Finite Dimensional Entropy Optimization, a Review

Razek Karnoub

December 12, 1995

Abstract

This paper reviews the basic entropy optimization principles and discusses several finite dimensional forms they might take. Methods to approach the optimal solution of those entropic forms and some applications to the Linear and Non-Linear Programming problems are summarized.

1 Introduction

The word entropy originated in the Thermodynamics literature about 150 years ago to represent a measure of the amount of energy in a thermodynamic system as a function of temperature and the heat that enters the system. The word belonged to the domain of physicists until 1948 when Claude Shannon working on his Theory of Communication coined the term, upon the suggestion of Von Neumann, to represent a measure of information he was developing[1]. With that, the concept of entropy penetrated a very wide range of disciplines. These include Statistical Mechanics, Thermodynamics, Statistical Inference, Economics, Business and Finance, Non-Linear Spectral Analysis, Pattern Recognition, Transportation, Urban and Regional Planning, Queueing Theory, Parameter Estimation and Linear and Non-Linear Programming, the focus of this review. It is worth noting that at the time Shannon introduced his concept of entropy no relationship, except for the similar mathematical expressions, was known between the Shannon entropy

[1] See Chapter 1 of Kapur and Kesavan and the references contained therein.

and the Thermodynamics entropy. The relationship was only established later[2].

The concept of entropy is closely tied to the concept of uncertainty embedded in probability distributions. In fact, entropy can be defined as a measure of probabilistic uncertainty. For example, suppose the probability distribution of the outcome of a coin flipping experiment is $(0.0001, 0.9999)$, 0.0001 being associated with a tail outcome. Here, one is likely to notice that there is much more "certainty" attached to the outcome of this experiment than "uncertainty": we are almost certain that the outcome will be a head. If, on the other hand, the probability distribution attached to that same experiment were $(0.5, 0.5)$, one would realize that there is less "certainty " and much more "uncertainty" involved. Generalizing this observation to the case of n possible outcomes, we conclude that the uniform distribution has the highest uncertainty out of all possible probability distributions. This implies that if we had to choose a probability distribution for a chance experiment without any prior knowledge about that distribution, it would seem reasonable to pick the uniform distribution because we have no reason to choose any other and because that distribution maximizes the "uncertainty" of the outcome. This is called Laplace's Principle of Insufficient Reason [12]. Note that we were able to justify it without resorting to a rigorous definition of "uncertainty". But what if we had some prior knowledge of the distribution? Suppose, for example, that we know of some constraints that the moments of that distribution have to satisfy. In that case a mathematical description of "uncertainty" is inescapable. It is here that Shannon measure of uncertainty, or entropy as he called it, plays its role.

To define that entropy, Shannon stated some axioms that he thought any measure of information should satisfy and deduced a unique function, up to a multiplicative constant, that satisfies them. It turned out that this function actually possesses many more properties that make it desirable and, in later years, many researchers modified and replaced those axioms in an effort to simplify them. They still, however, deduced that same function. Among those defining axioms, Kapur and Kesavan stated the following [12]:

> Let $(p_1, p_2, \ldots, p_n)$ be a probability distribution and S its entropy. Then

[2] For a detailed treatment of this relationship, see Chapter 3 of the book by Kapur and Kesavan.

1. S should be a function of all the p_i's.
2. S should be a continuous function of the p_i's.
3. S should be permutationally symmetric, i.e. if the p_i's are permuted then S remains the same.
4. $S(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$ should be a monotonic increasing function of n.
5. $S(p_1, p_2, \dots, p_n) = S(p_1+p_2, p_3, \dots, p_n)+(p_1+p_2)\, S(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$.

Properties $1, 2$ and 3 are obvious. Property 4 states that the maximum uncertainty of a probability distribution should increase as the number of possible outcomes increases. Property 5 is the least obvious but states that the uncertainty of a probability distribution is the sum of the uncertainty of the probability distribution that combines two of the outcomes and the uncertainty of the probability distribution consisting of only those two outcomes adjusted by the combined probabilities of the outcomes.

As it turns out, the unique family of functions that satisfy those properties is: $-k \sum\limits_{i=1}^{n} p_i \ln p_i$, where k is a positive constant [12]. Shannon chose $-\sum\limits_{i=1}^{n} p_i \ln p_i$ to represent his entropy. Among its many desirable properties we state the following:

(i) Shannon's measure is non-negative and concave in $p_1, \dots, p_n$.

(ii) The measure does not change with the inclusion of a zero probability event.

(iii) The entropy of the probability distribution representing a completely certain outcome is 0 and the entropy of the probability distribution representing uncertain outcomes is positive.

(iv) The maximum possible entropy is that of the uniform distribution.

(v) The entropy of the joint distribution of two independent distributions is the sum of the individual entropies.

(vi) The entropy of the joint distribution of two dependent distributions is less than or equal to the sum of the individual entropies.

Property (i) is desirable because it is much easier to maximize a concave function than a non concave one. Properties (ii) and (iii) are appealing because a zero probability event contributes nothing to uncertainty and neither does a completely certain outcome. Property (iv) was discussed earlier and properties (v) and (vi) state that joining two distributions does not affect the entropy if they are independent and may actually reduce it if they are not.

Back to the problem of choosing a probability distribution for a chance experiment knowing some constraints that the moments of that distribution have to satisfy. We said earlier that in the absence of moment constraints, the best course of action is to choose the distribution that maximizes the uncertainty. It seems natural then, now that we can mathematically describe that uncertainty, to generalize that reasoning and choose the distribution that maximizes uncertainty subject to the given moment constraints. In this way, we are making use of all the information given to us and are avoiding assumptions about information that is not available. This is what E.T. Jaynes argued[3]. He suggests that if we choose a distribution other than the one with the maximum entropy then the reduction in entropy can only come from some additional information we may have used. This argument provides the rational for enunciating the Principle of Maximum Entropy, also known as MaxEnt or Jaynes' Maximum Entropy Principle:

> Out of all possible distributions consistent with the moment constraints choose the one that has the maximum uncertainty.

