Maximum Entropy Distributions between Upper and Lower Bounds

Ali E. Abbas

School of Engineering, University of Illinois at Urbana-Champaign
104 South Mathews Ave, Transportation Building, MC-238, Urbana, IL, 61801.

Abstract. We discuss the formulation of discrete maximum entropy problems given upper and lower bounds on moments and probabilities. We show that with bounds on discrete probabilities, and bounds on cumulative probabilities, the solution is invariant to any additive concave objective function. This observation simplifies the analysis of the problem and unifies the solution of several generalized entropy expressions. We use this invariance result to provide an exact graphical solution to the maximum entropy distribution between upper and lower cumulative probability bounds. We also discuss the maximum entropy joint distribution with bounds on marginal probabilities and provide a graphical solution to the problem using properties of the entropy expression.

INTRODUCTION

In 1957, Edwin Jaynes proposed a method to assign probabilities based on partial information [1]. This is known as the maximum entropy principle and is stated from his original paper as follows: "In making inferences on the basis of partial information, we use the probability distribution which has maximum entropy subject to whatever is known". Maximum entropy applications have since found great popularity in many fields such as natural language processing [2], bioinformatics [3], medicine, thermodynamics [4], DNA sequence alignment, hydraulics [5] , decision analysis [6], [7], and many other fields.

A maximum entropy model is usually formulated to confirm to equality constraints on moments or cumulative probabilities of the distribution of a random variable, θ.

$$p^* = \arg\max_{p} \; -\sum_{i=1}^{n} p_i \ln(p_i)$$

$$\text{Such that} \qquad (1)$$

$$\sum_{i=1}^{n} h_j(\theta_i) p_i = \mu_j \;\; j=1,...m, \;\; \sum_{i=1}^{n} p_i = 1 \; and \; p_i \geq 0,$$

where $h_j(\theta_i)$ is either an indicator function over an interval for cumulative probability constraints, or θ_i raised to a certain power for moment constraints, and μ_i's are a given sequence of probabilities or moments.

CP803, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*
edited by Kevin H. Knuth, Ali E. Abbas, Robin D. Morris, and J. Patrick Castle

The solution to this problem has the well-known form

$$p_i = e^{-\alpha_0 - 1 - \sum_{i=1}^{n} \alpha_j h_j(\theta_i)}, \tag{2}$$

where α_j is the Lagrange multiplier for each probability or moment constraint.

In many cases that arise in practice, however, precise values for moments and probabilities are unavailable. For example, we may have (1) imprecision in certain measurements, such as temperature sensors needed to determine the average kinetic energy of molecules in a room; (2) insufficient data collected in data mining applications; (3) bounds on the assessed probabilities in a typical probability encoding in decision analysis; and (4) game theoretic formulations where we are uncertain about an opponent's beliefs, but have upper and lower bounds resulting from prior information.

When precise values for moments and probabilities are unavailable, the maximum entropy principle can be used to assign a representative probability distribution using upper and lower bounds. This formulation has several applications such as the maximum entropy joint distribution with lower-order marginal constraints [7], [8]. The mathematical formulation for the problem using vector notation is shown below. We choose (for convenience) to minimize the negative of the entropy function, $f(p) = \sum_{i=1}^{n} p_i \ln(p_i)$, instead of maximizing the entropy function.

$$\begin{aligned} & p^* = \arg\min_{p} f(p) \\ & \text{such that} \\ & e^T p = 1 \\ & ma \leq H(\theta) p \leq mb, \end{aligned} \tag{3}$$

where $H(\theta) \in R^{mxn}$ is a matrix whose rows are the constraints $h_j(\theta)$; ma and $mb \in R^{mx1}$ are vectors whose elements, ma_j and mb_j, are the lower and upper bounds (respectively) on the expected value of $h_j(\theta)$; $e \in R^{nx1}$ is a vector of ones; and e^T is the transpose of e.

Equation (3) is a generalization of (1) when the upper and lower bounds coincide. Our focus will, therefore, be on (3), while keeping in mind that the same results apply to (1). It is useful to think of the feasible region for (3) geometrically where each inequality constraint provides a region bounded by two cones: an outer cone whose projection on the vector $h_j(\theta)$ is the dot product $h_j(\theta)p = ma_j$ and is also the lower bound, and an inner cone whose projection on $h_j(\theta)$ is the dot product, $h_j(\theta)p = mb_j$, and is also the upper bound. If there were only equality constraints, the upper and lower cones would coincide. The equality constraint, $e^T p = 1$, limits the norm of the vector p to be unity (Figure 1). The feasible region for the problem is the intersection of the feasible regions of all the inequality constraints. We are interested in a vector in the feasible region that will minimize the convex objective function, $f(p)$.

