Maximum Entropy Aggregation of Individual Opinions

William B Levy, Member, IEEE, and Hakan Deliç

Abstract—This paper presents several formulations for aggregating opinions about the outcome of a future event—e.g., opinions in the form of probabilities are aggregated into a single probability. The approach used to derive these aggregation formulas is maximum entropy inference, with the assumption that the opinions and the event being predicted are joint random variables. As part of the presentation, the novel, axiomatically developed aggregation formula due to Bordley is used as a foil for contrasting and comparing the results developed here. It is shown that Bordley's result is overly restrictive when looked at in terms of maximum entropy derivations that are clearly using more information.

Initially, we suppose that the opinions given on the future event are in the form of odds; then we consider the case where these opinions are in the form of probabilities; and finally, we reach a general formulation in which they could be odds, probabilities or even binary, yes/no, predictions of a future event with interactions between experts. In regard to this last point, the use of Darroch and Ratcliff's method in maximum entropy aggregation is sketched.

I. INTRODUCTION

It is often desirable to aggregate a set of opinions into a single opinion (see, e.g., [1], [2], [8], [15]) that is, in some defined sense, best. One such aggregation problem starts with a group of experts giving the probability of the outcome of a future event. Then, the aggregation problem is to produce the single probability value that is the best group prediction of this future event. Bordley [1] axiomatically develops one particular aggregation formula. We shall, with different axiomatic underpinnings, consider this same problem when there is access to the prediction statistics of these experts (where for our purposes anyone willing to give opinions qualifies as an expert).

The approach of this article is to examine maximum entropy (ME) inferred aggregation formulas by comparing and contrasting them to Bordley's axiomatically justified method for aggregating opinions. Maximum entropy inference, which assumes equal priors, is a special case of minimum relative entropy inference [12], [13]. The latter inference method is an axiomatically derived procedure for inferring probability density functions, and here we will apply this procedure,

through ME formalism, to the aggregation of quantitative opinions. Relative entropy (also called directed divergence, cross entropy, etc.) has been axiomatically derived numerous times based on desirable mathematical properties (see, e.g., [11]). The axiomatic derivation of Shore and Johnson [12] is, however, special. They derive relative entropy as the uniquely correct measure for inference from statistics. Their derivation is not based on desirable mathematical properties of the measure, but rather, it is based on axioms that are clearly desirable properties of an inference procedure that infers a probability density function from statistical averages. Csiszar [3] has recently generalized and clarified the axiomatic bases of such inference procedures.

To use the ME inference, we need to assume that the expert opinions and the events in question are joint random variables that have been sampled enough to create reliable statistics on the opinions of these experts. These aggregation formulas are unique for any particular constraint set that meets the requirements of the ME inference procedure. Different constraint sets, however, result in different formulations.

The organization of the paper is as follows. In Section II, we introduce Bordley's axiomatic aggregation formula, and in Section III, we establish links between the ME inference procedure and the opinion aggregation problem. Sections IV and V present maximum entropy formulas for aggregating odds and probability assessments, respectively. It is also shown in Section V that, under certain restrictive assumptions, maximum entropy based aggregation formulas agree with Bordley's results. Discussions on aggregation formulas with more information and the case where experts interact can be found in Sections VI and VII, respectively. Some computational issues are addressed in Section VIII, and conclusions are presented in Section IX.

II. BORDLEY'S AXIOMATIC AGGREGATION

From an explicit set of axioms, Bordley [1] develops a formulation for aggregating the individual probability assessments of many experts. Each expert $i (i \in \{1, 2, \ldots, n\})$ predicts the probability, $P_{i}(Z = z) \equiv P_{i}$, of a single, binary event $Z$, and from these individual predictions the aggregation formula produces a single probability, $P_{O}(Z = z) \equiv P_{O}$. The general aggregation formula from the derivation is

$$P_{O} = \frac{P_{0} \prod_{i=1}^{n} (P_{i}/P_{0})^{w_{i}}}{P_{0} \prod_{i=1}^{n} (P_{i}/P_{0})^{w_{i}} + (1 - P_{0}) \prod_{i=1}^{n} ((1 - P_{i})/(1 - P_{0}))^{w_{i}}} \tag{1}$$

The authors are with the Department of Neurological Surgery, Health Sciences Center Box 420, University of Virginia, Charlottesville, Virginia 22908, USA.

IEEE Log Number 9400450.
where the exponents \( w_i \) are weighting coefficients that need to be determined, the multiple products are taken over all the participating individuals \( i \), and \( P_0 \) is a prior probability held in common by all experts.

A specialized version presented by Bordley [1] assumes \( \sum w_i = 1 \) or \( P_0 = 0.5 \) so that the aggregation formula in this case is

\[
P_G = \frac{\prod_{i=1}^{n} P_i^{w_i}}{\prod_{i=1}^{n} P_i^{w_i} + \prod_{i=1}^{n} (1 - P_i)^{w_i}}.
\]

(2)

Although there does not appear to be an ME inferred formulation corresponding to the more general (1), there is a corresponding ME inferred formulation corresponding to (2) when a rather unlikely assumption is made about the statistics of the environment.

