9, 1983.

The authors are with the Computer Science and Systems Branch, Information Technology Division, Naval Research Laboratory, Washington, DC 20375.

II. Minimum Cross-Entropy in the Discrete Case

Here is the new axiom.

Axiom V: Let I be the "null" constraint $I = (q^\dagger \in \mathscr{D})$, which is satisfied by any density $q \in \mathscr{D}$. Then, for any prior $p \in \mathscr{D}$,

$$p \circ I = p \tag{3}$$

holds.

Justification: In the absence of new information, we should not change our prior estimate.

In the discrete case, we have $\boldsymbol{q} = q_1, q_2, \cdots, q_n$, $\boldsymbol{p} = p_1, p_2, \cdots, p_n$, and the following version of Theorem II.

Theorem IIa: Let $H(\boldsymbol{q}, \boldsymbol{p})$ satisfy uniqueness, subset independence, invariance, and (3). Then H is equivalent to a function of the form

$$F(\boldsymbol{q}, \boldsymbol{p}) = \sum_{j=1}^{n} q_i h(q_j/p_j), \tag{4}$$

for some function h.

Proof: From Theorem I, H is equivalent to a function with the sum form

$$F(\boldsymbol{q}, \boldsymbol{p}) = \sum_{j=1}^{n} f(q_j, p_j), \tag{5}$$

for some function f. We begin by defining

$$u(x, y) = \frac{\partial f(x, y)}{\partial x} \tag{6}$$

and showing that $u(x, y)$ depends only on the ratio x/y.

We invoke subset independence in the case of null subset constraints $I_i = (\boldsymbol{q}^\dagger * \boldsymbol{S}_i \in \mathscr{S}_i)$. (See the definitions in Section III of the paper.[1]) Since $I \wedge M = M$, subset independence reduces in this case to

$$(\boldsymbol{p} \circ M) * \boldsymbol{S}_i = (\boldsymbol{p} * \boldsymbol{S}_i) \circ I_i.$$

Applying the new axiom (3) to the right-hand side yields

$$(\boldsymbol{p} \circ M) * \boldsymbol{S}_i = \boldsymbol{p} * \boldsymbol{S}_i. \tag{7}$$

Note that this is just a special case of subset aggregation—Property 9 in [1]. Let $\boldsymbol{q} = \boldsymbol{p} \circ M$. Then

$$\frac{q_j}{\sum_{k \in \boldsymbol{S}_i} q_k} = \frac{p_j}{\sum_{k \in \boldsymbol{S}_i} p_k}$$

holds when $j \in \boldsymbol{S}_i$. The ratio q_j/p_j is constant on each $\boldsymbol{S}_i$:

$$\frac{q_j}{p_j} = \frac{\sum_{k \in \boldsymbol{S}_i} q_k}{\sum_{k \in \boldsymbol{S}_i} p_k} = \frac{m_i}{\sum_{k \in \boldsymbol{S}_i} p_k}.$$

Now, the condition for a constrained minimum of F yields

$$\frac{\partial F(\boldsymbol{q}, \boldsymbol{p})}{\partial q_j} = \frac{\partial f(q_j, p_j)}{\partial q_j} = u(q_j, p_j) = -\lambda - \alpha m_i,$$

for some Lagrange multipliers λ, α. Since the right side is independent of j, it follows that $u(q_j, p_j)$ is constant on each $\boldsymbol{S}_i$.

At this point we know that q_j/p_j and $u(q_j, p_j)$ are constants for $j \in \boldsymbol{S}_i$. We can always arrange the numbering so that $1, 2 \in \boldsymbol{S}_1$. Then

$$\frac{q_1}{q_2} = \frac{p_1}{p_2}$$

and

$$u(q_1, p_1) = u(q_2, p_2)$$

both hold. We now show that $u(x, y) = u(x', y')$ for any positive numbers x, y, x', y' less than 1 and with equal ratios $x/y = x'/y'$. We choose positive numbers x'' and y'' that have the same ratio

$$\frac{x''}{y''} = \frac{x}{y} = \frac{x'}{y'}$$

and are so small that

$$x'' < 1 - x, \quad y'' < 1 - y,$$
$$x'' < 1 - x', \quad y'' < 1 - y'.$$

We can then construct $\boldsymbol{p}$ and choose m_1 so that

$$q_1 = x, \quad p_1 = y,$$
$$q_2 = x'', \quad p_2 = y''.$$

We find $u(x, y) = u(x'', y'')$. Similarly, with different choices for $\boldsymbol{p}$ and m_1, we find $u(x', y') = u(x'', y'')$. It follows that u depends only on the ratio of its arguments: $u(x, y) = u(x/y)$. Equation (6) therefore has the general solution $f(x, y) = xh(x/y) + v(y)$. Substitution of this solution into (5) yields

$$F(\boldsymbol{q}, \boldsymbol{p}) = \sum_{j=1}^{n} q_j h(q_j/p_j) + \sum_{j=1}^{n} v(p_j).$$

Since the second term depends only on the fixed prior, it cannot affect the minimization of F and can be dropped. This completes the proof of Theorem IIa.