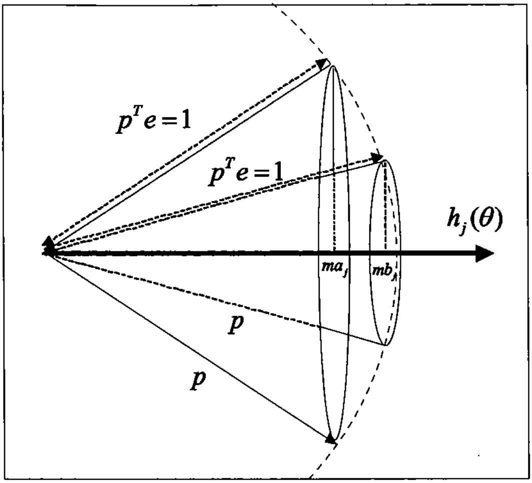

FIGURE 1. Feasible region implied by the upper and lower bounds of one inequality constraint.

We will discuss the optimality conditions for (3), and show that for (i) bounds on discrete probabilities and (ii) bounds on cumulative probabilities, the solution is invariant to any additive convex objective function. This observation simplifies the analysis of the problem significantly. For example, we can avoid some of the problems that occur with the singularity of the objective function $\phi(x) = x\ln(x)$ near the origin, or choose an objective function that enables an exact graphical solution to the problem. The invariance of (3) under the conditions stated above unifies the solutions of several generalized entropy expressions when bounds on probabilities are available.

We end this paper with a discussion of the maximum entropy joint distribution with bounds on marginal probabilities, and show how properties of the entropy expression enable the decomposition of this problem into a simpler formulation in terms of the sum of the entropies of the marginal distributions. This result simplifies the analysis and reduces the number of variables significantly.

OPTIMALITY CONDITIONS FOR THE MAXIMUM ENTROPY SOLUTION WITH UPPER AND LOWER BOUNDS

Note that the objective function in (3) is an additive convex function since each term, $p_i \ln(p_i)$, is strictly convex. Furthermore, the region defined by the inequality constraints is the intersection of a set of half spaces and is, therefore, a convex set. These conditions suffice the existence of a unique solution to the problem [9].

The first step in the analysis is to take the Lagrange operator of (3) to get

$$L(p,\alpha,\beta,\mu) = f(p) - \mu(e^T p - 1) - ([H(\theta)]p - ma)^T \alpha - (mb - [H(\theta)]p)^T \beta, \tag{4}$$

where $\alpha, \beta \in R^{m\mathrm{x}1}$ are vectors whose elements are the Lagrange multipliers for the lower and upper bounds respectively, and μ is a scalar Lagrange multiplier for the normalizing equality constraint. For convenience, we re-arrange (4) to get

$$L(p,\alpha,\beta,\mu) = [f(p) - p^T(\mu e + H(\theta)^T(\alpha - \beta))] + \mu + ma^T\alpha - mb^T\beta. \tag{5}$$

Now we consider the first-order optimality conditions for (3).

First-Order Optimality Conditions

(a) The First Partial Derivative $\frac{\partial}{\partial p} L(p,\alpha,\beta,\mu)\,|_{p^*} = 0$

This condition results in a solution, p^*, that minimizes $L(p,\alpha,\beta,\mu)$ and is also the maximum entropy solution. I.e.

$$p^* = \underset{p}{\operatorname{argmin}}\ L(p,\alpha,\beta,\mu). \tag{6}$$

Note that the $\underset{p}{\operatorname{argmin}}\ L(p,\alpha,\beta,\mu)$ operation will involve only the bracketed term in (5). Furthermore, the $\underset{p}{\operatorname{argmin}}$ operation performed on the negative of the bracketed term is (by definition) equal to the conjugate of the function $f(p)$, written as $f^*(.)$. See for example, Rockafeller [10], where

$$\begin{aligned} &\underset{p}{\operatorname{argmin}}[f(p) - p^T(\mu e + H(\theta)^T(\alpha - \beta))] \\ &\qquad = -\underset{p}{\operatorname{argmax}}[p^T(\mu e + H(\theta)^T(\alpha - \beta)) - f(p)] \triangleq -f^*(y^*), \end{aligned} \tag{7}$$

where $f^*(.)$ is the conjugate of $f(p)$ and y^* is by definition equal to

$$y^* = (\mu e + H(\theta)^T(\alpha - \beta)). \tag{8}$$

The conjugate of the negative of the entropy expression is $f^*(y^*) = \ln(\sum_{i=1}^{n} e^{y_i^*})$ as we show in the appendix. We can express the maximum entropy solution, p^*, in terms of y^* by taking the first partial derivative of (5) with respect to p and equating it to zero.

$$\nabla f(p^*) = (\mu e + H(\theta)^T(\alpha - \beta)) \triangleq y^*. \tag{9}$$

Note that the expression for y^* in (8) as well as the relation for the gradient $\nabla f(p^*) = y^*$ in (9) is the same if we minimize any additive convex objective function, $f(p) = \sum_{i=1}^{n} f(p_i)$, where $f(p_i)$ is strictly convex. As a result, (9) is also the same for any additive convex function in (3). Furthermore, the strict convexity of $f(p)$ implies that the argument of its gradient, $\nabla f(p^*)$, is uniquely determined by y^*. Hence, it is sufficient to determine y^* in order to determine the maximum entropy solution, p^*.