Regardless of which equation is to be used, there must be a way of placing a value on the \( w_i \). Although discussed in Bordley’s presentation, the way to actually value each weighting \( w_i \) is not resolved. In contrast, ME inferred formulations prescribe a way to value each of the weights, and, as we shall see, these values are often rather simple to obtain.

The aggregation of probabilities, (1), is first developed for an odds expression that is converted into a probability. The formulation developed for the group odds, \( O_G \), as a function of each experts odds assessment, \( O_i \), is

\[
O_G = \left( \frac{P_0}{1 - P_0} \right)^{1 - \sum w_i} \prod_{i=1}^{n} O_i^{w_i}.
\]

(3)

This aggregated odds assessment is then converted into a probability by the usual relationship for any random variable \( Z \) and particular outcome \( z : P(Z = z) = O(Z = z)/(1 + O(Z = z)) \) where \( O(Z = z) = P(Z = z)/(1 - P(Z = z)) \).

To develop (3), Bordley [1] starts with some decision theoretic axioms of additive conjoint measure (ACM) theory [5], [6], [7]. By themselves, the axioms of additive conjoint measure theory imply the generic aggregation

\[
O_G = F \left( \sum_{i=1}^{n} u_i(O_i) \right)
\]

(4)

where \( u_i \) and \( F \) are strictly monotonic and continuous functions. As we shall see, it will be easy to identify with (4) ME inferred aggregations of the same form.

The implications of the ACM theory axioms were further developed by the addition of two novel axioms. These two axioms, called the normalization axiom and the weak likelihood principle, are presented and discussed, to the extent of their original presentation, in Appendix I. We shall see that it is the weak likelihood principle, in particular, that leads to a formulation that need not agree with any ME inferred formulation. We shall also see that agreement is possible in the case of (2) but only under rather unlikely circumstances. Moreover, in taking the statistical perspective of ME inference, we are led to conclude that (1) and (2) unduly limit the information brought to bear on the aggregation problem.

III. MAXIMUM ENTROPY INFERENCE

Of critical importance is the recognition that ME inference produces a unique probability density function from a set of moment, i.e., expectation, constraints. As such, ME inference is a mapping that takes expectations of some random (possibly a vector) variable \( X \), or expectations of functions of \( X \), into a particular probability density function of \( X \). Consider a set of expectations, \( E[\cdot] \), over arbitrary functions \( f_h(X) \) that are not themselves expectations or inferable, as expectations, one from the others. Specify the set of available expectations as \( M = \{ E[f_h(X)] : h = 1, 2, \ldots, m \} \). Then, as given by Jaynes [10],

\[
ME: M \rightarrow P^*(X) = \exp \left\{ -\lambda_0 - \sum_{h=1}^{m} \lambda_h f_h(X) \right\}
\]

(5)

where \( P^*(X) \) is a density function dependent on the constraint set \( M \) and \( \int_{(X,P^*(X)>0)} dX P^*(X) = 1 \). The values in the constraint set and the last condition imply the values of the variables \( \lambda_0 \) through \( \lambda_m \). This inferred density function, \( P^* \), is the unique density that fits all the information in the constraint set \( M \) while implying no more information than what is in this constraint set. Considering \( P^* \) as the result, any other density function either fails to satisfy the statistics in the constraint set or implies a larger set of moment constraints than are actually in (or implied by) the given constraint set. Note, further, that the support intervals of the functions \( f_h(X) \) are implicit constraints, expressing themselves as integration limits when solving for Lagrange multipliers or marginal densities.

To apply ME inference to the aggregation problem, it is first necessary to identify the random variable(s) and their space of allowed values. It just so happens that in the aggregation problem the random variables of interest might be, in part, the probabilities given by the experts [15], and although a probability of a probability can lead to confusion, this concept causes no problems for the theory so long as we keep straight which probability is being inferred and which probabilities are the random variables.

Specifically, we only need consider each individual’s forecast, \( P_i \), joined with the outcome of the predicted event \( Z \) as a random variable. Then we can collect expectations over the set of experts, e.g., \( \{ E[P_i Z], \forall \} \) or, e.g., \( \{ E[Z \log P_i], \forall \} \). ME inference then creates a probability density function over the space of the \( P_i \) and the event in question, e.g., \( P^*(P_1 P_2 \ldots P_k Z) \). Margination and application of Bayes’s ideas produce the desired group prediction, i.e., the probability of \( Z \), conditioned on a relevant set of values of individual opinions, \( P^*(Z|P_1 \ldots P_k) \). As the examples developed below illustrate, different functions in the constraint set produce different formulations.

IV. AGGREGATING AN ODDS ASSESSMENT BY ME INFERENCE

In the first ME development, we start with individual odds \( O_i \), a function of the \( P_i \), instead of the \( P_i \) themselves because the odds variable is the beginning of and apparently central to, Bordley’s development.

Suppose there is a group of individuals \( i = 1, 2, \ldots, n \), e.g., weather forecasters or better yet bookmakers, who regularly
place odds, \( O_i \in [0, \infty) \), on a binary event \( Z \in \{a, b : 0 < a < b < \infty\} \). (Throughout the rest of the paper, \( O_i \) will denote \( O_i(Z = z) \).) We have access to each set of odds and outcome, and keep score for many trials so that we can expect sample averages to be very close to the population averages that are required for the constraint set. Specifically, the constraint set of this first development is \( \{E[O_i Z], \forall i, E[Z]\} \). As per (5), from this constraint set ME inference yields

\[
P^*(Q, Z) = \exp \left\{ -\lambda_0 - \lambda_i Z - \sum_{i=1}^{n} \lambda_i O_i Z \right\}
\]

where \( Q \) is the vector of individual odds \( O_i \).

