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:

\[ q = p \star I \] is unique.

II. Invariance:

\[ (I \star I) \star (I \star I) = (I \star I) \star (I \star I) \star I \star I \]

III. System Independence:

\[ (p \star I_1 \star I_2) = (p \star I_1 \star I_2) \star I \]

IV. Subset Independence:

\[ (p \star (I \star M)) \star S_i = (p \star S_i) \star I \star I \]

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 a function of the form)

\[ F(q, p) = \sum_{j=1}^{n} q_j h(q_j/p_j) \]

for some function h.

Proof: From Theorem I, H follows in the discrete case from a version of the axioms in this case to

\[ (p \star M) \star S_i = (p \star S_i) \star I \star I \]

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

\[ (p \star M) \star S_i = (p \star S_i) \star I \star I \]

Note that this is just a special case of subset aggregation—Property 9 in [1]. Let \( q = p \star M \). Then

\[ \sum_{k \in S_i} q_k = \sum_{k \in S_i} p_k \]

holds when \( j \in S_i \). The ratio \( q_j/p_j \) is constant on each \( S_i \):

\[ \frac{\sum_{k \in S_i} q_k}{p_j} = \frac{\sum_{k \in S_i} p_k}{m_i} \]

Now, the condition for a constrained minimum of F yields

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

for some Lagrange multipliers \( \lambda, \alpha \). Since the right side is independent of \( j \), it follows that \( u(q_j, p_j) \) is constant on each \( S_i \).

At this point we know that \( q_j/p_j \) and \( u(q_j, p_j) \) are constants for \( j \in S_i \). We can always arrange the numbering so that \( 1,2 \in S_1 \). Then

\[ q_1 = p_1 \]
\[ q_2 = p_2 \]

2 Manuscript received June 9, 1983.
3 The authors are with the Computer Science and Systems Branch, Information Technology Division, Naval Research Laboratory, Washington, DC 20375.
4 U.S. Government work not protected by U.S. copyright.
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
\[ x'' = x/y = y'' \]
and are so small that
\[ x'' < 1 - x, \quad y'' < 1 - y, \]
\[ x'' < 1 - x', \quad y'' < 1 - y'. \]
We can then construct \( p \) and choose \( m \) 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 \( p \) and \( m \), 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
\[ F(q, p) = \sum_{j=1}^{n} q_j h(q_j/p_j) + \sum_{j=1}^{n} v(p_j). \]
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.

### References

In the proof of Theorem III on page 31 of the paper, \( f(x) = Ax\log + Bx - A \) is supposed to be minimized. Fortunately, no conclusions are invalidated, since the constant terms in (48) do not affect the minimization of \( H \).

### References