(b) The Partial Derivative $\frac{\partial}{\partial \mu} L(p,\alpha,\beta,\mu) = 0$

Taking the first partial derivative of (5) with respect to μ and equating it to zero gives the normalizing constraint

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow e^T p = 1 \text{ or } \sum_{i=1}^{n} p_i = 1. \qquad (10)$$

(c) Complementary Slackness Conditions

The complementary slackness conditions for the inequality constraints in (3) require

$$([H(\theta)]p^* - ma)^T \alpha = 0 \qquad (11)$$

$$(mb - [H(\theta)]p^*)^T \beta = 0 \qquad (12)$$

Equations (11) and (12) can also be expressed on an element-by-element basis as

$$\alpha_j (h_j(\theta)^T p^* - ma_j) = 0 \qquad j = 1,\ldots,m \qquad (13)$$

$$\beta_j (mb_j - h_j(\theta)^T p^*) = 0 \qquad j = 1,\ldots,m \qquad (14)$$

(d) Non-negativity of the Lagrange Multipliers of the Inequality Constraints

These conditions can be expressed on an element-by-element basis for the vectors α and β as

$$\alpha_j \geq 0 \qquad j = 1,\ldots,m \qquad (15)$$

$$\beta_j \geq 0 \qquad j = 1,\ldots,m \qquad (16)$$

Equations (9), (10), (13), (14), (15), and (16) are sufficient to determine y^* and the maximum entropy solution to the problem. Once again, these equations apply to any additive convex function; however the relation $\nabla f(p^*) = y^*$ uniquely determines the particular p^* for a given additive convex function, $f(p)$. In the next section we discuss some important special cases of this formulation that appear in many applications in practice, and where the solution is invariant to any additive convex function, $f(p)$.

INVARIANCE OF THE MAXIMUM ENTROPY SOLUTION

Now we discuss some special cases of (3) where the solution is invariant to any additive convex function.

Bounds on Discrete Probabilities

When only bounds on discrete probabilities are available, the matrix $H(\theta)$ in (3) reduces to an identity matrix, and the product $h_j(\theta)^T p$ reduces to p_j to give inequality constraints of the form

$$ma_j \leq p_j \leq mb_j, \quad j = 1,...m \tag{17}$$

With no loss of generality, we can assume that m=n, since we can introduce a constraint $0 \leq p_j \leq 1$ if we have no other information about the probability, p_j. The expression for y^* in (8) reduces to

$$y^* = (\mu e + (\alpha - \beta)) \tag{18}$$

The first-order optimality conditions reduce to

$$\alpha_j (p_j^* - ma_j) = 0 \qquad j = 1,...,m \tag{19}$$

$$\beta_j (mb_j - p_j^*) = 0 \qquad j = 1,...,m \tag{20}$$

$$\alpha_j, \beta_j \geq 0 \qquad j = 1,...,m \tag{21}$$

The non-negativity of α_j, β_j in (21) implies three possible cases for each probability, p_j.

Case 1: $\alpha_j = 0, \beta_j \neq 0$

From (20), this implies that p_j^* is determined solely by the upper bound, where

$$p_j^* = mb_j. \tag{22}$$

Let us now determine the value of μ for this case. From (18), we have $y_j^* = \mu - \beta_j$, and from (9), we see that this condition implies

$$\frac{\partial f(p)}{\partial p_j}\Big|_{mb_j} \triangleq y_j^* = \mu - \beta_j. \tag{23}$$

The non-negativity of, β_j, implies that (22) is an optimal solution when

$$\frac{\partial f(p)}{\partial p_i}\Big|_{mb_i} \leq \mu. \tag{24}$$

Now we discuss the implications of (23) and (24) for the function $f(p_i) = p_i \ln(p_i)$ and for any general additive convex function.

(a) $f(p_i) = p_i \ln(p_i)$

This is the case of (3), where equation (23) becomes

$$1 + \ln(mb_i) = \mu - \beta_i. \tag{25}$$

Since $\beta_j \geq 0$, (25) occurs when

$$\mu \geq 1 + \ln(mb_i), \tag{26}$$

Using the substitution, $\delta = e^{\mu - 1}$, we can write (26) as

$$\delta \geq mb_i. \tag{27}$$

(b) $f(p) = \sum_{i=1}^{n} f(p_i)$, where $f(p_i)$ is any strictly convex function

Recall that y^* is still given by (8) for any additive convex function. Therefore, (23) is also valid for any additive convex function where

$$\frac{\partial f(p)}{\partial p_j}\Big|_{mb_j} \triangleq y_j^* = \mu - \beta_j, \tag{28}$$

which occurs when

$$\frac{\partial f(p)}{\partial p_j}\Big|_{mb_j} \leq \mu. \tag{29}$$

Now we introduce a general definition of δ as the inverse of the partial derivative $\left(\frac{\partial f(p_j)}{\partial p_j}\right)$ evaluated at μ. Using this general definition of δ, (29) can be written as

$$\delta \geq mb_j. \tag{30}$$

From the previous results, we see that for any additive convex function in (3), $p_j^* = mb_j$ when $\delta \geq mb_j$.