Mathematically, let X denote the random variable we are trying to find a probability distribution for, let $x_1, x_2, \ldots, x_n$ denote the values it takes with probabilities $p_1, p_2, \ldots, p_n$, respectively, and let $g_1(X), g_2(X), \ldots, g_m(X)$ be functions of X with expected values $E[g_1(X)] = a_1$, $E[g_2(X)] = a_2$, $\ldots$, $E[g_m(X)] = a_m$. Then the problem can be stated as :

$$\text{Max } -\sum_{i=1}^{n} p_i \ln p_i$$

subject to

$$\sum_{i=1}^{n} p_i = 1$$

$$\sum_{i=1}^{n} p_i\, g_r(x_i) = a_r,\ r = 1, \ldots, m$$

$$p_i \geq 0,\ i = 1, \ldots, n.$$

[3]See Chapter 2 of the book by Kapur and Kesvan and the references contained therin.

a concave programming problem with linear constraints where we would like the matrix $[g_r(x_i)]$ to have full rank. Note here that the non-negativity constraints are not binding for the optimal solution as each p_i^* can be expressed as an exponential in terms of the Lagrange multipliers of the problem restricted to the equality constraints. Note also that in the absence of the moment constraints, the solution to the problem is the uniform probability distribution with entropy $\ln n$. As such, the Maximum Entropy Principle can be seen as an extension of Laplace's Principle of Insufficient Reason.

Recall that the whole discussion of the Maximum Entropy Principle was motivated by having to choose a probability distribution with known moment constraints. Suppose now that in addition to those constraints we have an a priori probability distribution, q, that we think our probability distribution, p, should be close to. In fact in the absence of those constraint we might choose q for p. In that case, we would want the probability distribution that is "closest" to our a priori distribution and that at the same time satisfies the constraints. But to be able to do that we need a precise definition of what "close" is. In other words, we need to define some "sort of distance" or more precisely "directed divergence" [12] on the space of discrete probability distributions that we are dealing with. Notice that the we are deliberately avoiding calling this measure a "distance". For a distance measure should be symmetric and that is not crucial in this case: we can be content with a "one way distance", $D(p,q)$, from p to q. In fact, if a "one way distance" from p to q is not satisfactory, consider defining a symmetric measure as the sum of $D(p,q)$ and $D(q,p)$. A distance measure should also satisfy the triangular inequality and that is not extremely important in this context either. What is important for this "directed divergence" measure is the following:

1. $D(p,q)$ needs be non-negative for all p and q.
2. $D(p,q) = 0$ if and only if $p = q$.
3. $D(p,q)$ needs to be a convex function of $p_1, p_2, \ldots, p_n$.
4. When $D(p,q)$ is minimized subject to moment constraints and without the explicit presence of the non-negativity constraints, the resulting p_i's should be non-negative.

Property 1 is obvious: it assures that our measure is bounded from below. If property 2 were not satisfied then it would be possible to choose a p with zero directed divergence from q, i.e. as "close" to q as q itself yet different from q: a complicating factor that we would rather avoid. Property 3 makes minimizing the measure much simpler and property 4 spares us from explicitly considering n more constraints. Fortunately, there are many measures that satisfy those properties. We may even be able to find one that satisfies the triangle inequality. But simplicity of the measure is even more desirable. The simplest and most important of those measures is the Kullback-Liebler measure[4], also known as the measure of cross entropy: $D(p,q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$, where we assume that whenever q_i is 0 then p_i is 0 and $0 \ln \frac{0}{0}$ is defined as 0. In addition to its defining properties we state the following:

(i) $D(p,q)$ is a continuous function of p and q.

(ii) $D(p,q)$ is permutationally symmetric, i.e. the measure does not change if the pairs of (p_i, q_i) are permuted among themselves.

(iii) $D(p,q)$ is convex in both p and q.

(iv) $D(p,q)$ is not symmetric.

(v) If $p_1$and p_2 are independent and so are $q_1$and q_2 then

$$D(p_1 * p_2, q_1 * q_2) = D(p_1, q_1) + D(p_2, q_2)$$

where $p * q$ denotes the convolution of p and q.

(vi) In general, the Triangle Inequality does not hold but if distribution p minimizes $D(p,q)$ subject to some moment constraints and r is any other distribution that satisfies those same constraints then

$$D(r,q) = D(r,p) + D(p,q).$$

Thus, in this special case the Triangle Inequality holds as an equality.

With this definition of "closeness" the Kulback-Leibler Minimum Cross Entropy Principle, or MinxEnt, can be enunciated as follows [12]:

[4]See Chapter 4 of the book by Kapur and Kesavan.

> Out of all possible distributions consistent with the moment constraints choose the one that is closest to the given a priori distribution.

Mathematically,

$$\text{Min} \sum_{i=1}^{n} p_i \ln \tfrac{p_i}{q_i}$$

subject to

$$\sum_{i=1}^{n} p_i = 1$$

$$\sum_{i=1}^{n} p_i \, g_r(x_i) = a_{r,} \; r = 1, \ldots, m$$

$$p_i \geq 0, \; i = 1, \ldots, n.$$

Note that the non-negativity constraints are not binding as in the MaxEnt problem and for the same reason.

Now, if q is not given, one may think of trying the distribution that has the maximum uncertainty in its place. This is the uniform distribution, u. In that case,

$$D(p,q) = D(p,u) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{1/n} = \ln n + \sum_{i=1}^{n} p_i \ln p_i.$$

Since minimizing $\sum_{i=1}^{n} p_i \ln p_i$ is equivalent to maximizing $-\sum_{i=1}^{n} p_i \ln p_i$ we see that minimizing the closeness to the uniform distribution is equivalent to maximizing uncertainty and so MaxEnt is a special case of MinxEnt. Those two principles can now be combined into a general principle:

> Out of all probability distributions satisfying the given moment constraints choose the distribution that is closest to the given a priori distribution and in the absence of it choose the distribution that is closest to the uniform distribution.

It is worthwhile noting here that the MaxEnt and MinxEnt principles could be generalized to the case of continuous probability distributions. The objectives will be functionals and the constraints will be functional constraints, though. Also, special cases of those problems, continuous or discrete, lead to widely known distributions that have been discovered independently of MaxEnt or MinxEnt.

