Mathematical Programming 44 (1989) 85-96
North-Holland

QUADRATICALLY CONSTRAINED MINIMUM CROSS-ENTROPY ANALYSIS

S.C. FANG
IE&OR, North Carolina State University, Box 7913, Raleigh, NC 27695, USA

J.R. RAJASEKERA
AT&T Technologies, Engineering Research Center, Box 900, Princeton, NJ 08540, USA

Received 16 November 1987

Quadratically constrained minimum cross-entropy problem has recently been studied by Zhang and Brockett through an elaborately constructed dual. In this paper, we take a geometric programming approach to analyze this problem. Unlike Zhang and Brockett, we separate the probability constraint from general quadratic constraints and use two simple geometric inequalities to derive its dual problem. Furthermore, by using the dual perturbation method, we directly prove the "strong duality theorem" and derive a "dual-to-primal" conversion formula. As a by-product, the perturbation proof gives us insights to develop a computation procedure that avoids dual non-differentiability and allows us to use a general purpose optimizer to find an ε-optimal solution for the quadratically constrained minimum cross-entropy analysis.

Key words: Minimum cross-entropy, maximum entropy, information theory, geometric programming, duality theory.

1. Introduction

Let $\{q_i\}$, $i=1,2,\ldots,n$, be a probability distribution function over a given finite state space and $\{p_i\}$ be an a priori distribution that estimates $\{q_i\}$, the "cross-entropy" or "discrimination information", between these two distributions is defined by $\sum_{i=1}^n q_i \log(q_i/p_i)$. When certain estimated values or bounds of the expectation functions (first order information), $\boldsymbol{b}_k^{\mathrm{t}}\boldsymbol{q} = \sum_{i=1}^n q_i b_{k,i}$, are given, the "minimum cross-entropy", or "minimum discrimination information (MDI)", strategy is to choose the one with the least cross-entropy among all those feasible probability distributions. It is a well-known result that the minimum cross-entropy strategy is the same as the "principle of maximum entropy" when the a priori distribution $\{p_i\}$ is a uniform one.

Minimum cross-entropy analysis has been applied successfully to various fields including information theory, pattern recognition, statistical mechanics, thermodynamics, game theory, and actuarial science. Good examples can be found in [1, 2, 6, 7, 10, 11].

Recently, Zhang and Brockett [12] extended the minimum cross-entropy analysis to include the second order information, $\frac{1}{2}\boldsymbol{q}^{\mathrm{t}}\boldsymbol{H}_k\boldsymbol{q} + \boldsymbol{b}_k^{\mathrm{t}}\boldsymbol{q} + c_k$, where $\boldsymbol{H}_k$ is a positive semi-definite symmetric matrix, in the constraints. Their elaborately devised approach was based on the dual form of a linearly constrained minimum-cross entropy problem as studied in [1, 2]. They took three major steps to derive a dual problem: (1) approximating the quadratically constrained feasible region by a set of supporting hyperplanes at points of a dense countable set; (2) finding the dual of the approximated, linearly constrained minimum cross entropy problem; (3) taking the limit of the approximate results and simplifying the expression. However, the dual objective function they obtained was neither concave nor convex over its feasible domain.

In this paper, we take a new approach to study the quadratically constrained minimum cross-entropy analysis. In Section 2, we use two simple inequalities to derive a geometric dual problem with concave objective function and linear constraints. The "weak duality theorem" is also derived. Then, in Section 3, we prove the "strong duality theorem" (or "existence theorem") for the quadratically constrained minimum cross-entropy problem. The dual-to-primal conversion is studied in Section 4. The converting formula constructs an "ε-optimal" solution for the original problem once a proper dual problem is solved. Finally, the computational issues are discussed in Section 5. The perturbational approach derived in Sections 3 and 4 provides us an opportunity to solve the quadratically constrained minimum cross-entropy problem by using a general purpose optimizer like MINOS [8]. The computation procedure is similar to our findings in [4, 5]. Our dual program can also be transformed to a special form that is similar to Zhang and Brockett's [12]. But the non-convexity (or non-concavity) property of the objective function in this form may make it less desirable.