Case 2: $\alpha_j \neq 0, \beta_j = 0$

From (19), this case implies that

$$p_j^* = ma_j, \tag{31}$$

and from (18), we have,

$$\frac{\partial f(p)}{\partial p_j}|_{ma_j} \triangleq y_j^* = \mu + \alpha_j. \tag{32}$$

The non-negativity of α_j implies that (31) is an optimal solution to (3) when

$$\mu \leq 1 + \ln(ma_j). \tag{33}$$

Furthermore, (31) is an optimal solution to any additive convex function, $f(p)$, when

$$\delta \leq ma_j. \tag{34}$$

Case 3: $\alpha_j = 0, \beta_j = 0$

From (18), this case implies that

$$y_j^* = \mu, \tag{35}$$

and from (9), we have

$$\frac{\partial f(p)}{\partial p_j}|_{ma_j} = \mu. \tag{36}$$

Using the general definition of δ, we can write (36) as

$$p_j^* = \delta, \tag{37}$$

which occurs in the remaining interval $ma_j \le \delta \le mb_j$.

We now summarize the solution to the maximum entropy problem (or any additive concave function) with bounds on discrete probabilities below.

$$p_j^* = \begin{cases} ma_j & \delta \le ma_j \\ \delta & ma_j \le \delta \le mb_j \\ mb_j & \delta \ge mb_j \end{cases}, \tag{38}$$

Note that, from (38), we have another interpretation for δ as a constant that normalizes the sum of the probabilities to equal one. By summing the probabilities, p_j^*, and using sensitivity analysis, we can find the value of δ at which the sum is equal to one. Equation (38) is thus determined solely by the upper bounds; lower bounds; and a normalizing parameter, δ. Furthermore, we have seen that (38) is valid for any additive convex function in (3). In the next sections, we will present several implications of this result, but first we summarize our main conclusions about the maximum entropy solution with bounds on discrete probabilities below.

Proposition 1: Invariance of the Maximum Entropy Solution with Probability Bounds

The maximum entropy distribution between upper and lower bounds on discrete probabilities is invariant to any additive concave function of the probabilities.

The invariance of the maximum entropy solution presented above provides additional justification for the use of the maximum entropy distribution with bounds on discrete probabilities, since the solution is invariant to any information measure that is an additive concave function.

Freund and Saxena [11] presented an algorithm to solve the maximum entropy problem with bounds on discrete probabilities. In this section, we have shown that this problem is invariant to any additive concave function. This observation simplifies the analysis of the maximum entropy formulation. For example, we can maximize the objective function, $f(p) = \sum_{i=1}^{n} -p_i^2$ instead of the function $f(p) = \sum_{i=1}^{n} -p_i \ln(p_i)$ in these types of problems and obtain the same result.

Bounds on Cumulative Probabilities

In many situations that arise in practice, such as in probability encoding in decision analysis, we assess a cumulative probability distribution for a variable of interest rather than a discrete event probability for each of its outcomes (Spetzler, C.S. and

Von Holstein [12]). In these situations, the decision maker may provide upper and lower bounds during the probability assessment, or he may consult an expert who provides another cumulative distribution for the variable of interest. Faced with two cumulative distributions, or upper and lower cumulative probability bounds, we are interested in the maximum entropy distribution that lies between them. The mathematical formulation for the problem is

$$p^* = \arg\min_p f(p)$$

$$\textit{such that} \qquad (39)$$

$$e^T p = 1, p \geq 0$$

$$ma \leq H(\theta)p \leq mb,$$

where $H(\theta)$ is an upper (or lower) triangular matrix whose non-zero elements are equal to one.

Let us now discuss (39) in more detail when $H(\theta)$ is a lower triangular matrix. The upper and lower bounds on cumulative probabilities can be written on an element-by-element basis as

$$ma_1 \leq p_1 \leq mb_1 \qquad (40)$$

$$ma_2 \leq p_1 + p_2 \leq mb_2 \text{ ...etc.} \qquad (41)$$

Equations (40) and (41) also provide upper and lower bounds on p_2 as follows

$$\max(ma_2 - mb_1, 0) \leq p_2 \leq \min(mb_2 - ma_1, 1)\,. \qquad (42)$$

Similarly, we can reduce the remaining cumulative probability constraints in (39) into upper and lower bounds on the discrete probabilities, where

$$\max(ma_j - mb_{j-1}, 0) \leq p_j \leq \min(mb_j - ma_{j-1}, 1), \quad j = 1,\ldots,m\,. \qquad (43)$$