If we can relate the expectations that make up the constraint set to the lambdas of (6), then we can explicitly calculate aggregate odds \( O^*(Z = a|Q) \) via the relationships

\[
P^*(Q) = P^*(Q, Z = a) + P^*(Q, Z = b) = \exp \left\{ -\lambda_0 - \lambda_a a - a \Sigma_i O_i \right\} + \exp \left\{ -\lambda_0 - \lambda_b b - b \Sigma_i O_i \right\} \quad (7)
\]

\[
P^*(Z = a|Q) = P^*(Q, Z = a)/P^*(Q) = 1/(1 + \exp \{-(a-b)/(\lambda_a + \Sigma_i O_i)\}) \quad (9)
\]

and

\[
O^*(Z = a|Q) = P^*(Q, Z = a)/P^*(Q) = \exp \{b-a)/(\lambda_a + \Sigma_i O_i)\} \quad (10)
\]

Note that the density function \( P^*(Q) \) is the unique density over the space \( Q \) consistent with the full constraint set because it is formed by marginating \( P^*(Q, Z) \) which is the unique density function over the space \( (Q, Z) \) consistent with the full constraint set. Thus, the \( P^* \) designation is justified, i.e., all the inferred density functions are consistent with each other and the constraint set (see [13] properties 6–9). As a result of this consistency, we are free to use these probabilities in Bayes's theorem, or in an appropriate part of it, to produce another legitimate \( P^* \) density. The equalities (8–10) follow from the definition of conditional probability, odds, and substitution of the appropriate form of (6).

Now we solve (10) for the \( \lambda_i \)'s in terms of the constraint set. Further margination of (6) allows us to take care of \( \lambda_Z \) and \( \lambda_0 \).

\[
P^*(Z) = \int_{O_j=0}^{\infty} \cdots \int_{O_n=0}^{\infty} \exp \left\{ -\lambda_0 - \lambda_Z Z - \sum_{i=1}^{n} \lambda_i O_i Z \right\} dO_1 \cdots dO_n
\]

\[
= \exp \left\{ -\lambda_0 - \lambda_Z Z \right\} \prod_{i=1}^{n} \int_{O_i=0}^{\infty} \exp \left\{ -\lambda_i O_i Z \right\} dO_i \quad (11)
\]

From (11) follows

\[
P^*(Z = a) = \exp \left\{ -\lambda_0 - a \lambda_Z \right\} \prod_{i=1}^{n} \int_{O_i=0}^{\infty} \exp \left\{ -a \lambda_i O_i \right\} dO_i
\]

\[
= \exp \left\{ -\lambda_0 - a \lambda_Z \right\} \prod_{i=1}^{n} (a \lambda_i)^{-1} \quad (12)
\]

and

\[
P^*(Z = b) = \exp \left\{ -\lambda_0 - b \lambda_Z \right\} \prod_{i=1}^{n} (b \lambda_i)^{-1}. \quad (13)
\]

Notice that the integral in (11) diverges for \( Z \leq 0 \) and hence the initial constraint on valuating \( Z \). For the binary event \( Z \),

\[
P^*(Z = a) = E[Z] - b + (a-b) P^*(Z = a) = a - (a-b) P^*(Z = b),
\]

\[
P^*(Z = b) = O^*(Z = a) = E[Z] - b \quad (14)
\]

and via (12), (13) and the first equality of (14)

\[
\lambda_Z = \frac{1}{b-a} \left[ \log O^*(Z = a) + n \log (a/b) \right]. \quad (15)
\]

Each \( \lambda_j \) is solved by working with the appropriate marginal density of \( P^*(Q, Z) \), e.g.,

\[
P^*(O_j, Z) = \exp \left\{ -\lambda_0 - \lambda_j Z - \lambda_j O_j Z \right\} \prod_{i \neq j} \int_{O_i=0}^{\infty} \exp \left\{ -\lambda_i O_i Z \right\} dO_i,
\]

\[
= \exp \left\{ -\lambda_0 - \lambda_j Z - \lambda_j O_j Z \right\} \prod_{i \neq j} \left( \lambda_i Z \right)^{-1}. \quad (16)
\]

Because \( Z \) is binary, averaging is not difficult

\[
E[O_j Z] = \exp \left\{ -\lambda_0 \right\} \left[ \frac{\exp \left\{ -\lambda_j a \right\} \prod_{i \neq j} \left( \lambda_i a \right)^{-1}}{\lambda_j^2 a} + \frac{\exp \left\{ -\lambda_j b \right\} \prod_{i \neq j} \left( \lambda_i b \right)^{-1}}{\lambda_j^2 b} \right]. \quad (17)
\]