In the remainder of this paper, we review several types of entropy maximization and cross entropy minimization problems, the methods that are employed to solve them and applications to the Linear and Non-Linear Programming problems.

2 Problems and Algorithms

The first problem discussed is that of maximizing the entropy subject to linear constraints, equality and inequality. The second problem is an LP attached to an entropic constraint. The third is the Linear Programming problem solved via entropic perturbation. The fourth is a cross-entropy problem with cross-entropic constraints. The fifth is a cross-entropy problem with convex constraints and the last is an application of entropy optimization to solve the Non-Linear Programming problem.

2.1 Linearly Constrained Entropy Maximization

In their paper, Censor *et al* [3], considered the following problem:

$$\begin{aligned}
&\text{Max } -\sum_{i=1}^{n} x_i . \ln x_i \\
&\text{subject to} \\
&\sum_{i=1}^{n} x_i = 1 \\
&x \in Q \\
&x_i \geq 0,\, i = 1, \ldots, n.
\end{aligned}$$

where $Q \subset \mathbf{R}^n$ is a constraints set of the types:

$$\begin{aligned}
Q_1 &= \{x \in \mathbf{R}^n : A\,x = b\} \\
Q_2 &= \{x \in \mathbf{R}^n : A\,x \leq b\} \\
Q_3 &= \{x \in \mathbf{R}^n : c \leq A\,x \leq b\}
\end{aligned}$$

The algorithms they presented to solve those problems are suitable for large sparse systems. They are "row-action" algorithms in the sense that during each iteration only one row of the constraint matrix is used; eventually, of course, all the rows are expected to be used. The typical choice of rows for the different iterations of a row-action algorithm, also called control strategy, is the following: if k is the iteration number then the row to be used is row $i(k) = 1 + k \bmod m$, m being the numbers of rows of A. In fact, any other cyclical strategy for choosing the rows can be reduced to it by, simply, renumbering the rows. We will adopt this control strategy for all the row-action algorithms described in this section. Additionally, the idea behind several of those algorithms is to take the hyperplane determined by a constraint, "project" onto it the current iterate and use that "projection" to decide what the next iterate should be. The "projections" used are not the simple orthogonal projections but the so called "entropic-projections", special cases of a more general type of projections developed by Bregman for the Convex Programming problem called "D-projections" [2]. We note that the D-projection of a point onto a hyperplane is determined not only by the point and hyperplane but also by the objective function being considered and that in the case where the objective is the Entropy function, the D-projection is called entropic-projection and may be defined in this way:

> Given $y \in \mathbf{R}^n_+$ and a hyperplane $H = \{x \in \mathbf{R}^n :< a, x >= b\}$, the entropic projection of y onto $H, \hat{P}_H(y)$, is the point $y' \in \mathbf{R}^n$ such that
>
> $$\begin{aligned} y'_j &= y_j \exp(B\, a_j) \\ < a, y' > &= b \end{aligned}$$
>
> for some $B \in \mathbf{R}$.

Here, it can be shown that y' and B exist and are unique [5]; whence, the real number B is called the entropy projection coefficient associated with projecting y onto H. It can also be shown that if H_1 and H_2 are two parallel hyperplanes then $\hat{P}_{H_2}(\hat{P}_{H_1}(y)) = \hat{P}_{H_2}(y)$ and $B_2 = B_1 + \bar{B}_2$, where B_i is the entropic projection coefficient associated with projecting y onto $H_i, i = 1, 2$, and $\bar{B}_2$ is the entropy projection coefficient associated with projecting $\hat{P}_{H_1}(y)$ onto H_2.

In the subsections that follow, descriptions of the iterates generated are given. The algorithms could be stopped when the norm of the difference of two successive iterates is less than a prescribed acceptable error.

2.1.1 Algorithms for Entropy Maximization over Linear Equality Constraints

Two algorithms are mentioned. The first is called MART for Multiplicative Algebraic Reconstruction Technique and dates back to 1970 [11] at which time it was proposed as a heuristic for image reconstruction. It updates the iterate x^k as follows:

$$x_j^{k+1} = x_j^k \left(\frac{b_i}{< a^i, x^k >} \right)^{a_j^i}, j = 1, 2, \ldots, n$$

where a^i denotes the i^{th} column of A, i being the $i(k)$ defined previously.
Convergence of this algorithm, to the unique solution, was first proven by Lent [13] provided $Q_1 \cap \mathbf{R}_+^n \neq \emptyset, a_j^i \in [0,1], b_i > 0$ and $x_j^0 = e^{-1}$ for all i and j; conditions that are not too restrictive in many applications as the data can preprocessed to fit the requirements. In his proof, Lent formulates the ordinary programming dual problem, obtains dual-primal conversion formulas and shows that the above changes in the primal variable are induced by changes in the dual variable made to optimize the dual objective.

The second algorithm is a special case of a general scheme developed by Bregman for the Convex Programming problem [2]. Its convergence is assured under the assumption that all iterates stay in $\mathbf{R}_+^n$. Those iterates are constructed by entropic projections:

$$x^{k+1} = \hat{P}_{H_i}(x^k)$$

where $i = i(k)$ and $H_i = \{x \in \mathbf{R}^n :< x, a^i >= b_i\}$.
It may be worthwhile mentioning that to compute x^{k+1} in the second algorithm, we can first determine B by solving the equation

$$\sum_{j=1}^{n} a_j^i x_j^k \exp(B a_j^i) = b_i$$

and use it to obtain x^{k+1}. Another important point from a theoretical point of view, one that is easy to verify and that could possibly help shed some light on the relationship between those two algorithms is that if A consists of 0's and 1's then the two algorithms produce the same iterates; given, of course, that they are started at the same initial point. The case of the arbitrary A has reportedly been investigated by Censor, De Pierro *et al* [4].