Necessary Conditions for the Upper and Lower Bounds

Since the upper and lower bounds on each discrete probability must lie between zero and one, the bounds obtained in (43) must satisfy

$$ma_j - mb_{j-m} \leq 1, \ \forall m \leq j \leq n \qquad (44)$$

$$mb_j \text{ - } ma_{j-m} \geq 0, \ \forall m \leq j \leq n \qquad (45)$$

Equations (44) and (45) are necessary conditions for feasibility. Furthermore, since a cumulative distribution function is non-decreasing, we can assume, with no loss of generality, that both upper and lower bounds are monotonically non-decreasing, i.e

$$lb_{j-m} \le lb_j \text{ and } ub_{j-m} \le ub_j, \quad \forall m \le j \le n. \tag{46}$$

The normalization of the cumulative distribution over a sufficiently large support also requires

$$\begin{aligned} ma_0 &= mb_0 = 0 \\ ma_n &= mb_n = 1 \end{aligned} \tag{47}$$

Sufficient Conditions for the Upper and Lower Bounds

Equations (45), (44), (46), and (47) are sufficient conditions for the existence of a unique solution to (39).

Proof. Note that if (44), (45), (46), and (47) apply, then we can reduced the problem of bounds on cumulative probabilities into that of bounds on discrete probabilities discussed earlier. As a result, the solution to this problem is also invariant to any additive concave function of probability. We summarize this result below.

Proposition 2: Invariance of the Maximum Entropy Solution with Cumulative Probability Bounds

The maximum entropy distribution between upper and lower cumulative probability bounds is invariant to any additive concave function of the probabilities.

Proposition 2 allows for a simple graphical solution for the maximum entropy distribution between upper and lower cumulative probability bounds. We illustrate this result below.

Graphical Solution between Two Cumulative Distributions

The invariance of the maximum entropy solution described above allows us to minimize the additive convex objective function, $f(p) = \sum_{i=1}^{n} \sqrt{p_i^2 + \Delta^2}$, instead of $f(p) = \sum_{i=1}^{n} p_i \ln(p_i)$ when bounds on cumulative probabilities are present. In this new objective function, Δ is the discretizing interval for the variable, θ, for which successive cumulative probability assessments are made.

With no loss of generality, we will now focus on the maximum entropy distribution between upper and lower bounds of two cumulative probability distributions R and Q, since they satisfy (45), (44), (46), and (47). We start with the case where there is stochastic dominance between the two distributions.

a) The Two Distributions, R and Q, have Stochastic Dominance

Observe that the new objective function, $f(p)=\sum_{i=1}^{n}\sqrt{p_i^2+\Delta^2}$, is the distance (path) in the plane of the cumulative probability distribution starting from point A to point B (Figure 2). As a result, (39) is equivalent to the problem of finding the shortest path in the plane that lies between the two distributions. To find this shortest path, imagine pins in the plane at the points (θ_i, R_i) and $(\theta_i, Q_i)\, for\ 0 \le i \le n,$ where θ_i is the value of the variable at which the assessment took place, and R_i, and Q_i are the value of the cumulative distributions at θ_i. Now thread a string between the pins at (θ_i, R_i) and (θ_i, Q_i) for $0 \le i \le n,$ and pull the string taut. The taut string traces out the shortest path and is also the maximum entropy solution. Note that the taut string does not have to be linear but takes the shape of the shortest path, which could be one of the bounds themselves. We summarize this result below.

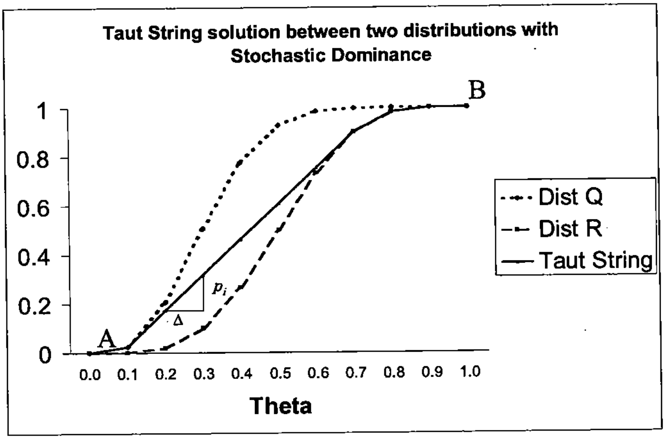

FIGURE 2. Graphical determination of maximum entropy solution between upper and lower cumulative bounds.

Proposition 3: Maximum Entropy Solution is the Shortest Path Solution

The maximum entropy distribution between upper and lower cumulative probability bounds is the shortest path in the plane that lies between the upper and lower bounds.