2. Quadratically constrained minimum cross-entropy problem

The quadratically constrained minimum cross-entropy problem to be studied in this paper has the following form:

Program P

$$\text{Min} \quad g(\boldsymbol{q}) = \sum_{i=1}^{n} q_i \log(q_i/p_i) \quad (p_i > 0 \text{ are given}) \tag{1}$$

$$\text{s.t.} \quad \boldsymbol{q} \geqslant \boldsymbol{0}, \ \sum_{i=1}^{n} q_i = 1, \text{ and} \tag{2}$$

$$g_k(\boldsymbol{q}) = \tfrac{1}{2}\boldsymbol{q}^{\mathrm{t}}\boldsymbol{H}_k\boldsymbol{q} + \boldsymbol{b}_k^{\mathrm{t}}\boldsymbol{q} + c_k \leqslant 0, \ \ k = 1, 2, \ldots, r, \tag{3}$$

where $\boldsymbol{H}_k$ is a positive semidefinite, symmetric, $n \times n$ matrix, $\boldsymbol{q}$ and $\boldsymbol{b}_k$ are n-vectors, and c_k is a constant for each k.

Notice that, contrary to Zhang and Brockett [12], we explicitly separated the probability distribution constraint $\sum_{i=1}^{n} q_i = 1$ from the quadratic constraints set. This will give us better insights in formulating the dual problem.

For easy cross-reference, we stick to those notations that had been used extensively in the geometric programming literature (e.g., [3, 4]). We first express the positive semi-definite, symmetric matrix $\boldsymbol{H}_k$ as $\boldsymbol{H}_k = \boldsymbol{A}_k^{\mathrm{t}}\boldsymbol{A}_k$, where $\boldsymbol{A}_k$ is an $m_k \times n$ matrix and m_k is the rank of $\boldsymbol{H}_k$, for $k = 1, 2, \ldots, r$. Then define $\boldsymbol{I}_n$ to be the $n \times n$ identity matrix, and the index notations

$$[0] = \{1, 2, \ldots, n\}, \quad]0[= n, \tag{4}$$

$$[k] = \{]k-1[+1, \ldots,]k-1[+m_k\},]k[=]k-1[+m_k+1 \quad \text{for } k = 1, \ldots, r, \tag{5}$$

$$]r[= m. \tag{6}$$

Moreover, we let X denote the column space of the $m \times n$ matrix M defined by

$$M^{\mathrm{t}} = (\boldsymbol{I}_n, \boldsymbol{A}_1^{\mathrm{t}}, \boldsymbol{b}_1, \ldots, \boldsymbol{A}_r^{\mathrm{t}}, \boldsymbol{b}_r). \tag{7}$$

Hence each $\boldsymbol{x} \in X$ is an m-vector that can be expressed in Cartesian product form as $\boldsymbol{x} = \times_{k=0}^{r} \boldsymbol{x}^k$, where $\boldsymbol{x}^0 = (x_1, \ldots, x_{]0[})^{\mathrm{t}}$, and $\boldsymbol{x}^k = (x_{]k-1[+1}, \ldots, x_{]k[})^{\mathrm{t}}$, for $k = 1, \ldots r$. Then it is a simple routine to check that Program P is equivalent to the following "primal" geometric program (PGP):

Program PGP

$$\text{Min} \quad \boldsymbol{G}(\boldsymbol{x}) = \sum_{i \in [0]} x_i \log(x_i/p_i) \ \ (p_i > 0) \tag{8}$$

$$\text{s.t.} \quad x_i \geqslant 0 \ \text{ for } i \in [0], \quad \sum_{i \in [0]} x_i = 1, \tag{9}$$

$$\boldsymbol{G}_k(\boldsymbol{x}) = \tfrac{1}{2} \sum_{i \in [k]} x_i^2 + x_{]k[} + c_k \leqslant 0, \ \ k = 1, 2, \ldots, r, \tag{10}$$

$$\boldsymbol{x} \in X. \tag{11}$$

In order to construct a geometric dual problem, corresponding to primal vectors $\boldsymbol{x}$, we define dual vectors $\boldsymbol{y} = \times_{k=0}^{r} \boldsymbol{y}^k$, where $\boldsymbol{y}^0 = (y_1, \ldots, y_{]0[})^{\mathrm{t}}$, and $\boldsymbol{y}^k = (y_{]k-1[+1}, \ldots, y_{]k[})^{\mathrm{t}}$, for $k = 1, \ldots, r$. In this way, $\boldsymbol{y}$ is a column vector of size m. Then consider the arithmetic-geometric inequality used in [3]:

$$\sum_{i \in [0]} p_i \exp(y_i) \geqslant \prod_{i \in [0]} \{p_i \exp(y_i)/x_i\}^{x_i}. \tag{12}$$