2.1.2 Algorithms for Entropy Maximization over Linear Inequality Constraints

Again, two algorithms are considered. The first is due to Bregman too [2]:

$$
\begin{aligned}
x_j^{k+1} &= x_j^k \exp(c_k a_j^i), j = 1, 2, \ldots, n \\
z^{k+1} &= z^k - c_k e^i
\end{aligned}
$$

where $c_k \doteq \min(z_i^k, B_k)$, B_k is the entropy projection coefficient associated with projecting x^k onto H_i, as defined previously, and e^i is the i^{th} unit vector in $\mathbf{R}^m$.

Note that if $c_k = B_k$ then $x^{k+1} = \hat{P}_{H_i}(x^k)$ and that if $c_k = z_i^k$ then $x^{k+1} = \hat{P}_{\bar{H}_i}(x^k)$ where $\bar{H}_i$ and H_i are parallel.

The second algorithm is due to Censor *et al* [3]. Its advantage is that it eliminates the need to compute B_k:

$$
\begin{aligned}
x_j^{k+1} &= x_j^k \exp(g_k a_j^i), j = 1, 2, \ldots, n \\
z^{k+1} &= z^k - g_k e^i
\end{aligned}
$$

where $g_k \doteq \min(z_i^k, M_k)$, $M_k = \ln \frac{b_i}{<a^i, x^k>}$ and $i = i(k)$.

Aside from the apparent relief from computing B_k, in the second algorithm, no mention of how those two algorithms compare in practice could be found.

2.1.3 Algorithm for Entropy Maximization over Linear Interval Constraints

The study of this problem is motivated by practical situations where the consistency of the system $A\,x = b$ cannot be guaranteed due, for example, to measurement errors or idealizing assumptions. This forces the replacement of the linear equations with the linear interval constraints: $b - \varepsilon \leq A\,x \leq b + \varepsilon$. One way to approach the problem is to transform it to a problem with linear inequalities. But the next algorithm, also due to Censor *et al* [3], deals with this situation more efficiently:

$$
\begin{aligned}
x_j^{k+1} &= x_j^k \exp(h_k a_j^i), j = 1, 2, \ldots, n \\
z^{k+1} &= z^k - h_k e^i
\end{aligned}
$$

where h_k =mid $(z_i^k, \Delta_k, \Gamma_k)$, mid stands for median and Δ_k and Γ_k stand for the entropy projection coefficients associated with the entropy projection of x^k onto $\{x \in \mathbf{R}^n :< a^i, x >= b + \varepsilon_i\}$ and $\{x \in \mathbf{R}^n :< a^i, x >= b - \varepsilon_i\}$, respectively.

2.2 Entropy Constrained LP's

The problem studied in this class has the form:

$$\begin{aligned}
&\text{Min } c^T x \\
&\text{subject to} \\
&Ax = b, x \geq 0 \\
&-\sum_{j=1}^{n} x_j \ln x_j \geq H
\end{aligned}$$

where it is assumed that $\sum_{i=1}^{n} x_i = 1$ is built in the constraints and that A has full rank m.

This problem was first studied by Erlander [6]. It is a standard form LP attached to an entropy constraint introduced to keep a minimum level of spread or smoothness in the solution. In fact, it is felt in many situations that the LP model oversimplifies the problem by producing a basic solution that eliminates the contributions of many of the variables. The entropy constraint is meant to restore to the model a complexity, or perhaps uncertainty, that was lost in the process of building the model itself.

Note that if $H \leq 0$, the solution to the problem is independent of H. Also, if H exceeds $\ln n$ then a solution does not exist. Clearly, then, for the problem to be meaningful we must have H inside some interval $[H_{\min}, H_{\max}]$. It is not hard to see, in this case, that

$$\begin{aligned}
H_{\min} &= \max\{H(x) : x \text{ is optimal for Min } c^T x \text{ s.t. } Ax = b, x \geq 0\} \\
&\quad \text{and} \\
H_{\max} &= \max\{H(x) : Ax = b, x \geq 0\}.
\end{aligned}$$

The following theorem characterizes the optimal solution to this problem:
Theorem [6]: The entropy constrained problem above has for:

(i) $H \leq H_{\min}$: the optimal solution of the corresponding LP.

(ii) $H_{\min} < H < H_{\max}$: the unique regular optimal solution

$$x_j^* = \exp\{(\beta^T a^j - c_j)/\eta\}, \eta > 0, H(x^*) = H, j = 1, \ldots, n.$$

(iii) $H = H_{\max}$: the unique non-regular optimal solution

$$x_j^* = \exp\{\lambda^T a^j\}, j = 1, .., n.$$

(iv) $H_{\max} < H$: no solution.

where $\beta, \lambda \in \mathbf{R}^m, \eta$ is real and a^j is the j^{th} column of A.

We note that in cases (ii) and (iii) the optimal solution is expressed analytically in terms of the dual variables which effectively reduces the number of variables of the problem from n to $m+1$ and could motivate the search for the optimal solution of the dual problem. In fact, it can be shown that when $H_{\min} < H < H_{\max}$, the dual of our problem is:

$$\text{Max } D(\beta, \eta) = \beta^T b - \eta \sum_{j=1}^{n} \exp\{(\beta^T a^j - c_j)/\eta\} + \eta(H + 1)$$

subject to $\eta \geq 0$.

a convex programming problem with only one non-negativity constraint. Erlander [6] did not attack this problem, though. Instead, he noticed that the binding entropy constraint together with the linear constraints lead to $m+1$ equations, the same as the number of unknowns. He proposed eliminating the x variable to obtain $m+1$ equations in β and η. He, then, solved that system of equations using a Newton procedure.

2.3 Linear Programming Solution by Entropic Perturbation

Two approaches to solving the Linear Programming problem are presented. Both depend on introducing an entropic perturbation to the objective function whose sole purpose is a computational benefit; no interpretation of it in terms of the parameters of the model is meant.

