MMA and GCMMA – Fortran versions March 2013

Krister Svanberg

KTH, Royal Institute of Technology

This note describes the algorithms used in the author's latest fortran implementations of MMA and GCMMA. The first versions of these methods were published in [1] and [2].

1 Considered optimization problem

The fortran implementations of the author's MMA and GCMMA codes are based on the assumption that the users optimization problem is written on the following form, where the optimization variables are $\mathbf{x} = (x_1, \ldots, x_n)^\mathsf{T}$, $\mathbf{y} = (y_1, \ldots, y_m)^\mathsf{T}$ and z.

$$\begin{aligned}
&\text{minimize} && f_0(\mathbf{x}) + z + \sum_{i=1}^{m}(c_i y_i + \tfrac{1}{2}y_i^2) && \\
&\text{subject to} && f_i(\mathbf{x}) - a_i z - y_i \leq f_i^{\max}, && i = 1, \ldots, m \\
& && x_j^{\min} \leq x_j \leq x_j^{\max}, && j = 1, \ldots, n \\
& && y_i \geq 0, && i = 1, \ldots, m \\
& && z \geq 0, &&
\end{aligned} \tag{1.1}$$

where $f_0, f_1, \ldots, f_m$ are given differentiable functions, while $x_j^{\min}$, $x_j^{\max}$, a_i, c_i and $f_i^{\max}$ are given real numbers which satisfy $x_j^{\min} < x_j^{\max}$, $a_i \geq 0$ and $c_i \geq 0$.

In problem (1.1), the "true" optimization variables are $x_1, \ldots, x_n$, while $y_1, \ldots, y_m$ and z are "artificial" optimization variables which should make it easier for the user to formulate and solve certain subclasses of problems, like least squares problems and minmax problems.

As a first example of how to transform a given problem to the form (1.1), assume that the user wants to solve a problem on the following "standard" form for nonlinear programming.

$$\begin{aligned}
&\text{minimize} && f_0(\mathbf{x}) && \\
&\text{subject to} && f_i(\mathbf{x}) \leq f_i^{\max}, && i = 1, \ldots, m \\
& && x_j^{\min} \leq x_j \leq x_j^{\max}, && j = 1, \ldots, n,
\end{aligned} \tag{1.2}$$

where $f_0, f_1, \ldots, f_m$ are given differentiable functions. To make problem (1.1) (almost) equivalent to this problem (1.2), first let $a_i = 0$ for all $i > 0$. Then $z = 0$ in any optimal solution to (1.1). Further, for each i, let $c_i =$ "a large number", so that the variables y_i become very "expensive". Then typically $\mathbf{y} = \mathbf{0}$ in any optimal solution to (1.1), and the corresponding $\mathbf{x}$ is an optimal solution to (1.2).

It should be noted that the problem (1.1) always has feasible solutions, and in fact also at least one optimal solution. This holds even if the user's problem (1.2) does not have any feasible solutions, in which case some $y_i > 0$ in the optimal solution of (1.1).

Now some practical considerations:

The user should preferably scale the constraints in such a way that $1 \le f_i^{\max} \le 100$ for each i (and not $f_i^{\max} = 10^{10}$). The objective function $f_0(\mathbf{x})$ should preferably be scaled such that $1 \le f_0(\mathbf{x}) \le 100$ for reasonable values on the variables. The variables x_j should preferably be scaled such that $0.1 \le x_j^{\max} - x_j^{\min} \le 100$, for all j.

Concerning the "large numbers" on the coefficients c_i mentioned above, the user should for numerical reasons try to avoid "extremely large" values on these coefficients (like 10^{10}). It is better to start with "reasonably large" values and then, if it turns out that not all $y_i = 0$ in the optimal solution of (1.1), increase the corresponding values of c_i by e.g. a factor 100 and solve the problem again, etc. If the functions and the variables have been scaled according to above, then "resonably large" values on the parameters c_i could be, say, $c_i = 1000$ or 10000.