Because

\[
\sum_{Z \in \{a, b\}} \int_{O_j=0}^{\infty} P^*(O_j, Z) dO_j = 1
\]

we also find that

\[
\exp \left\{ -\lambda_0 \right\} = \frac{1}{\lambda_j} \left[ \frac{\exp \left\{ -\lambda_j a \right\} \prod_{i=1}^{n} \left( \lambda_i a \right)^{-1}}{\prod_{i=1}^{n} \left( \lambda_i b \right)^{-1}} \right]. \quad (18)
\]

Combining (16) and (18) gives

\[
E[O_j Z] = 1/\lambda_j \quad (19)
\]

or

\[
\lambda_j = 1/E[O_j Z]. \quad (20)
\]

Finally, we use the results of (15) and the second equality of (14) to rewrite (10) as

\[
O^*(Z = a|Q) = O^*(Z = a) \left( \frac{a}{b} \right) \prod_{i=1}^{n} \exp \left\{ -(a-b) \lambda_i O_i \right\} \quad (21)
\]
or, taking advantage of (20), to show the solution written explicitly in terms of the constraint set

\[
O^*(Z = a|O) = \frac{E[Z] - b}{a - E[Z]} \left( \frac{a}{b} \right)^n \prod_{i=1}^{n} \exp \{- (a - b)O_i / E[O_i] \}. \tag{22}
\]

We now have our first ME prescription for aggregating opinions. In the form of (21) we can see some similarities with Bordley’s result, (3), in that the aggregation is a product. This product includes one of the possible prior odds terms, \((E[Z] - b)/(a - E[Z])\), and it includes some function of the individual assessments \(O_i\) raised to the power \(-(a - b)\lambda_i\) where each of these powers corresponds to a \(w_i\) of (3). However, \(\exp\{-O_i\}\) is certainly not the same as \(O_i\) in (3), and the scaled product of priors, e.g., \(P_0^{w_i}\), does not appear in the product.

Because Bordley’s [1] development applies some of its axioms successively, it is possible to identify the axiom that causes his final result to be different than the ME inferred result. That is, (21) and (22) produced by ME inference are consistent with the ACM theory form, (4), even though (21) and (22) are not consistent with Bordley’s odds grouping, i.e., (3).

To see the equivalence of (21) to (4) let

\[
F(t) = \left( \frac{a}{b} \right)^n \cdot \frac{E[Z] - b}{a - E[Z]} \cdot e^t
\]

and then let

\[
u_i(O_i) = \log (\exp \{- (a - b)O_i \lambda_i \}) = -(a - b)O_i \lambda_i.
\]

Because application of the ACM theory axioms comes before application of the weak likelihood ratio axiom, it is possible to conclude that this new axiom must be the one producing the formulation that differs from the ME result. It is also important to note the dependence of the posterior probability in (22) on particular values of the variable \(Z\). This effect certainly comes as no surprise, since ME inference of many functions of a continuous variable are affected by the range of the variable, and now we have multiplied a binary variable by a continuous variable. As a result, for each different choice of the \(\{a, b\}\) pair, there corresponds a different joint probability density function. This is unfortunate, because the numerically-valued symbols which are selected to encode the occurrence or non-occurrence of an event should not affect the final aggregate assessment. For the particular constraint set under consideration, the only way to remedy this ambiguity is to introduce axioms that will uniquely determine \(a\) and \(b\). Such axioms can be devised—for example, to ignore the opinions of a useless expert, or to avoid the dominance of one expert over the rest—but the list of such possible axioms is long, and the careful exploration and comparison of them is beyond the goals of this paper.

As we show in Appendix II, there is a way around this problem by proper selection of the constraint set. Moreover, the formulation presented in Section VI does not depend on the valuation of \(\{a, b\}\).

V. AGGREGATING PROBABILITY ASSESSMENTS

In the above derivation we used odds, \(O_a\) and \(O_b\), because Bordley [2] developed his formulation using these variables. However, why not use individual probabilities to get an aggregate probability? After all, in the United States, weather forecasters are producing probabilities.

Although producing a different ME aggregation formulation, just changing the space from odds to probability does not produce Bordley’s result. The ME probability density function on the constraint set \(\{E[P_i, Z], \forall i, E[Z]\}\) for individual probabilities \(P_i \in [0,1]\), leads to equations identical to (6)-(10) because we need only replace \(O_i\) with \(P_i\) and \(Q\) with the corresponding probability vector \(P\). However, when it is time to integrate, we see that numerical techniques are required and that it is not possible to derive an aggregation formulation with explicitly valued weights as found in (22). (Because the required integration of the expectations that lead to the calculation of the \(\lambda_i's\) are integrations of an exponential form between zero and one, transcendental equations, in analogy to (16)-(20), must be solved to resolve the \(\lambda_i's\).) Even so, because the aggregation formulas are the same as (9) and (10), we still can say that the ME solution differs from Bordley’s [1] final answer but do coincide with the ACM theory derived form.

Thus, the conclusions are so far: 1) ME inferred aggregations agree with ACM theory aggregations and 2) when using an ME inferred aggregation, there is a unique specification of the weighting factors \(\lambda_i\) which are left unspecified in Bordley’s ACM derived theory.

A. Steps Towards Bordley’s Result

So far, our attempts to understand (1) as a potential ME inferred result have been rather naive in the sense that we have taken simple averages of the random variables (i.e., odds or probabilities) and then applied ME inference on the random variable. Now we will change the perspective slightly and try to produce the desired form by working backwards. In addition, we will assume that \(Z \in \{0,1\}\) to provide a direct connection to Bordley’s result.