The first approach, due to Fang [7] [14], considers the LP problem in Karmarkar's standard form:

Min $c^T x$

subject to

$Ax = 0$

$e^T x = 1$

$x \geq 0$

where it is assumed that A has full rank m and that an interior solution exists.
The dual of this problem has the form:

Max w_{m+1}

subject to

$\sum_{i=1}^{m} a_{ij} w_i + w_{m+1} \leq c_j, j = 1, 2, \ldots, n,$

$w_i \in \mathbf{R}, i = 1, 2, \ldots, m+1.$

Now, recall the Geometric Inequality:

If $x, y \in \mathbf{R}^n, x > 0$ and $\sum_{j=1}^{n} x_j = 1$ then $\sum_{j=1}^{n} \exp(y_j) \geq \prod_{j=1}^{n} \{\exp(y_j)/x_j\}^{x_j}$ with equality if and only if $x = \lambda \exp(y_j), j = 1, 2, .., n$, for some $\lambda > 0$.

Making use of this inequality with $y_j = (\sum_{i=1}^{m} a_{ij} w_i - c_j)/\mu, j = 1, 2, \ldots, n, \mu > 0$ and assuming that x is primal feasible, one can show that

$$-\mu \ln \left\{ \sum_{j=1}^{n} \exp \left[\left(\sum_{i=1}^{m} a_{ij} w_i - c_j \right) / \mu \right] \right\} \leq c^T x + \mu \sum_{j=1}^{n} x_j \ln x_j \tag{1}$$

with equality if and only if

$$x_j = x_j^* \doteq \frac{\exp \left[\left(\sum_{i=1}^{m} a_{ij} w_i^* - c_j \right) / \mu \right]}{\sum_{j=1}^{n} \exp \left[\left(\sum_{i=1}^{m} a_{ij} w_i^* - c_j \right) / \mu \right]}, j = 1, 2, \ldots, n \tag{2}$$

where w^* is the unique unrestricted maximizer of the left-hand side of (1), denoted by $h(w, \mu)$.

Moreover, it can be shown that

$$h(w^*, \mu) = c^T x^* + \mu \sum_{j=1}^{n} x_j^* \ln x_j^*.$$

Now letting

$$w_{m+1}^* = \min_{j=1,\ldots,n} \left\{ c_j - \sum_{i=1}^{m} a_{ij} w_i^* \right\}$$

we can easily see that $(w_1^*, \ldots, w_{m+1}^*)$ is dual feasible and prove that

$$0 \leq c^T x^* - w_{m+1}^* \leq \mu \ln n.$$

The following theorem sharpens that result and provides the motivation for considering a perturbation of the objective function of the primal problem with an entropic term $\mu \sum_{j=1}^{n} x_j \ln x_j$.

Theorem[14]: Let w^* be the unique maximum of the concave function $h(w, \mu)$ where $\mu = \varepsilon / \ln n, \varepsilon > 0$. If x^* is as defined by equation (2) then

$$0 \leq c^T x^* - w_{m+1}^* \leq \varepsilon$$

and (x^*, w^*) is an ε-optimal primal-dual solution pair.

In short, what Fang proposed is to attack the LP indirectly: first introduce the perturbing entropic term $\frac{\varepsilon}{\ln n} \sum_{j=1}^{n} x_j \ln x_j$, where $\varepsilon > 0$ and small, to its objective function and thereby obtain, via the Karush-Kuhn-Tucker conditions, a non-linear system of equation in terms of the primal variable and the Lagrange multipliers. Second, manipulate that system slightly to obtain a simpler system [14]. Third, apply Newton's method to the modified system and get the optimal solution of the LP.

The second approach used to solve the LP is due to Fang and Tsao [8] . It considers the LP in its standard form. Problem P is:

Min $c^T x$

subject to

$Ax = b, x \geq 0$

and its dual, D:

Max $b^T w$

subject to

$A^T w \leq b, w \in \mathbf{R}^m$.

Adding the entropic perturbation to the objective function of P we get problem P_μ:

Min $c^T x + \mu \sum_{j=1}^{n} x_j \ln x_j$

subject to

$Ax = b, x > 0$

where $\mu > 0$ and, as before, an interior solution is assumed to exist.

Using the simple inequality

$$\ln z \leq z - 1, \text{ for } z > 0$$

and the substitution

$$z_j = \frac{\exp\left\{\left[\left(\sum_{i=1}^{m} a_{ij} w_i - c_j\right)/\mu\right] - 1\right\}}{x_j}, \text{for } j = 1, 2, \ldots, n$$

it is shown that the dual of P_μ, D_μ, can be defined as the convex programming problem:

Max $\sum_{i=1}^{m} b_i w_i - \mu \sum_{i=1}^{m} \exp\left\{\left[\left(\sum_{i=1}^{m} a_{ij} w_i - c_j\right)/\mu\right] - 1\right\}$
subject to $w \in \mathbf{R}^m$.

It can also be shown that $\text{Min}(P_\mu) \geq \text{Max}(D_\mu)$ and that as $\mu \to 0$, $w^*(\mu)$, the optimal solution of D_μ, approaches w^*, the optimal solution of D. We, furthermore, have the following two theorems:

Theorem [8] : If problem P has an interior feasible solution and a full row-rank constraints matrix A, then problem D_μ has a unique optimal solution $w^*(\mu) \in \mathbf{R}^m$. In this case the optimal solution of P_μ, $x^*(\mu)$, is given by

$$x_j^*(\mu) = \exp\left\{\left[\left(\sum_{i=1}^{m} a_{ij} w_i^*(\mu) - c_j\right)/\mu\right] - 1\right\}, \text{for } j = 1, 2, \ldots, n. \qquad (3)$$

Moreover, $\mathrm{Min}(P_\mu) = \mathrm{Max}(D_\mu)$.

and

Theorem [8] : Under the assumption that problem P has a bounded feasible domain whose interior is non-empty, choose $\varepsilon > 0$ and let

$$\mu = \varepsilon / 2n \max\{e^{-1}, |M \ln M|\}$$

where $\|x\|_2 \leq M$ for all x in the feasible domain. Then $x^*(\mu)$, as given in equation (3) is an $\varepsilon-$optimal solution to problem P.

The algorithm to solve problem P should be clear by now. First choose an $\varepsilon > 0$ and μ according to the previous theorem. Second, solve the unconstrained convex programming problem D_μ. Third and last, use the dual to primal conversion formula (3) to obtain an $\varepsilon-$optimal solution to problem P. The only step that is not straight-forward in this scheme is step 2. To solve the convex programming problem, Fang and Tsao propose a customized curved search method. This is a descent method where instead of the iterates being chosen along the steepest descent direction, they are chosen along a quadratic curve determined by the gradient vector and the Hessian matrix of the objective function [8].