This inequality is true for any real numbers y_i, $x_i > 0$, $i \in [0]$, and $\sum_{i \in [0]} x_i = 1$. By taking the logarithm of both sides and simplifying it, we have

$$\sum_{i \in [0]} x_i y_i \leqslant \sum_{i \in [0]} x_i \log(x_i/p_i) + \log\left(\sum_{i \in [0]} p_i \exp(y_i)\right). \tag{13}$$

With the understanding that $0 \times \log 0 = 0$, the above inequality is true for all y_i, $x_i \geqslant 0$, $i \in [0]$, and $\sum_{i \in [0]} x_i = 1$; and with equality holding if and only if $p_i \exp(y_i) = K \times x_i$, for $i \in [0]$, where K is a constant.

Next consider the quadratic inequality used in [4]:

$$\sum_{i\in[k]} x_i z_i \leqslant \tfrac{1}{2} \sum_{i\in[k]} (x_i^2 + z_i^2) \quad \text{for real numbers } x_i,\, z_i. \tag{14}$$

Replacing z_i by $y_i/y_{]k[}$, multiplying both sides by $y_{]k[}$, adding $x_{]k[}y_{]k[}$ to both sides, and adding then subtracting the term $c_k y_{]k[}$ on the right hand side, we have

$$\sum_{i\in[k]} x_i y_i + x_{]k[}y_{]k[} \leqslant y_{]k[}\left\{\tfrac{1}{2}\sum_{i\in[k]} x_i^2 + x_{]k[} + c_k\right\} + \left\{\tfrac{1}{2}\sum_{i\in[k]} y_i^2/y_{]k[} - c_k y_{]k[}\right\}. \tag{15}$$

With the understanding that $y_{]k[}=0$ only if $y_i=0$ for each $i\in[k]$, this expression holds for all $y_{]k[}\geqslant 0$ and $k=1,2,\ldots,r$.

Combining the inequalities (13) and (15) (for $k=1,2,\ldots,r$), we obtain that

$$\sum_{i=1}^{m} x_i y_i \leqslant \sum_{i\in[0]} x_i \log(x_i/p_i) + \sum_{k=1}^{r} y_{]k[}\left\{\tfrac{1}{2}\sum_{i\in[k]} x_i^2 + x_{]k[} + c_k\right\}$$
$$+\log\left(\sum_{i\in[0]} p_i \exp(y_i)\right) + \sum_{k=1}^{r}\left\{\tfrac{1}{2}\sum_{i\in[k]} y_i^2/y_{]k[} - c_k y_{]k[}\right\}. \tag{16}$$

If $\boldsymbol{x}$ is primal feasible, $y_{]k[}\geqslant 0$ (and $y_{]k[}=0$ only if $y_i=0$ for each $i\in[k]$, for $k=1,2,\ldots,r$), and $M^{\mathrm{t}}\boldsymbol{y}=0$, then we know that $\sum_{i=1}^{m} x_i y_i = 0$, and $y_{]k[}\{\tfrac{1}{2}\sum_{i\in[k]} x_i^2 + x_{]k[} + c_k\}\leqslant 0$, for $k=1,\ldots,r$. Consequently,

$$-\log\left(\sum_{i\in[0]} p_i \exp(y_i)\right) + \sum_{k=1}^{r} c_k y_{]k[} - \tfrac{1}{2}\sum_{k=1}^{r}\sum_{i\in[k]} y_i^2/y_{]k[} \leqslant \sum_{i\in[0]} x_i \log(x_i/p_i). \tag{17}$$

Based on this inequality, we can define the "dual program (DGP)" of Program PGP as follows:

Program DGP

$$\text{Max} \quad V(\boldsymbol{y}) = -\log\left(\sum_{i\in[0]} p_i \exp(y_i)\right) + \sum_{k=1}^{r} c_k y_{]k[} - \tfrac{1}{2}\sum_{k=1}^{r}\sum_{i\in[k]} y_i^2/y_{]k[} \tag{18}$$

$$\text{s.t.} \quad M^{\mathrm{t}}\boldsymbol{y} = 0, \tag{19}$$

$$y_{]k[} \geqslant 0, \quad k=1,2,\ldots,r, \tag{20}$$

where $\boldsymbol{y}\in R_n$, with M, $[k]$, $]k[$ and c_k defined as before.