The Bayesian-like reversal of the conditioning and conditioned variables seems to produce the type fraction sought (i.e., we get the form \(P_{\text{group}} = \text{Prob}(A)/(\text{Prob}(A) + \text{Prob}(B))\), so we are on the right track here. Furthermore, it seems inevitable that we must hypothesize that the weights, \(w_i\), of (1) must correspond to the Lagrange multipliers, \(\lambda_i\), of (5). The inclusion of \(E[Z]\) in the constraint set seems to produce terms corresponding to \(P_0\) and \(1 - P_0\). Finally, to derive terms such as \(P_i^{\lambda_i}\) instead of \(\exp\{-\lambda_i P_i\}\) we will need something like expectations of logarithms of \(P_i\) in the constraint set. It even seems possible to get terms corresponding to \(P_{0}^{w_i}\) and \((1 - P_0)^{w_i}\) by collecting expectations such as

\[
E\left[ Z \log \frac{P_i}{P_0} \right] \text{ and } E\left[ (1 - Z) \log \frac{1 - P_i}{1 - P_0} \right].
\]

Unfortunately, the closest we can come to (1) with this line
of thought is

$$P^*(Z = 1|X) = \frac{\prod_{i=1}^{n} \left( \frac{P_i}{P_0} \right)^{-\lambda_{i,1}}}{\prod_{i=1}^{n} \left( \frac{P_i}{P_0} \right)^{-\lambda_{i,1}} + \prod_{i=1}^{n} \left( \frac{1-P_i}{1-P_0} \right)^{-\lambda_{i,0}}}$$

when the constraint set is

$$\left\{ E\left[ Z \log \frac{P_i}{P_0} \right], \forall i, E\left[ (1-Z) \log \frac{1-P_i}{1-P_0} \right], \forall i \right\}.$$  \(23\)

Thus, for (23), there are two differences with (1): the appearance of different weighting factors (Lagrange multipliers) for the two terms in the denominator \(()ri,lVS.Xi,o\) and the inability to get both the weighted and unweighted PO terms to appear at once. Indeed, this last problem seems insurmountable. The ME development will not allow us to put arbitrary unweighted constants in a distributional form because the only term that is of this form (i.e., neither weighted nor a function of the random variable) is the normalization term that corresponds to \(\exp\{-A_0\}\) of (5) or, none at all when constraints are conditional on \(f(2)\). In the case of \(\exp\{-A_0\}\), this normalization term cannot be arbitrarily set in advance but is a function of the values in the constraint set. In the other case, when \(P_0\) is multiplicatively applied to the random variable \(Z\) and to the random variables \(X_i\), it will disappear.

B. Additional Assumptions Produce Bordley's Result

Still, we might finesse the problem of the differences; if the terms with \(P_0\) and \(1-P_0\) are the problem, perhaps they can be ignored. That is, \(P_0\) drops out of Bordley's [11] formulation by either letting \(\Sigma w_i = 1\) or assuming fully naive priors, i.e., \(P_0 = 1-P_0 = 0.5\). In either case, all such prior terms cancel and (1) becomes (2) while (3) becomes

$$O_2 = 1^{(1-\sum w_i)} \prod_{i=1}^{n} O_i^{w_i} = \prod_{i=1}^{n} O_i^{w_i}.$$  \(24\)

These forms are attainable under the ME regimen combined with Bayes's ideas if a particular constraint set is chosen.

Let \(P_i \in [0,1]\) be the probability of an individual expert about event \(Z \in [0,1]\) and let \(P\) be the \(n\)-dimensional vector made up of the \(P_i\)'s. Suppose we collect statistics so that the constraint set is

$$\left\{ E\left[ Z \log \frac{P_i}{1-P_i} \right], \forall i, E\left[ (1-Z) \log \frac{1-P_i}{1-P_0} \right], \forall i \right\}.$$  \(25\)

In general, ME inference on this constraint set gives

$$P^*(P, Z) = \exp\left\{ -A_0 + \sum_{i=1}^{n} \lambda_i Z \log \frac{P_i}{1-P_i} \right\}$$

$$= e^{-A_0} \prod_{i=1}^{n} \left( \frac{P_i}{1-P_i} \right)^{\lambda_i} Z.$$  \(26\)

Substituting for the two possible values of \(Z\) gives

$$P^*(P, Z = 0) = e^{-A_0}$$

$$P^*(P, Z = 1) = e^{-A_0} \prod_{i=1}^{n} \left( \frac{P_i}{1-P_i} \right)^{\lambda_i} Z.$$  \(27\)

Again, because of the consistency of the \(P^*\) designation throughout, we can use ideas from Bayes to produce another unique, consistent \(P^*\) density. That is, substituting (24) into the appropriate conditional form gives

$$P^*(Z = 1|P) = \frac{P^*(P, Z = 1)}{P^*(P)} = \frac{P^*(P, Z = 1)}{P^*(P, Z = 0) + P^*(P, Z = 1)}$$

$$= \frac{\prod_{i=1}^{n} P_i^{\lambda_i}}{\prod_{i=1}^{n} (1-P_i)^{\lambda_i} + \prod_{i=1}^{n} P_i^{\lambda_i}},$$  \(28\)

which coincides with Bordley's formula in (2).