Note that no mention of Phase 1 iterations was given; a definite advantage. Another important observation was reported from computational experiments that were performed: the total number of iterations for the convex programming solver grows slightly with the size of the problem.

2.4 Cross-Entropy Minimization with Cross-Entropic Constraints

The question to be addressed here is that of finding the probability distribution p that is "closest" to a given a priori distribution, p^0, and that is at the same time within e_k "distance" from distribution p^k. Motivation to study this question comes from Information Theory where it has significant applications [1]. In mathematical terms, the problem considered, call it P, is of the form:

$$\text{Min } g_0(q) = \sum_{i=1}^{n} q_i \ln\left(\frac{q_i}{p_i^0}\right)$$

subject to

$$\sum_{i=1}^{n} q_i = 1, q \geq 0$$

$$g_k(q) = \sum_{i=1}^{n} q_i \ln\left(\frac{q_i}{p_i^k}\right) \leq e_k, k = 1, 2, .., r.$$

To solve this problem Fang *et al* [9] set this problem in a geometric programming form:

Let M be the $(r+1)n \times n$ dimensional matrix $(I, I, \ldots, I)^T$ where each I is the $n \times n$ identity matrix and let X be the column space of M. A vector $x \in X$ is the Cartesian product of $(r+1)$ identical n-vectors $x^k, k = 0, 1, \ldots, r$ and so problem P is equivalent to the following problem PGP:

$$\text{Min } G_0(x) = \sum_{i=1}^{n} x_i^0 \ln\left(\frac{x_i^0}{p_i^0}\right)$$

subject to

$$G_k(x) = \sum_{i=1}^{n} x_i^k \ln\left(\frac{x_i^k}{p_i^0}\right) - e_k \leq 0, k = 1, 2, \ldots, r$$

$$\sum_{i=1}^{n} x_i^k = 1, x^k \geq 0, k = 0, 1, \ldots, n$$

$$x \in X$$

where $G_k(x) = g_k(x^k) - e_k$.

Using the Arithmetic-Geometric Inequality and some algebra and analysis, the dual of PGP, problem DGP, can be defined as:

$$\text{Max } V(y, \lambda) = -\ln\left[\sum_{i=1}^{n} p_i^0 \exp\left(y_i^0\right)\right] - \sum_{k=1}^{r} \lambda_k \left[\ln\left(\sum_{i=1}^{n} p_i^k \exp\left(\frac{y_i^k}{\lambda_k}\right)\right) + e_k\right]$$

subject to

$$M^T y = 0$$

$$\lambda_k \geq 0, k = 1, 2, \ldots, r$$

Here, the objective function is still concave and y is the Cartesian product of $(r+1)$ n vectors $y^k, k = 0, 1, \ldots, r$. Since $M^T y = 0$ we get that $y^0 = -\sum_{k=1}^{r} y^k$. Using this observation in the objective function we obtain a dual problem in n less variables and no y constraint. Further and above, the concavity of the objective function is preserved which may suggest that the new problem is easier than DGP. Numerical experience shows otherwise, though, and so DGP remains the problem to solve [9].
A weak duality theorem is next:

Theorem[9] : If x is a primal feasible solution of PGP and (y, λ) is a dual feasible solution of DGP, then $V(y, \lambda) \leq G_0(x)$.

Note that both problems PGP and DGP have objective functions with the desirable convexity/concavity property. But whereas DGP has linear constraints , PGP,by contrast, has non-linear constraints. Clearly, then, PGP is a more difficult problem than DGP is. The trouble with DGP, however, is that its objective function does not possess partial derivatives at some of the boundaries of its feasible region; namely, when $\lambda_k = 0$ for any k. To overcome this difficulty, Fang *et al* perturb the λ_k constraints away from the boundary. They consider the following problem $DGP(l)$:

$$\text{Max } V(y, \lambda) = -\ln\left[\sum_{i=1}^{n} p_i^0 \exp\left(y_i^0\right)\right] - \sum_{k=1}^{r} \lambda_k \left[\ln\left(\sum_{i=1}^{n} p_i^k \exp\left(\frac{y_i^k}{\lambda_k}\right)\right) + e_k\right]$$

subject to

$$M^T y = 0$$

$$\lambda_k \geq l_k > 0, k = 1, 2, \ldots, r$$

The techniques of Convex Programming can now be used on $DGP(l)$. But before doing that, we need to determine how close the optimal solutions of PGP, $DGP(l)$ and DGP are and what duality gap, if any, exists between PGP and DGP. The following three theorems do just that.

Theorem[9] **:** If problem DGP has a finite supremum, then problem PGP is consistent. Moreover, for any given $\varepsilon > 0$, if $\delta \in (0, 1)$, V is an upper bound on DGP, (y^+, λ^+) is one of its feasible solutions with $\lambda^+ > 0$ and

$$l_k(\varepsilon) = \frac{\varepsilon\delta\lambda_k^+}{V - V(y^+, \lambda^+)}, k = 1, 2 \ldots, n \tag{4}$$

then the perturbed problem $DGP\,(l\,(\varepsilon))$ has an optimal solution (y, λ) and problem PGP has a feasible solution x such that

$$0 \le G_0(x) - V(y, \lambda) \le \varepsilon.$$

Moreover,

Theorem[9] **:** (Strong Duality Theorem) Problem PGP is consistent if and only if its dual problem DGP has a finite optimum value. In this case, the two problems attain a common optimum value.

and

Theorem[9] : For any given $\varepsilon > 0$, if $l(\varepsilon)$ is defined by (4) and (y, λ) is an optimal solution to problem $DGP\,(l\,(\varepsilon))$ then x defined by

$$\begin{aligned} x_i^0 &= \frac{p_i^0 \exp\left(y_i^0\right)}{\sum_{j=1}^n p_j^0 \exp\left(y_j^0\right)}, \text{ for } i = 1, \ldots, n, \qquad (5) \\ x_i^k &= \frac{p_i^k \exp\left(y_i^k/\lambda_k\right)}{\sum_{j=1}^n p_j^k \exp\left(y_j^k/\lambda_k\right)}, \text{ for } k = 1, \ldots, r \text{ and } i = 1, \ldots, n \end{aligned}$$

is an ε-optimal solution of problem PGP.