Notice that we were able to replace the constraint "$y_{]k[}\geqslant 0$ and $y_{]k[}=0$ only if $y_i=0$, for each $i\in[k]$" by a simple constraint "$y_{]k[}\geqslant 0$", since if $y_{]k[}=0$ and $y_i\neq 0$ for some $i\in[k]$, then $V(\boldsymbol{y})$ would become $-\infty$.

There are several observations that can be made here. First, as an immediate consequence of the inequality (17), we can easily show the "weak duality theorem", i.e.:

Theorem 1. *If x is a primal feasible solution of Program* PGP *and y is a dual feasible solution of Program* DGP, *then* $V(y) \leqslant G(x)$.

Second, we follow the well-established geometric program theory to define a "canonical program" as a program with a feasible dual solution y^+ such that $y^+_{]k[} > 0$, for $k = 1, \ldots, r$. Then, since matrix M is defined by equation (7), if we take vector y^+ with components $(y^+)^0 = -\sum_{k=1}^{r} b_k$, $y^+_{]k-1[+1} = \cdots = y^+_{]k-1[+m_k} = 0$, and $y^+_{]k[} = 1$, for $k = 1, \ldots, r$, equation (19) is automatically satisfied for y^+. Therefore, we know that the following result is true.

Theorem 2. *Program* DGP *is a canonical program that is always consistent.*

The third observation is a direct consequence of the first two.

Theorem 3. *Program* DGP *has a finite optimum value, if Program* PGP *is consistent.*

Moreover, just like most geometric duals [4, 5], it is easy to see that $V(y)$ is concave over the dual feasible region. However, it is not differentiable at the boundary of the constraints $y_{]k[} \geqslant 0$, for $k = 1, 2, \ldots, r$. This issue will be explicitly addressed in later sections.

3. Existence theorem

Compared to the primal program PGP, its dual program DGP has a concave objective function with linear constraints. This makes the dual approach very attractive. However, as we observed in the last section, the dual problem is non-differentiable at some of its boundary points. This would cause at least two difficulties: (1) applying Lagrangian theory is difficult without resorting to subgradient techniques as developed in [9]; (2) developing a dual based solution procedure becomes more troublesome.

Fortunately, the most recently developed "controlled dual perturbation" approach [4, 5] can be applied here. We can perturb constraints (20) by a well-controlled positive amount l_k, for $k = 1, 2, \ldots, r$, and construct the following "perturbed dual" program:

Program DGP(l)

$$\text{Max} \quad V(y) = -\log\left(\sum_{i\in[0]} p_i \exp(y_i)\right) + \sum_{k=1}^{r} c_k y_{]k[} - \tfrac{1}{2} \sum_{k=1}^{r} \sum_{i\in[k]} y_i^2 / y_{]k[} \tag{21}$$

$$\text{s.t.} \quad M^t y = 0, \tag{22}$$

$$y_{]k[} \geqslant l_k > 0, \quad k = 1, 2, \ldots, r, \tag{23}$$

where all notations are defined as before, and those positive l_k's are called "perturbations".

There is one interesting observation that can be made on this perturbed dual program. Since Theorem 2 indicates that Program DGP is canonical with a feasible solution y^+ such that $y^+_{]k[} > 0$, for $k = 1, \ldots, r$, we can always choose a perturbation vector $l = (l_1, \ldots, l_r)$ with $y^+_{]k[} > l_k > 0$, for $k = 1, \ldots, r$. In this way, the corresponding Program DGP(l) is always feasible. Moreover, Program DGP(l) has a concave differentiable objective function over its linearly contained non-void feasible region. This will make the dual approach very favorable since any general purpose optimizer like MINOS can be used to solve it without much difficulty. However, we have to show that there is no duality gap before we actually devote to the computations.

The following theorem will lead us to this goal:

Theorem 4. *If Program* DGP *has a finite supremum, then program* PGP *is consistent. Moreover, for any given* $\varepsilon > 0$, *we can choose a proper perturbation vector* $l(\varepsilon) = (l(\varepsilon)_1, \ldots, l(\varepsilon)_r)$ *with* $l(\varepsilon)_k > 0$, *for* $k = 1, \ldots, r$, *such that the perturbed dual program* DGP($l(\varepsilon)$) *has an optimal solution* y *and the Program* PGP *has a feasible solution* x *satisfying the condition* $0 \leqslant G(x) - V(y) \leqslant \varepsilon$.