VI. AGGREGATION FORMULAS WITH MORE INFORMATION

As should be obvious from preceding developments, an ME inferred aggregation formula depends on the elements of the constraint set, and this observation spawns at least two points worth considering. First, there are as many possible formulas as there are constraint sets that imply a probability density function. Second, given two constraints \(M_1\) and \(M_2\) such that \(M_1 \subset M_2\), then \(M_2\) has more information than \(M_1\) where more information means the inferred density function has less entropy. Thus, we can do better than (23) or (25) by increasing the information in the constraint set. For example, we can expand the constraint set defined previously as \(M_a\) by including unconditioned information about \(Z\) (i.e., \(E[Z]\)). Specifically let the enlarged constraint set be

$$M_{a+} = \{ E[-Z \log P_i], \forall i, E[-(1-Z) \log P_i], \forall i, E[-Z \log (1-P_i)], \forall i, E[-(1-Z) \log (1-P_i)], \forall i, E[Z] \}.$$  \(29\)

Even though the enlarged set has additional information about each expert's prediction and has unconditioned information about the event itself, it still produces a tractable ME probability calculation. This expanded constraint set produces the binary mass function \(P^*(Z)\), and Beta density forms for each \(P^*(P_i|Z = 1)\) and \(P^*(P_i|Z = 0)\) so that the aggregate probability is (see (26) at the bottom of the page) where \(B(\cdot, \cdot)\) is the beta function. All the lower case Greek parameters are calculated from the digamma function based on the valued expectations in the constraint set as explained by Tribus [14].

From the point of view of ME inference, the formulation in (26) is an improvement over Bordley's. That is,
the constraining equalities that helped generate Bordley's formulation are unnecessary restrictions (and are unlikely to be true), there is more information about each individual's prediction history (i.e., the additional expectations such as \(E[-(1 - Z) \log(1 - P_i)]\) and \(E[-Z \log(1 - P_i)]\), and some information about the event itself is now used (i.e. \(E[Z]\) is now in the constraint set).

Unlike the earlier ME aggregations of this paper, the particular values chosen to encode \(Z\) do not affect the ME aggregation of probabilities in (26) (see Appendix II).

VII. INTERACTIONS BETWEEN EXPERTS

Still other results are possible with other constraint sets. Most obvious and desirable is to extend the elements of the constraint set to include expectations that compensate for statistical dependence between and among the individuals. There are many instances in which it seems important to compensate for such interactions, for example, when using opinion givers that might share information sources, educational backgrounds, or theoretical dispositions relative to the problem at hand.

Pair-wise dependencies can be taken into account for an odds aggregation. The constraint set of odds \(\{E[O_i|Z], \forall i\) for \(Z \in \{0, 1\}, E[O_iO_j|Z], \forall i, j\) for \(Z \in \{0, 1\}, E[Z]\) produces a legitimate density function which takes into account pair-wise interactions between opinion givers although the implied density function is difficult to integrate. (Apparently special functions are needed to perform the integration.)

A more tempting constraint set that also takes into account pair-wise interactions is

\[
\{E(Z), E[Z \cdot f(P_i)], \forall i, E[(1 - Z) \cdot f(P_i)], \forall i,
E[Z \cdot f(P_i) \cdot f(P_j)], \forall i, j, E[(1 - Z) \cdot f(P_i) \cdot f(P_j)], \forall i, j\}
\]

where

\[
f(P_i) = \log \frac{P_i}{1 - P_i}
\]

and

\[P_i\]

is restricted to \((0, 1)\).

Note that \(f(P_i)\) maps a probability from the open unit interval to \((-\infty, \infty)\) so that this set produces the multivariate normal density functions, \(P^*(Z = 0, f(P))\) and \(P^*(Z = 1, f(P))\). In this case the density parameters can be explicitly solved as a function of the constraint set \([14]\), and \(P^*(Z = 1|X)\) is then produced by applying Bayes' ideas.

It is difficult to continue on in this vein, that is to infer density functions while using continuous variables and filling the constraint set with multivariate correlational terms of arbitrary order. The problem is that some of the integrals needed for the expectation equations diverge in such situations. In a real world setting, however, the restriction to continuous variables should be reconsidered because it does not seem sensible to ask our experts to produce probabilities of infinite precision.

VIII. A COMPUTATIONAL APPROACH ON A DISCRETE SPACE IS UNRESTRICTED

Finally, while trying to increase the information in the constraint set, we take a more practical viewpoint. There is no reason why a formula with the parameters written as explicit functions of the constraint set should be the best prescription for aggregating opinions. That is, an explicit computational procedure should be good enough for our purposes, especially if it can take higher order interactions into account.

As pointed out by the authors in their last paragraph, the iterative proportional fitting procedure (IPFP) of Darroch and Ratcliff \([4]\) is an ME-consistent approach (note that they use the older terminology of minimum discrimination information). The IPFP can take advantage of much more information than any of the previous constraint sets using continuous variables (where information is used in the sense of incorporating correlations of arbitrary orders between and among individuals), and it does not require transformations or dropping end points. For example, it is, in principle, possible for the IPFP to use correlations such as \(E[P_iP_jP_kZ]\).