Finally, concerning the simple bound constraints $x_j^{\min} \le x_j \le x_j^{\max}$, it may sometimes be the case that some variables x_j do not have any prescribed upper and/or lower bounds. In that case, it is in practice always possible to choose "artificial" bounds $x_j^{\min}$ and $x_j^{\max}$ such that every realistic solution $\mathbf{x}$ satisfies the corresponding bound constraints. The user should then preferably avoid choosing $x_j^{\max} - x_j^{\min}$ unnecessarily large. It is better to try some reasonable bounds and then, if it turns out that some variable x_j becomes equal to such an "artificial" bound in the optimal solution of (1.1), change this bound and solve the problem again (starting from the recently obtained solution), etc.

As a second example of how to transform a given problem to the form (1.1), assume that the user wants to solve a constrained least squares problem on the form

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\sum_{i=1}^{p}(h_i(\mathbf{x}) - \bar{h}_i)^2 \\ \text{subject to} & g_i(\mathbf{x}) \le g_i^{\max}, \qquad i = 1,\dots,q \\ & x_j^{\min} \le x_j \le x_j^{\max}, \quad j = 1,\dots,n \end{array} \tag{1.3}$$

where h_i and g_i are given differentiable functions, while $\bar{h}_i$ and $g_i^{\max}$ are given constants.

Problem (1.3) may equivalently be written on the following form with variables $\mathbf{x} \in I\!R^n$ and $y_1,\dots,y_{2p} \in I\!R$:

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\sum_{i=1}^{p}(y_i^2 + y_{p+i}^2) & \\ \text{subject to} & y_i \ge h_i(\mathbf{x}) - \bar{h}_i, & i = 1,\dots,p \\ & y_{p+i} \ge \bar{h}_i - h_i(\mathbf{x}), & i = 1,\dots,p \\ & g_i(\mathbf{x}) \le g_i^{\max}, & i = 1,\dots,q \\ & x_j^{\min} \le x_j \le x_j^{\max}, & j = 1,\dots,n \\ & y_i \ge 0, & i = 1,\dots,2p. \end{array} \tag{1.4}$$

To make problem (1.1) (almost) equivalent to this problem (1.4), let

$$\begin{array}{lcll}
m & = & 2p+q, & \\
f_0(\mathbf{x}) & = & 0, & \\
f_i(\mathbf{x}) & = & h_i(\mathbf{x}), & i=1,\ldots,p \\
f_i^{\max} & = & \bar{h}_i, & i=1,\ldots,p \\
f_{p+i}(\mathbf{x}) & = & -h_i(\mathbf{x}), & i=1,\ldots,p \\
f_{p+i}^{\max} & = & -\bar{h}_i, & i=1,\ldots,p \\
f_{2p+i}(\mathbf{x}) & = & g_i(\mathbf{x}), & i=1,\ldots,q \\
f_{2p+i}^{\max} & = & g_i^{\max}, & i=1,\ldots,q \\
a_i & = & 0, & i=1,\ldots,m \\
c_i & = & 0, & i=1,\ldots,2p \\
c_{2p+i} & = & \text{large number}, & i=1,\ldots,q.
\end{array}$$

As a third example of how to transform a given problem to the form (1.1), assume that the user wants to solve a "min-max" problem on the form

$$\begin{array}{rll}
\text{minimize} & \max\limits_{i=1,..,p}\{h_i(\mathbf{x})\} & \\
\text{subject to} & g_i(\mathbf{x}) \le g_i^{\max}, & i=1,\ldots,q \\
& x_j^{\min} \le x_j \le x_j^{\max}, & j=1,\ldots,n
\end{array} \tag{1.5}$$

where h_i and g_i are given differentiable functions, while $g_i^{\max}$ are given constants. For each given $\mathbf{x}$, the value of the objective function in problem (1.5) is the largest of the p real numbers $h_1(\mathbf{x}),\ldots,h_p(\mathbf{x})$. If there is a known number $h^{\min}$ such that $h_i(\mathbf{x}) \ge h^{\min}$ for all i and all feasible $\mathbf{x}$, then problem (1.5) may equivalently be written on the following form with variables $\mathbf{x} \in I\!R^n$ and $z \in I\!R$:

$$\begin{array}{rll}
\text{minimize} & z + h^{\min} & \\
\text{subject to} & z \ge h_i(\mathbf{x}) - h^{\min}, & i=1,\ldots,p \\
& g_i(\mathbf{x}) \le g_i^{\max}, & i=1,\ldots,q \\
& x_j^{\min} \le x_j \le x_j^{\max}, & j=1,\ldots,n \\
& z \ge 0. &
\end{array} \tag{1.6}$$