Proof. Let y^+ be a feasible solution of Program DGP with $y^+_{]k[} > 0$, for $k = 1, \ldots, r$. We consider Program DGP(l) with $l_k < y^+_{]k[}$, for $k = 1, \ldots r$. Since program DGP has a finite supremum, so does program DGP(l). Now, consider the Lagrangian defined as

$$L = V(y) + \sum_{k=1}^{r} \lambda_k (y_{]k[} - l_k) + \gamma^t M^t y, \tag{24}$$

where γ (a column vector of size n) and λ_k, $k = 1, 2, \ldots, r$ are Lagrange multipliers of the constraints (22) and (23) respectively.

Because program DGP(l) is differentiable and concave, there exist solutions to the following Kuhn-Tucker conditions:

$$\partial L / \partial y_i = -(p_i \exp(y_i)) / \sum_{i \in [0]} (p_i \exp(y_i)) + \gamma^t M_i = 0 \quad \text{for } i \in [0], \tag{25}$$

$$\partial L / \partial y_i = -y_i / y_{]k[} + \gamma^t M_i = 0 \quad \text{for } i \in [k],\ k = 1, 2, \ldots, r, \tag{26}$$

$$\partial L / \partial y_{]k[} = c_k + \tfrac{1}{2} \sum_{i \in [k]} y_i^2 / y_{]k[}^2 + \lambda_k + \gamma^t M_{]k[} = 0 \quad \text{for } k = 1, 2, \ldots, r, \tag{27}$$

$$\lambda_k (y_{]k[} - l_k) = 0, \quad k = 1, 2, \ldots, r, \tag{28}$$

$$\lambda_k \geqslant 0, \quad k = 1, 2, \ldots, r, \tag{29}$$

plus conditions (22) and (23), where M_i (a column vector of size n) is the ith column of matrix M^t.

If we let

$$x_i = \gamma^t M_i \quad \text{for } i = 1, \ldots, m, \tag{30}$$

then, by definition, we know $x \in X$ (the column space of M). Moreover, by condition (25), we know $x_i \geq 0$ and $\sum_{i\in[0]} x_i = 1$; and by condition (26), $y_i / y_{]k[} = x_i$, for $i \in [k]$, $k = 1, \ldots, r$. Plugging into condition (27), we have

$$\tfrac{1}{2} \sum_{i\in[k]} x_i^2 + x_{]k[} + c_k + \lambda_k = 0, \quad k = 1, 2, \ldots, r. \tag{31}$$

Since $\lambda_k \geq 0$ is given by condition (29), we know (10) is satisfied by this vector x. Hence Program PGP has a feasible solution x.

To prove the second part of this theorem, let us consider the right hand side of the Lagrangian defined by (24), since condition (28) holds and x and y are complementary, we have

$$L = V(y).$$

On the other hand, if we plug in the value of $V(y)$ given by definition (21) to the right hand side of (24) and rearrange terms, we have

$$\begin{aligned} L = &-\log\Big(\sum_{i\in[0]} p_i \exp(y_i)\Big) + \sum_{i\in[0]} \gamma^t M_i y_i \\ &+ \sum_{k=1}^{r} y_{]k[} \Big\{ c_k - \tfrac{1}{2} \sum_{i\in[k]} y_i^2 / y_{]k[}^2 + \lambda_k + \gamma^t M_{]k[} \Big\} \\ &+ \sum_{k=1}^{r} \sum_{i\in[k]} \gamma^t M_i y_i - \sum_{k=1}^{r} \lambda_k l_k. \end{aligned}$$

In other words,

$$\begin{aligned} V(y) = &-\log\Big(\sum_{i\in[0]} p_i \exp(y_i)\Big) + \sum_{i\in[0]} \gamma^t M_i y_i \\ &+ \sum_{k=1}^{r} y_{]k[} \Big\{ c_k - \tfrac{1}{2} \sum_{i\in[k]} y_i^2 / y_{]k[}^2 + \lambda_k + \gamma^t M_{]k[} \Big\} \\ &+ \sum_{k=1}^{r} \sum_{i\in[k]} \gamma^t M_i y_i - \sum_{k=1}^{r} \lambda_k l_k. \end{aligned} \tag{32}$$