To use the IPFP, we need one rather mild restriction: opinions are limited to predefined values of finite accuracy. For example, it might be sensible to collect probabilities where values are restricted to the set \(\{0, 1, 2, \cdots, 9, 1.0\}\).

In addition to the predefined points, we must decide on the statistics to collect. We should always collect the most basic statistics \(E[Z]\) and the \(E[P_iZ]\) but, for example, we might wish to compensate for pair-wise, for three way, or for higher order interactions among the experts. Then we might, just for example, collect the following statistics as the constraint set \(\{E[P_iZ], \forall i \in \{0, 1\}, E[P_iP_jZ], \forall i, j \in \{0, 1\}, E[P_iP_jP_kZ], \forall i, j, k \in \{0, 1\}, E[Z]\}\)

where \(i, j, k\) range over the \(n\) opinion givers.

For any such constraint set, the IPFP can be applied to create \(P^*(Z, P)\) from which \(P^*(Z = 1|E)\) is inferred as before with (8)

\[
P^*(Z = 1|E) = \frac{P^*(Z = 1, E)}{P^*(Z = 1, E)P^*(Z = 0, E)}.
\]

To help the reader more quickly bridge between the present article and the work by Darroch and Ratcliff \([4]\), note that when they are considering a point in a distribution as \(P_i = \pi_i\mu \prod_{s=1}^k \beta_i^s\), their \(P_i\) corresponds to our \(P^*(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n, Z = z)\) and their \(\mu\) corresponds to \(e^{-\lambda_i}\) as in (5) here. Thus, their probability mass function is the correct form of an ME inferred density function for a discrete-valued variable.

As versatile as such a procedure is, it does have its practical limitations. Eventually, with the inclusion of many experts and many higher order interactions, this approach will be limited by the accuracy of sample means as approximations of population means. That is, sample sizes will tend to be smaller and smaller for the higher order interactions of many experts. As a result, sample averages will not always be good approximations of population averages and thus, the application of ME inference can be called into doubt. In
addition a computational tractability problem occurs with many experts and too many higher order interactions because, in the case that all interactions are used, memory requirements are of size \(2^n\) even when opinions are restricted to a binary situation.

**IX. CONCLUSION**

We have examined maximum entropy (ME) inferred aggregation formulas for combining expert opinions in the form of odds or probabilities. We have shown that the ME and ACM are compatible aggregation formulas; but, ME goes further, as discussed in the next paragraph. In this regard, Bordley’s weak likelihood ratio axiom is too restrictive as better forms exist in terms of using more information, such as the ME inference on the enlarged constraint set that was discussed.

The problem of aggregating expert information assuming that only the moments of the density functions are available was also investigated by Genest and Schervish [9]. Following non-axiomatic Bayesian inference, they obtain a formula which coincides with Bordley’s (and ours for a special case). Genest and Schervish assume, however, that the experts specify their prior probabilities (in addition to moment constraints), whereas in ME development, prior opinions are wiped out by the expectations of the actual behavior of the experts. Moreover, ME incorporates the probability of the event being predicted in the form of its expected value. Finally, the ME approach extends to situations where the decisions of the experts are dependent, a quality that distinguishes the ME technique from other axiomatic derivations.

The ultimate suggestion of the paper is to use the iterative proportional fitting procedure of Darroch and Ratcliff so that as much information about the opinions and their interactions can be included in the aggregation calculations as possible.

**APPENDIX I**

The weak likelihood axiom is built around the idea of supplying each expert with the same correct likelihood value, \(L_i\), in some way that will not affect their individual opinions. That is, after each expert produces \(P_i(E|A)\) we define for all \(i\)

\[
L_i = L = \frac{P_i(A| E)}{P(A \text{ not } E)} = O_i(E| A) \cdot O_i(E)^{-1}
\]

or equivalently

\[
O(E) = L^{-1} O(E| A)
\]

and

\[
O_i(E| A) = L \cdot O(E)
\]

where \(E\) is the event being predicted and \(A\) is some other piece of information that precedes \(E\). Next the axiom develops by assuming the same relationship holds for the group odds. In particular, from the equality

\[
O_G(E| A) = F^{E|A} \left( \sum_{i=1}^{n} u_i^{E|A}(O_i(E| A)) \right)
\]

the axiom states

\[
F^{E|A} \left( \sum_{i=1}^{n} u_i^{E|A}(L O_i) \right) = L F^E \left( \sum_{i=1}^{n} u_i^{E}(O_i) \right)
\]

Apparently this condition severely restricts the possible forms that are allowed by the ACM theory development down to two.

Finally, one of these two is easily eliminated by the, rather natural, normalization axiom which just requires

\[
O_G(E) \cdot O_G(E) = 1
\]

\[
O_k(E) \cdot O_k(E) = 1 \text{ for all experts } k.
\]

**APPENDIX II**

In the ME odds assessment of (22), which was developed from the constraint set \{\(E[Z], E[Z|O_1], \ldots\)\}, we see an aggregation prescription that is dependent on the values selected for \(a\) and \(b\). This is quite undesirable because we seek a methodology which lacks arbitrary prescriptions and because many prediction problems do not have a natural numeric valuation (e.g., {win, lose}, {rain, no rain}, \(\{X < 5, X \geq 5\}, \text{ etc.}\) ). Fortunately, this valuation problem is not a general property of ME inference on binary outcomes, and it is avoided by selecting the appropriate constraint set.