We note that taut string solutions exist in many applications in practice. For example, Modigliani and Hohn [13] discuss a string solution for problems in production planning over time; Dantzig [14] discusses a string solution for special cases in control theory, and Veinott [15] discusses a taut string solution for optimal flow in network problems.

b) Stochastic Dominance Does Not Exist

When stochastic dominance does not exist between the two distributions, we may be interested in the maximum entropy distribution between their upper bound, $\max(R_i, Q_i)$, and lower bound, $\min(R_i, Q_i)$. These bounds also satisfy (45), (44), (46), and (47). When stochastic dominance does not exist, the two distributions will cross at least once. The crossing points comprise both upper and lower bounds at that value of θ so the taut string must pass through them. Figure 3 the maximum entropy distribution between two distributions with no stochastic dominance.

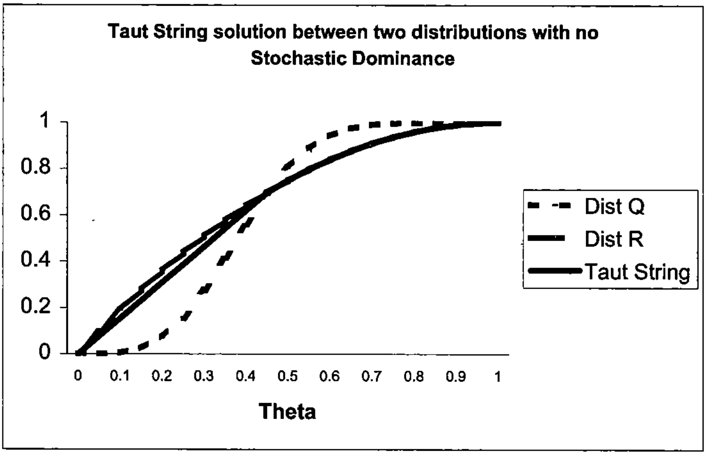

FIGURE 3. The maximum entropy distribution between two distributions with no stochastic dominance.

c) Maximum Entropy Distribution with Cumulative Probability Constraints

As a special case of the graphical solution discussed above, we refer to the maximum entropy distribution given precise cumulative probability constraints. As we noted earlier, we can think of equality constraints as upper and lower bounds that coincide. This implies that the taut string must pass through them. When the string is taut, the solution traces out the well known piece-wise linear cumulative distribution that is also invariant to any additive concave function of probability.

MAXIMUM ENTROPY JOINT DISTRIBUTION WITH BOUNDS ON MARGINAL PROBABILITIES

Now we present another application of (38), and consider the maximum entropy joint distribution given bounds on the marginal probabilities of the variables. For example, the maximum entropy formulation for a three-variable joint distribution in terms of bounds on its marginal probabilities is

$$\max \ -\sum_{i,j,k} p_{ijk} \ln(p_{ijk})$$

such that

$$ma_i \leq \sum_{j,k} p_{ijk} \leq mb_i \ \forall i, \ ma_j \leq \sum_{i,k} p_{ijk} \leq mb_j \ \forall j \qquad (48)$$

$$ma_k \leq \sum_{i,j} p_{ijk} \leq mb_k \ \forall k, \ p_{ijk} \geq 0 \ \forall i, j, k,$$

$$\sum_{i,j,k} p_{ijk} = 1,$$

where the subscripts i ,j ,k represent the i^{th} outcome of the first variable, the j^{th} outcome of the second variable, and the k^{th} outcome of the third variable respectively; p_{ijk} is the joint probability of i ,j ,k; ma_i, ma_j, ma_k are lower bounds; and mb_i, mb_j, mb_k are upper bounds.

We assume, with no loss of generality, that the upper and lower bounds range from zero and one. Furthermore, we define lb_n and ub_n as the sum of the lower and upper bounds for each variable n respectively. For example, $lb_1 = \sum_i ma_i$, $ub_1 = \sum_i mb_i$, $lb_2 = \sum_j ma_j$,...etc.

It is well known in information theory that the maximum entropy joint distribution given its marginal distributions is equal to the product of the marginal distributions (see for example Cover and Thomas [16]). The rationale for this assignment is that the marginal constraints do not contain any information about the dependence relations between the variables present. Furthermore, properties of the entropy expression show that the value of this maximum entropy is equal to the sum of the entropies of the marginal distributions.

When bounds on the marginal probabilities are available, the maximum entropy joint distribution is equal to the product of the maximum entropy marginal distributions given bounds on their marginal probabilities. We can now use (38) to determine the maximum entropy joint distribution with bounds on marginal probabilities. First we present the necessary and sufficient conditions for the existence of a feasible solution to this problem.

Proposition 4: Necessary and Sufficient Conditions

The conditions $lb_n \leq 1$ and $ub_n \geq 1 \ \forall n$ are necessary and sufficient for the existence of a unique solution to (48).