Now, by (25), we know that $p_i \exp(y_i) = x_i \times \sum_{i\in[0]} (p_i \exp(y_i))$, for each $i \in [0]$. But $\sum_{i\in[0]} (p_i \exp(y_i))$ is a constant number, hence inequality (13) becomes equality and

$$\sum_{i\in[0]} \gamma^t M_i y_i = \sum_{i\in[0]} x_i y_i = \log\Big(\sum_{i\in[0]} p_i \exp y_i\Big) + \sum_{i\in[0]} x_i \log(x_i / p_i). \tag{33}$$

In addition, by (26), we have

$$\sum_{k=1}^{r} \sum_{i\in[k]} \gamma^t M_i y_i = \sum_{k=1}^{r} \sum_{i\in[k]} y_i^2 / y_{]k[}. \tag{34}$$

Substituting these values into (32) and noticing (27), we show that

$$V(y) = G(x) - \sum_{k=1}^{r} \lambda_k l_k.$$

By Theorem 1, we further have

$$0 \leqslant G(x) - V(y) = \sum_{k=1}^{r} \lambda_k l_k. \tag{35}$$

Now substitute the components of the vector y^+ in inequalities (13) and (15). By adding these inequalities together and using condition (31), for $k = 1, \ldots, r$, we have

$$0 \leqslant -\sum_{k=1}^{r} \lambda_k y^+_{]k[} - V(y^+) + G(x). \tag{36}$$

Substituting for $G(x)$ from (35), we get

$$0 \leqslant \sum_{k=1}^{r} \lambda_k (y^+_{]k[} - l_k) \leqslant V(y) - V(y^+) \leqslant V - V(y^+) \tag{37}$$

where V is an upper bound for the dual program as assumed for this theorem.

Now, for any given $\varepsilon > 0$, if we choose l_k such that

$$l_k = l_k(\varepsilon) = \varepsilon\delta y^+_{]k[} / (V - V(y^+) + \varepsilon) \quad \text{for any } 0 < \delta < 1 \text{ and } k = 1, 2, \ldots, r, \tag{38}$$

then there are two possible cases for consideration:

Case 1: If $V - V(y^+) = 0$, then since $l_k = \delta y^+_{]k[}$ and $y^+_{]k[} - l_k = (1-\delta) y^+_{]k[} > 0$, for each k, therefore λ_k in (37) has to be 0. Consequently, by (35), we have $G(x) - V(y) = 0$.

Case 2: If $V - V(y^+) > 0$, by (38), $y^+_{]k[} = [(V - V(y^+) + \varepsilon)/\varepsilon\delta] l_k$. Plugging into (37), we obtain

$$\sum_{k=1}^{r} \lambda_k [((V - V(y^+))/\delta\varepsilon) + (1/\delta) - 1] l_k \leqslant V - V(y^+).$$

This implies that $\sum_{k=1}^{r} \lambda_k l_k \leqslant \delta\varepsilon < \varepsilon$. Therefore, by (35), we have $0 \leqslant G(x) - V(y) < \varepsilon$. This completes the proof.

We now are ready to prove the "strong duality theorem" (or "existence theorem") for the quadratically constrained minimum cross-entropy problem.

Theorem 5. *Program* PGP *is consistent if and only if its dual Program* DGP *has a finite optimum value. In this case, Programs* PGP *and* DGP *attain a common optimum value.*

Proof. Combine Theorems 3 and 4, we know Program PGP has a feasible solution if and only if Program DGP has a finite optimum value. Moreover, for any given $\varepsilon > 0$, we can find feasible x and y such that $0 \leqslant G(x) - V(y) \leqslant \varepsilon$. Hence we have $V(y) \leqslant V^* \leqslant G^* \leqslant G(x)$, where V^* and G^* are optimum values of DGP and PGP respectively. Consequently, $0 \leqslant G^* - V^* \leqslant G(x) - V(y) \leqslant \varepsilon$, for any $\varepsilon > 0$. Therefore, $G^* = V^*$ and we complete the proof.

4. Dual-to-primal conversion

In this section, we show that an ε-optimal solution to the original quadratically constrained cross-entropy problem (Program P) can be derived by solving Program (DGP($l(\varepsilon)$)) with $l(\varepsilon)$ properly selected according to equation (38). The first result is as follows:

Theorem 6. *If γ is a vector of Lagrange multipliers associated with the constraint set $M^t y = 0$, of Program* DGP($l(\varepsilon)$), *then γ is a feasible solution for Program* P *and* $0 \leqslant g(\gamma) - V(y) \leqslant \varepsilon$.