Summarizing all those results, the algorithm to solve problem PGP is as follows:

step 1: Choose an $\varepsilon > 0$, a dual feasible vector (y^+, λ^+) with $\lambda^+ > 0$ and a dual upper bound V of $V(y, \lambda)$.

step 2: Choose a perturbation vector $l(\varepsilon)$ according to (4).

step 3: Find an optimal solution (y, λ) for the perturbed dual problem $DGP\,(l\,(\varepsilon))$ using a any non-linear optimizer.

step 4: Plug in (y, λ) into (5) to generate an ε-optimal solution x.

2.5 Entropy Optimization Methods with Convex Constraints

The problem considered in this class by Fang and Rajasekera [10] is a generalization of the problem in the previous subsection. Problem P is:

$$\text{Min } g_0(q) = \sum_{i=1}^{n} q_i \ln\left(\frac{q_i}{p_i^0}\right)$$

subject to

$$\sum_{i=1}^{n} q_i = 1, q \geq 0$$

$$g_k(q) = h_k\left(A_k q\right) + b_k^T q + c_k \leq 0, k = 1, 2, .., r.$$

where A_k is an $m_k \times n$ matrix with full rank, b_k is n-dimensional, c_k is a constant and h_k is a convex function on $\mathbf{R}^{m_k}$ for each k. h_k is also assumed to possess partial derivatives and to be closed convex with an epigraph containing no non-vertical lines.

Note that when $h_k\left(A_k q\right) = 0$, we get linear constraints, when $h_k\left(A_k q\right) = \frac{1}{2} q^T A_k^T A_k q$ we get quadratic constraints and that when $A_k = I$, $h_k\left(q\right) = \sum_{i=1}^{n} q_i \ln\left(\frac{q_i}{p_i^k}\right)$, for a distribution q, and $b_k = 0$ we get the entropic constraints.

The strategy followed to solve this problem is exactly the same as the one employed to solve the Cross-Entropy minimization problem with entropic constraints. First, the problem is reformulated as a geometric programming problem. Second, using the Arithmetic-Geometric Inequality and the Conjugate Inequality, a partial dual of the problem is defined which turns out also to be a convex programming problem with linear constraints. This new problem is a partial dual and not a dual because some of the dual variables satisfy the positivity constraint instead of the non-negativity constraint. The new problem suffers, also, from the same troubles as the corresponding one in the case of the entropic constraints: the objective function has no partial derivatives at some of the boundary points. Third, to deal with this, a perturbation of the partial dual is introduced that is similar to the perturbation in the case of the entropic constraint; the result is a perturbed dual problem. Fourth, a solution of the perturbed problem is sought using any general nonlinear solver. Fifth and last, dual to primal conversion formulas are used to generate an ε-optimal solution to the original problem.[10]

2.6 Maximum Entropy and Constrained Optimization

The problem we consider here is a general constrained non-linear programming problem. P:

$$
\begin{aligned}
&\text{Min } f(x)\\
&\text{subject to}\\
&g_j(x) \leq 0, j = 1, \ldots, m\\
&x \in \mathbf{R}^n
\end{aligned}
$$

where f and the g_j's are real valued functions with continuous partial derivatives.

One approach to solve this problem, due to Templeman and Li [15] and called constraints surrogation, is to consider problem S:

$$
\begin{aligned}
&\text{Min } f(x)\\
&\text{subject to}\\
&\sum_{j=1}^{m} \lambda_j g_j(x) \leq 0\\
&x \in \mathbf{R}^n
\end{aligned}
$$

where the vector $\lambda \in \mathbf{R}^n_+$ is appropriately chosen and, without loss of generality, normalized. The idea is for the single constraint of S to replace the m constraints of P in the sense that solving problem S yields the solutions of problem P. The problem is, then, to find x and the correct values of λ.

A necessary condition for the equivalence of the two problems is for the Lagrangian of S

$$L_S\,(x, \alpha, \lambda) = f(x) + \alpha \sum_{j=1}^{m} \lambda_j g_j(x)$$

to satisfy the saddle-point condition:

$$L_S\,(x, \alpha^*, \lambda^*) \geq L_S\,(x^*, \alpha^*, \lambda^*) \geq L_S\,(x^*, \alpha, \lambda)$$

where x^*, α^*, λ^* are the optimal values of x, α, λ, respectively.

This saddle point condition suggests a two-phase iterative approach for solving P. Starting with an estimate λ^0 of λ, solve problem S to get x^0. This corresponds to minimizing the left-hand side of the saddle point condition. With the x^0, update λ^0, by a maximization process corresponding to the right-hand side of the saddle point condition, to get λ^1. Then, repeat this sequence until convergence. To find the λ updates note that if any of the constraints of P is binding at the optimal solution then the constraint of S

is binding too and so can be treated as an equality. Of course, if none of the constraints of P is binding then the problem is an unconstrained problem to begin with; and so, much easier to solve. We will therefore assume that the constraint of S is binding. In this case, and in light of the fact that given x^0 we need to find λ^1 with $0 \leq \lambda_j^1 \leq 1$ for all j, the constraint of S may be seen as a moment constraint on the probability distribution λ^1. Since we have no prior information about λ^1, we may apply MaxEnt to find it *i.e.* we solve:

$$\text{Max } -K\sum_{j=1}^{m}\lambda_j^1 \ln \lambda_j^1$$

subject to

$$\sum_{j=1}^{m}\lambda_j^1 g_j(x^o) = \varepsilon_1$$

$$\sum_{j=1}^{m}\lambda_j^1 = 1$$

$\lambda_j^1 \geq 0$ for all j.

where K and $\varepsilon_1 > 0$ are constants. ε_1 corresponds to the fact that a prior estimate of x is being used and not the x corresponding to λ^1. The solution to this problem is

$$\lambda_j^1 = \frac{\exp\left(\beta g_j\left(x^0\right)/K\right)}{\sum\limits_{j=1}^{m}\exp\left(\beta g_j\left(x^0\right)/K\right)}, \text{ for } j = 1, 2, \ldots, n$$

where β is the Lagrange multiplier corresponding to the moment constraint of the MaxEnt problem. Here, β/K can be considered as a parameter and the sequence of β/K 's should be increasing to infinity as the number of iterations increase [15].