In the foregoing proof of Theorem IIa for the discrete case, we used the new axiom (3) and subset independence to derive (7), a special case of subset aggregation. As one might expect, subset aggregation can be adopted as an axiom instead of (3). We chose (3) because it expresses a weaker property than subset aggregation, because its role in forcing the posterior to depend on the prior is intuitively clear, and because its justification is compelling.

III. Other Corrections and Comments

In the proof of Theorem III (on page 31) of the paper,[1] (36) can be obtained more simply without (30) and (35)—instead of differentiating (35) with respect to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, we differentiate (34) after substituting $r_1 r_2$ for r.

Similarly, in the derivation of maximum entropy in Section V-B, we can use

$$u(q_i r_k) = -\alpha' a_i - \beta' b_k - \lambda'$$

instead of the more complicated equation at the bottom of the left column on page 33.

An error occurs on page 33 of the paper,[1] in the text just before (48). We state that the integration of $u(x) = A \log(x) + B$ yields $f(x) = Ax \log(x) + Bx - A$, where $u(x) = f'(x)$. This is wrong; the correct result is $f(x) = Ax \log(x) + (B - A)x + C$. Fortunately, no conclusions are invalidated, since the constant terms in (48) do not affect the minimization of H.

References

[1] J. E. Shore and R. W. Johnson, "Properties of cross-entropy minimization," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 472–482, July 1981.

Dissertation Abstracts

Alex Netch, "Pairwise Phase-Locked Loop Coupling for Tracking Improvement in Nonrigid Retrodirective Arrays," Ph.D., Dept. of Electrical Engineering, University of Southern California, Los Angeles, CA, March 1983. Advisor: Robert A. Scholtz.

The study addressed one aspect of adaptive array antennas, namely, the ability of a retrodirective array (RDA) to form a beam in the direction of a pilot signal, using the phase conjugacy principle. Methods of pairwise phase-locked loop (PLL) coupling for tracking improvement in structurally static and dynamically distorting arrays were developed and analyzed.

Two new types of PLL tracking networks were developed, based on the maximum likelihood principle. The delta coupling method was designed to track electrical phase angle differential increments. The geometry coupled approach exploited assumed differences in the rates of array distortion and pilot signal directional motion by tracking the changing array geometry and constructing the electrical phase from the estimated geometric variable. Both tracking networks were realized with generic pairwise coupling modules, common in form to all coupling modes, and resulting in networks having a binary tree structure.

Practical acquisition and ambiguity controlling algorithms were developed. To resolve stable ambiguous lock points and speed up acquisitions, a combination of restricted search and sequential acquisition was proposed. The pairwise coupled modular design was expanded to include the acquisition logic.

A linearized analysis was used to evaluate the reduction in estimated phase noise, achieved through coupling, and the increased in transient decay time constants. A significant improvement in performance was observed if array distortion varied slowly, relative to the dynamics of the carrier phase. The linearized analysis was verified with digital Monte-Carlo simulation for a small array (8 elements) and high pilot signal-to-noise ratio (SNR = 14 dB).

Typical simulated antenna patterns were illustrated for large geometrically coupled arrays (64 elements) and compared to simulated uncoupled RDA antenna patterns for an uncoupled PLL SNR of 4 dB (0.61 rads^2 phase variance). The comparison showed a significant coupled beamforming improvement.

Sabah A. Al-Quaddoomi, "Two-Dimensional Binary Codes with Good Autocorrelation," Ph.D., Dept. of Electrical Engineering, University of Southern California, Los Angeles, CA, December 1982. Advisor: Robert A. Scholtz.

In this dissertation the problem of finding two-dimensional binary arrays which possess optimal doubly aperiodic autocorrelation properties or aperiodic-periodic autocorrelation properties, is investigated. Optimal arrays with up to four rows and seven columns are generated by means of an exhaustive computer search. Novel shortcuts are implemented in the search program, including special storage techniques and backtracking, in order to reduce CPU execution time. Gold and Kasami codes of length seven and fifteen respectively are implemented to generate four-by-seven and four-by-fifteen arrays having optimal aperiodic-periodic autocorrelation.