Proof. Since the first n columns of matrix M form an identity matrix I_n, by (30), we conclude that $x_i = \gamma_i$, for $i \in [0]$, and $G(x) = g(\gamma)$. By (25), we know $\gamma_i \geqslant 0$ and $\sum_{i\in[0]} \gamma_i = 1$. Moreover, putting the structure of matrix M and equation (30) together, then substituting for x_i ($i \in [k]$) and $x_{]k[}$ ($k = 1, \ldots, r$) in (31) and noticing the fact that $\lambda_k \geqslant 0$, we can conclude that $g_k(\gamma) = \frac{1}{2}\gamma^t H_k \gamma + b_k^t \gamma + c_k \leqslant 0$ for $k = 1, 2, \ldots, r$. Hence γ is a solution to Program P. Then, by Theorem 4, it is clear that $0 \leqslant g(\gamma) - V(y) \leqslant \varepsilon$. This completes the proof.

The second result is to express the primal solution γ in terms of dual solutions. To do so, we define

$$\mu = \log\left(\sum_{i\in[0]} p_i \exp(y_i)\right). \tag{39}$$

The constraint $M^t y = 0$ implies that

$$y^0 = -\sum_{k=1}^{r} \{A_k^t \bar{y}^k + b_k y_{]k[}\}, \tag{40}$$

where $\bar{y}^k$ is a column vector with elements y_i such that $i \in [k]$, for $k = 1, \ldots, r$. In other words, $\bar{y}^k$ is the same as y^k but with one less element $y_{]k[}$. Hence,

$$y_i = -\sum_{k=1}^{r} \{A_{k,i}^t \bar{y}^k + b_{k,i} y_{]k[}\} \quad \text{for } i \in [0], \tag{41}$$

where $A_{k,i}^t$ is the ith row of A_k^t and $b_{k,i}$ is the ith component of b_k.

Using the definition (30) in (25), we have $p_i \exp(y_i) = \gamma_i \exp \mu$, for $i \in [0]$. Now, substituting for y_i in this expression, we have

$$\gamma_i = p_i \exp(-\mu - \sum_{k=1}^{r} \{A_{k,i}^t \bar{y}^k + b_{k,i} y_{]k[}\}) \quad \text{for } i \in [0]. \tag{42}$$

Hence we have the following theorem:

Theorem 7. *For any given $\varepsilon > 0$, if $l(\varepsilon)$ is defined by equation* (38) *and y solves Program* DGP($l(\varepsilon)$), *then equations* (39) *and* (42) *generate an ε-optimal solution to the original quadratically constrained minimum cross-entropy analysis Program* P.

5. Computational issues

As a by-product of the theoretical insights we obtained in previous sections, we can outline a dual-based computation procedure as follows:

Step 1: Given an $\varepsilon > 0$, a dual feasible vector y^+ with $y^+_{]k[} > 0$, for $k = 1, \ldots, n$, and a dual upper bound V of $V(y)$.

Step 2: Choose a perturbation vector $l(\varepsilon) = (l(\varepsilon)_1, \ldots, l(\varepsilon)_r)$ according to equation (38).

Step 3: Find an optimal solution y for the perturbed dual program DGP($l(\varepsilon)$).

Step 4: Plug in y into equations (39) and (42) to generate an ε-optimal solution γ, then

$$0 \leqslant g(\gamma) - V(y) \leqslant \varepsilon.$$

Step 5: Stop.

There are several observations that can be made here.

1. Compared to the primal program P, the perturbed dual program DGP($l(\varepsilon)$) is relatively simple since it possesses a concave differentiable objective function and linear constraints. We may use a commercial nonlinear optimizer, eg., MINOS, to solve this dual problem. The additional customization of codes is minimal.

2. In Step 1, y^+ can be easily obtained as we derived for Theorem 2, i.e.,

$$(y^+)^0 = -\sum_{k=1}^{r} b_k, \quad y^+_{]k-1[+1} = \cdots = y^+_{]k-1[+m_k} = 0 \text{ and } y^+_{]k[} = 1 \quad \text{for } k = 1, \ldots, r.$$

3. Usually a dual bound V can be estimated easily. If it is not provided, we can always solve a Program DGP(l) with $l_k = \delta y^+_{]k[}$ first, then use equations (39) and (42) to find a corresponding primal solution γ^*, then we know $V \leqslant g(\gamma^*)$.