Proof. The proof of necessity is straightforward; if either $lb_n \geq 1$ or $ub_n \leq 1$, it will lead to solution whose marginal probabilities do not sum to one. The proof of sufficiency is more involved. From (38), we can see that the marginal probability for the outcomes of each variable, *n*, is equal either to either its lower bound; upper bound; or a normalizing constant, δ_n. Note also that the sum of marginal probabilities in (38) is a monotonically non-decreasing function of δ_n. For example, when $\delta_n = 0$, the sum of marginal probabilities for variable *n* is equal to the sum of lower bounds, and when $\delta_n = 1$, the sum of marginal probabilities for variable *n* is equal to the sum of the upper bounds. Monotonicity of (38) implies there is a value of δ_n for which the sum of probabilities is equal to one. Using sensitivity analysis, we can determine this value of δ_n such that the marginal probabilities of variable *n* sum to one.

The proof of the previous lemma provides a convenient method to solve (48) graphically. We illustrate this result using the following example.

Example: We are interested in the maximum entropy joint distribution for three variables, each discretized to three outcomes, and whose marginal probabilities have the following lower and upper bounds (Table 1). We use the notation $p_{i..}$ for the marginal probability of the i^{th} outcome of the first variable; $p_{.j.}$ for the marginal probability of the j^{th} outcome of the second variable; and $p_{..k}$ for the marginal probability of the k^{th} outcome of the third variable.

TABLE 1. Upper and lower bounds on marginal probabilities.

	P1..	P2..	P3..	P.1.	P.2.	P.3.	P..1	P..2	P..3
ma	0.05	0.45	0.35	0.016	0.39	0.192	0.28	0.34	0.2
mb	0.17	0.6	0.8	0.25	0.5	0.592	0.4	0.37	0.3

The first step is to check feasibility of the lower bounds,

$$\begin{aligned} lb_1 &= 0.05 + 0.45 + 0.35 = 0.85 < 1 \\ lb_2 &= 0.016 + 0.39 + 0.192 = 0.597 < 1 \\ lb_3 &= 0.28 + 0.34 + 0.2 = 0.82 < 1 \end{aligned} \tag{49}$$

The second step is to check the feasibility of the upper bounds,

$$\begin{aligned} ub_1 &= 0.17 + 0.16 + 0.8 = 1.57 > 1 \\ ub_2 &= 0.25 + 0.5 + 0.592 = 1.342 > 1 \\ ub_3 &= 0.4 + 0.37 + 0.3 = 1.07 > 1 \end{aligned} \tag{50}$$

From Lemma 1, we know that there exists a unique solution to this problem. The sum of marginal probabilities for each variable, *n*, can be plotted vs. δ. The value at which the sum is equal to one determines the value of δ_n for that variable. The marginal probability for each outcome of variable, *n*, can now be determined using

(38) and the value of δ_n. Figure 4 shows a sensitivity analysis to the sum of marginal probabilities for each variable, n, and the corresponding values of δ_n.

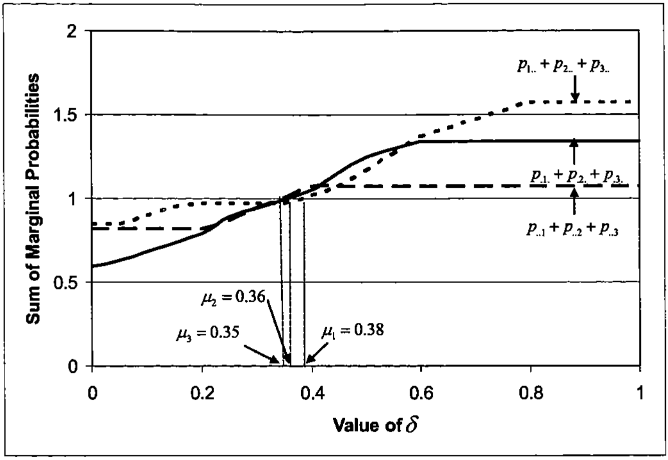

FIGURE 4. Sensitivity analysis to the value of δ_n.

We now use (38) to determine the maximum entropy marginal probabilities as shown below.

TABLE 2. Maximum entropy marginal probabilities.

P1..	P2..	P3..	P.1.	P.2.	P.3.	P..1	P..2	P..3
0.17	0.45	0.38	0.25	0.39	0.36	0.35	0.35	0.3
ub	*lb*	δ_1	*ub*	*lb*	δ_2	δ_3	δ_3	*ub*

The maximum entropy joint distribution is equal to the product of the marginal distributions.

CONCLUSIONS

We have shown that the maximum entropy solution with bounds on discrete probabilities and bounds on cumulative probabilities is invariant to any additive concave objective function. This observation simplifies the analysis of the problem, by choosing simpler objective functions, and unifies the results of several concave information measures for these situations. We used this invariance result to provide a graphical solution to the maximum entropy distribution between upper and lower cumulative probability bounds. We also discussed the maximum entropy joint distribution given bounds on the marginal probabilities, and illustrated how the formulation can be decomposed into a simpler formulation with bounds on discrete probabilities using properties of the entropy expression.