The key to avoiding a dependency on the valuation of \(\{a, b\}\) is the invariance property of ME inference. This property says that any two constraint sets \(M_1\) and \(M_2\) are equivalent if we can create \(M_1\) from \(M_2\) and vice versa by coordinate transformations [12], [13] on the expectations. If \(Z' \in \{a, b\}\), then the coordinate transformation we will use on \(Z'\) and the expectations maps \(f : Z' \rightarrow Z\) such that \(f(a) = 0\) and \(f(b) = 1\). Some examples will help; let \(Z' \in \{a, b\}\) and \(f(Z') \in \{0, 1\}\).

**Example 1:** \(M_1 = \{E[Z'], E[Z'|O_1], \ldots\}\) and \(M_2 = \{E[f(Z')], E[f(Z')|O_1], \ldots\}\). Because \(f\) is one-to-one and \(f(Z') \in \{0, 1\}\), \(P(Z' = b) = P(f(Z') = 1) = E[f(Z')]\), Also, \(E[Z'] = aP(Z' = a) + bP(Z' = b) = a + (b - a)P(Z' = b)\). Therefore,

\[
E[f(Z')] = (E[Z'] - a)/(b - a)
\]

Note that, by (A1), \(E[f(Z')]\) is a function of \(E[Z']\) and \(\{a, b\}\), which is just what we needed to prove the equivalence of \(M_1\) and \(M_2\). In fact, because this same argument is reversible, any \(\{a, b\}\) can be transformed to any other \(\{c, d\}\), so all valuations must be equivalent.

In the next example, we add one more constraint to example (1) to get a set that is equivalent to a simplified version of \(M_{a+b}\).

**Example 2:** \(M_1 = \{E[(Z' - a)g(X)], E[(Z' - b)g(X)]\} \equiv M_2 = \{E[f(Z')g(X)], g\} \), where \(g\) is an arbitrary function. We need only note that \(E[(Z' - a)g(X)](b - a)^{-1} = E[f(Z')g(X)]\) because \(Z' - a \in \{0, b - a\}\).

Thus, we could have been more general in section VI by writing \(M_{a+b}\) with expectations such as \(E[(Z' - a)(b - a)^{-1} \log P_i]\) and \(E[(b - Z')(b - a) \log P_i]\), but it is hardly worth the effort because of the equivalence to the set actually used.
Example 3: \[ M_1 = \{ E[Z'], E[Z'g(X)]Z' = b \} \equiv M_2 = \{ E[f(Z')], E[f(Z')g(X)]f(Z') = 1 \}. \]
Note that only \( Z' = b \), not \( Z = a \), contributes to \( E[Z'g(X)]Z' = b \) so the constant \( b \) can be factored out, and because \( P(Z' = b) = P(f(Z') = 1) \),
the result follows.
Thus, if \( E[Z'] \) is present and the other moments are conditioned on \( Z' \), then we can connect them merely as a function of themselves and \( \{a, b\} \) to a set dependent on any other binary pair \( \{a, b\} \).

We can go even further if both \( E[g(X)|Z' = a] \) and \( E[g(X)|Z' = b] \) are available in the constraint set in addition to \( E[Z'] \). Then, we also have access to \( E[Z'g(X)] \) through the equality \( E[Z'g(X)] = aE[g(X)|Z' = a]P(Z' = a) + bE[g(X)|Z' = b]P(Z' = b) \).
Thus, there are several types of constraint sets that avoid dependency of the valuation of \( \{a, b\} \).

ACKNOWLEDGMENT
The authors would also like to thank the reviewers for showing a way closer to Bordley's result and finding an integration error in the original manuscript. Mr. Steve Friedman and Dr. Donald St. P. Richards provided help and encouragement in the calculation of the derivations.

REFERENCES

Hakan Deliç was born in Bonn, Germany, on October 28, 1966. He received the B.S. degree (with honors) in electrical engineering from Bogaziçi University, Istanbul, Turkey, in 1988 and the M.S. and Ph.D. degrees in electrical engineering from the University of Virginia, Charlottesville, VA, in 1990 and 1992, respectively. He is currently with the University of Virginia Health Sciences Center. Dr. Deliç's research interests include detection and estimation theory, robust statistics and its applications, distributed decision making, communication networks, and neural networks.

William B Levy (M'85) was born in St. Louis, Missouri, on March 28, 1947. He received the A.B. degree in psychology from Princeton University, N.J., in 1969 and the Ph.D. degree in biological sciences from the University of California at Irvine in 1973. He is currently a Research Professor of Neurosurgery and, since 1989, the Director of Neurosurgery Research Laboratories at the University of Virginia Health Sciences Center. Dr. Levy is a member of the Society for Neuroscience, the American Society for Neurochemistry, the Cognitive Sciences Society, the American Association of Anatomists, the International Neural Network Society, and the American Psychological Society. His current research interests include prediction by both artificial and natural neural networks and the neural basis of cognition.