4. To compare with Zhang and Brocketts result in [12], let us consider Program DGP. If we define $y_i = w_i y_{]k[}$, for $i \in [k]$ and $k = 1, 2, \ldots, r$, then the constraint (19) $M^t y = 0$ implies that

$$y^0 = -\sum_{k=1}^{r} y_{]k[}\{A_k^t w^k + b_k\}. \tag{43}$$

In other words,

$$y_i = -\sum_{k=1}^{r} y_{]k[}\{A_{k,i}^t w^k + b_{k,i}\} \quad \text{for } i \in [0], \tag{44}$$

where $A_{k,i}^t$ is the ith row of the matrix A_k^t, $b_{k,i}$ is the ith component of vector b_k, and w^k is the column vector in R_{m_k} with components w_i, for $i \in [k]$.

Consequently, if we follow the notation $\sum r_i \exp s_i = r^t \exp s$, Program DGP becomes equivalent to

Program D

$$\text{Sup} \quad -\log\left[\boldsymbol{p}^{\mathrm{t}} \exp\left(-\sum_{k=1}^{r} y_{]k[}\{\boldsymbol{A}_k^{\mathrm{t}}\boldsymbol{w}^k + \boldsymbol{b}_k\}\right)\right] + \sum_{k=1}^{r} c_k y_{]k[} - \tfrac{1}{2}\sum_{k=1}^{r} y_{]k[}\|\boldsymbol{w}^k\|^2 \tag{45}$$

$$\text{s.t.} \quad y_{]k[} \geqslant 0 \text{ and } \boldsymbol{w}^k \in R_{m_k} \text{ for } k=1,\ldots,r. \tag{46}$$

This formulation is very close to Zhang and Brockett's dual program in [12], except the first part of the objective function is in a "log" form. The reason for this fact can be attributed to our original formulation of Program P which, unlike [12], explicitly separates the probability distribution requirement from quadratic constraints. But the basic characteristics of these two duals are essentially the same and there is no difference in computational complexity. For this form (Program D), although the constraints are even simpler than those of Program DGP, the objective function is neither convex nor concave. When it comes to applying a routine nonlinear optimizer that can pause a problem. In this sense, a program of the form DGP (actually, DGP($l(\varepsilon)$)) is preferred.

6. Conclusion

In this paper we have provided a geometric programming approach to the quadratically constrained minimum cross-entropy problem. A well-controlled dual perturbational method not only showed its theoretical value in proving the weak and strong duality theorems, but also led to a potential computational procedure that would allow us to use a general purpose optimizer to solve the problem.

References

[1] A. Charnes and W.W. Cooper, "An extremal principle for accounting balance of a resource-value transfer economy: existence, uniqueness and computation," *Rendiconte Accademia Nazionale dei Lincei, Rome* (8) 4 (1974) 556–578.

[2] A. Charnes, W.W. Cooper and L. Seiford, "Extremal principles and optimization qualities for Khinchin-Kullback-Leibler estimation," *Statistics* (*Zentralinstitut für Mathematik und Mechanik*) 9 (1978) 21–29.

[3] R.J. Duffin, E.L. Peterson and C. Zener, *Geometric Programming—Theory and Applications* (John Wiley, New York, 1967).

[4] S.C. Fang and J.R. Rajasekera, "Controlled perturbations for quadratically constrained quadratic programs," *Mathematical Programming* 36 (1986) 276–289.

[5] S.C. Fang and J.R. Rajasekera, Controlled dual perturbations for l_p-programming," *Zeitschrift für Operations Research* 30 (1986) A29–A42.

[6] S. Guiasu, *Information Theory with Applications* (McGraw-Hill, New York, 1977).

[7] S. Kullback, *Information Theory and Statistics* (John Wiley, New York, 1959).

[8] B.A. Murtagh and M.A. Saunders, "MINOS users guide," Technical Report 77-9, Systems Optimization Laboratory, Stanford University (1977).

[9] R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970).

[10] J.E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29(2) (1981) 230–237.

[11] J.E. Shore and R.W. Johnson, "Axiomatic derivation of principle of maximum entropy and principle of minimum cross-entropy," *IEEE Transactions on Information Theory* 26(1) (1980) 26–37.

[12] J. Zhang and P.L. Brockett, "Quadratically constrained information theoretic analysis," *SIAM Journal on Applied Mathematics* 47 (1987) 871–885.