APPENDIX: CONJUGATE OF NEGATIVE OF THE ENTROPY

The conjugate of the negative of the entropy expression is by definition

$$f^*(y^*) \triangleq \sup_p [p^T(\mu e + H(\theta)^T(\alpha - \beta)) - f(p)] \qquad \text{such that } e^T p = 1, \tag{51}$$

where $f(p) = \sum_{i=1}^{n} p_i \ln(p_i)$. To solve (51) we evaluate the Lagrange operator as

$$L(p, y^*, \lambda) = p^T y^* - f(p) - \lambda(e^T p - 1) \tag{52}$$

The first partial of (52) with respect to p gives

$$\nabla f(p^*) = y^* - \lambda e \tag{53}$$

Re-arranging (53) and substituting for the gradient of the negative of the entropy expression, $\frac{\partial f(p)}{\partial p_i} = 1 + \ln(p_i)$, for each element p_i gives

$$p_i^* = e^{y_i^* - 1 - \lambda} \qquad i = 1,..,n \tag{54}$$

The normalizing constraint in (51) gives

$$\sum_{i=1}^{n} p_i^* = \sum_{i=1}^{n} e^{y_i^* - 1 - \lambda^*} = 1 \tag{55}$$

Re-arranging (55) gives

$$\lambda^* = \ln(\sum_{i=1}^{n} e^{y_i^* - 1}) \tag{56}$$

Substituting (54) and (56) into (52) gives

$$\begin{aligned} L(p^*, y^*, \lambda^*) &= p^T y^* - f(p) \\ &= \sum_{i=1}^{n} p_i y_i^* - \sum_{i=1}^{n} p_i \ln(p_i) \\ &= \sum_{i=1}^{n} p_i y_i^* - \sum_{i=1}^{n} p_i (y_i^* - 1 - \lambda^*) \\ &= \sum_{i=1}^{n} p_i (1 + \lambda^*) = (1 + \lambda^*) \end{aligned} \tag{57}$$

Substituting for λ^* from (56) gives

$$\begin{aligned} L(p^*, y^*, \lambda^*) &= 1 + \ln(\sum_{i=1}^{n} e^{y_i^* - 1}) \\ &= 1 + \ln(e^{-1}) + \ln(\sum_{i=1}^{n} e^{y_i^*}) \\ &= \ln(\sum_{i=1}^{n} e^{y_i^*}) \\ &\triangleq f^*(y^*) \end{aligned} \tag{58}$$

The conjugate of the negative of the entropy expression is thus $f^*(y) = \ln(\sum_{i=1}^{n} e^{y_i})$.

Q.E.D.

REFERENCES

1. Jaynes, E.T. "Information theory and statistical mechanics". *Phys. Rev.*, 108:171, (1957).
2. Berger, A, Della Pietra, S and Della Pietra, V. "A maximum entropy approach to natural language processing". *Computational Linguistics*, 22,1, (1996).
3. Baldi, P and Brunak, S. 1998. *Bioinformatics: The Machine Learning Approach,* MIT Press , Boston, MA. (1998).
4. Tribus, M. *Rational Descriptions, Decisions and Designs*, Pergamon Press, Elmsford, NY, (1969)
5. Singh, V.P. *Entropy-Based Parameter Estimation*. Kluwer Academic Publishers. Netherlands. (1998).
6. Thomas, M. 1979. "A generalized maximum entropy principle". *Operations Research*. 27. 6.
7. Abbas, A. E. *Entropy Methods in Decision Analysis*. PhD dissertation, Stanford University. (2003).
8. Abbas, A.E. "Entropy Methods for Joint distributions in Decision Analysis". Forthcoming in *IEEE Transactions on Engineering Management.* (2005)
9. Nash, S. G, and Sofer, A. *Linear and nonlinear programming.* McGraw-Hill. (1996)
10. Rockafeller, R.T. *Convex Analysis*. Princeton University Press. Princeton, N.J. (1970)
11. Freund, D and Saxena, U. "An algorithm for a class of discrete maximum entropy problems". *Operations Research,* 32, 1. (1984)
12. Spetzler, C.S. and Von Holstein . "Probability encoding in decision analysis". *Readings on the Principles and Applications of Decision Analysis*, vol.2, Strategic Decisions Group, Menlo Park, CA. (1972)
13. Modigliani, F, and Hohn, F. "Production planning over time and the nature of the expectation and planning horizon", *Econometrica*, 23, 253-292.(1955)
14. Dantzig, D. "A control problem of Bellman". *Management Science*, 17, 9, 542-546. (1971).
15. Veinott, A. F. "Least d-majorized network flows with inventory and statistical applications". *Management Science*, 17, 9. 547-567. (1971)
16. Cover, T, and Thomas, J. *Elements of Information Theory.* Wiley, New York. (1991)