Templeman and Li also presented another approach to the surrogate problem. They directly incorporate the entropy function into problem S to get problem SA:

$$\text{Min}_x \max{}_\lambda\ f(x) - \tfrac{1}{p}\sum_{j=1}^{m}\lambda_j \ln \lambda_j$$

subject to

$$\sum_{j=1}^{m} \lambda_j g_j(x) = 0$$

$$\sum_{j=1}^{m} \lambda_j = 1$$

$$x \in \mathbf{R}^n, \lambda \geq 0$$

The Lagrangian of this problem is:

$$L_{SA} = f(x) + \alpha \sum_{j=1}^{m} \lambda_j g_j(x) + \gamma \left[\sum_{j=1}^{m} \lambda_j = 1 \right] - \frac{1}{p} \sum_{j=1}^{m} \lambda_j \ln \lambda_j .$$

Stationarity of this Lagrangian with respect to λ and γ yields

$$\lambda_j^* = \frac{\exp\left(p\alpha g_j\left(x\right)\right)}{\sum\limits_{j=1}^{m} \exp\left(p\alpha g_j\left(x\right)\right)}, for j = 1, 2, \ldots, n$$

which is similar to the previous expression for λ^1.
L_{SA} is now:

$$L_{SA}^* = f(x) + \frac{1}{p} \sum_{j=1}^{m} \exp\left(p\alpha g_j\left(x\right)\right)$$

Templeman and Li, then, suggested that minimizing this expression with $p\alpha$ taking on an increasing positive sequence of values yields the optimal value for x. But probably with further analysis, an ε-optimal solution x could be generated for an appropriately large $p\alpha$.

Finally, it is worthwhile mentioning that Templeman and Li provided a proof of the validity of incorporating the entropy function here that is completely independent of probabilistic interpretations. Their proof is based on the substitution

$$U_j = \exp\left(p\left[f(x) + \alpha g_j(x)\right]\right)$$

in the inequality

$$\ln \left[\sum_{j=1}^{m} U_j \right] \geq \sum_{j=1}^{m} \lambda_j \ln \lambda_j - \sum_{j=1}^{m} \lambda_j \ln U_j$$

for $U, \lambda \geq 0$ and $\sum_{j=1}^{m} \lambda_j = 1$, with equality if and only if the right-hand side is maximized over λ.

References

[1] Ben-Tal, A. , M. Teboulle and A. Charnes (1988), "The Role of Duality in Optimization Involving Entropy Functionals with Applications to Information Theory", Journal of Optimization Theory and Applications 58(2), pp 209 − 223.

[2] Bregman, L.M. (1967), "The relaxation Method of Finding the Common Point of Convex Sets and its Application to the Solution of Problems in Convex Programming", USSR Computational Mathematics and Mathematical Physics 7(3), pp 200 − 217.

[3] Censor, Y., T. Elfving and G.T. Herman, "Special Purpose Algorithms for Linearly Constrained Optimization", *Maximum-Entropy and Bayesian Spectral Analysis and estimation Problems,* edited by C. Ray Smith and Gary J. Erickson, D. Reidel Publishing Company, 1987.

[4] Censor Y., A.R. De Pierro, T. Elfving, G.T. Herman and A.N. Isseum (1986), "On Maximization of Entropies and Generalization of Bregman's Method of Convex Programming", Technical Report MIPG 113, Medical Image Processing Group, Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, Pa.

[5] Censor,Y. and A. Lent (1981), "An Iterative Row-Action Method for Interval Convex Programming", Journal of Optimization Theory and Applications 34, pp 321 − 353.

[6] Erlander, Sven (1981), "Entropy in Linear Programs", Mathematical Programming 21, pp 137 − 151.

[7] Fang, S.C. (1990),"A New Unconstrained Convex Programming Approach to Linear Programming", OR Report No. 243, North Carolina State University, Raleigh,NC, Zeitschrift fur Operations Research 36, pp 149 − 161 (1992).

[8] Fang, S. C. and H. S. J. Tsao (1993), "Linear Programming with Entropic Perturbation", Zeitschrift fur Operations Research (37), pp 171 − 186.

[9] Fang, S. C., E.L. Peterson and J.R. Rajasekera (1992), "Minimum Cross-Entropy Analysis with Entropy-Type Constraints", Journal of Computational and Applied Mathematics 39, pp 165 − 178.

[10] Fang, S. C. and J.R. Rajasekera (1995), "Minimum Cross-Entropy Analysis with Convex Constraints", Information and Computation 116(2), pp 304 − 311.

[11] Gordon, R., R. Bender and G.T. Herman (1970), "Algebraic Reconstruction Techniques (ART) for Three Dimensional Electron Microscopy and X-Ray Photography", Journal of Theoretical Biology 29, pp 471 − 481.

[12] Kapur, J.N. and H.K. Kesavan, *Entropy Optimization Principles with Applications,* Academic Press, 1992.

[13] Lent, Arnold, "A Convergent Algorithm for Maximum Entropy Image Restoration, with a Medical X-Ray Application", *Image Analysis and Evaluation,* Conference Proceedings of the Society of Photographic Scientists and Engineers, R. Shaw ed., 1977, pp 249 − 257.

[14] Rajasekera, R. J. and S. C. Fang (1992), "Deriving an Unconstrained Convex Program for Linear Programming", Journal of Optimization Theory and Applications 75(3), pp 603 − 612.

[15] Templeman, A. B. and Li Xingsi, "Maximum Entropy and Constrained Optimization", *Maximum-Entropy and Bayesian Methods,* edited by J. Skilling, Kluwer Academic Publishers, 1989.