To make problem (1.1) (almost) equivalent to this problem (1.6), let

$$\begin{array}{lcll}
m & = & p+q, & \\
f_0(\mathbf{x}) & = & h^{\min}, \ \ \text{(a constant!)} & \\
f_i(\mathbf{x}) & = & h_i(\mathbf{x}), & i=1,\ldots,p \\
f_i^{\max} & = & h^{\min}, & i=1,\ldots,p \\
f_{p+i}(\mathbf{x}) & = & g_i(\mathbf{x}), & i=1,\ldots,q \\
f_{p+i}^{\max} & = & g_i^{\max}, & i=1,\ldots,q \\
a_i & = & 1, & i=1,\ldots,p \\
a_{p+i} & = & 0, & i=1,\ldots,q \\
c_i & = & \text{large number}, & i=1,\ldots,m
\end{array}$$

2 The ordinary MMA

MMA is a method for solving problems on the form (1.1), using the following approach: In each iteration $k \in \{1, 2, \ldots\}$, the current iteration point $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}, z^{(k)})$ is given. Then an approximating subproblem, in which the original functions f_i are replaced by certain convex functions $\tilde{f}_i^{(k)}$, is generated. The choice of these approximating functions is based mainly on gradient information at the current iteration point, but also on some parameters $u_j^{(k)}$ and $l_j^{(k)}$ ("moving asymptotes") which are updated in each iteration based on information from previous iteration points. The subproblem is solved, and the unique optimal solution becomes the next iteration point $(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)}, z^{(k+1)})$. Then a new subproblem is generated, etc.
A possible convergence criterium is that the iteration process is stopped when $|x_j^{(k+1)} - x_j^{(k)}| <$ $< \varepsilon \cdot (x_j^{\max} - x_j^{\min})$ for all $j = 1, \ldots, n$, where ε is a small number. The default value of this parameter ε, called XCHTOL in the fortran code, is 10^{-4}. Other convergence criteria are of course also possible.

The MMA subproblem looks as follows:

$$\begin{aligned}
&\text{minimize} && \tilde{f}_0^{(k)}(\mathbf{x}) + z + \tfrac{1}{2}d_0 z^2 + \sum_{i=1}^{m}(c_i y_i + \tfrac{1}{2}y_i^2) && \\
&\text{subject to} && \tilde{f}_i^{(k)}(\mathbf{x}) - a_i z - y_i \le f_i^{\max}, && i = 1, \ldots, m \\
& && \alpha_j^{(k)} \le x_j \le \beta_j^{(k)}, && j = 1, \ldots, n, \\
& && y_i \ge 0, && i = 1, \ldots, m \\
& && z \ge 0, &&
\end{aligned} \tag{2.1}$$

In this subproblem (2.1), the approximating functions $\tilde{f}_i^{(k)}(\mathbf{x})$ are chosen as

$$\tilde{f}_i^{(k)}(\mathbf{x}) = \sum_{j=1}^{n} \left(\frac{p_{ij}^{(k)}}{u_j^{(k)} - x_j} + \frac{q_{ij}^{(k)}}{x_j - l_j^{(k)}} \right) + r_i^{(k)}, \quad i = 0, 1, \ldots, m, \tag{2.2}$$

where

$$p_{ij}^{(k)} = (u_j^{(k)} - x_j^{(k)})^2 \left(1.001 \left(\frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \right)^+ + \, 0.001 \left(\frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \right)^- + \, \frac{10^{-5}}{x_j^{\max} - x_j^{\min}} \right), \tag{2.3}$$

$$q_{ij}^{(k)} = (x_j^{(k)} - l_j^{(k)})^2 \left(0.001 \left(\frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \right)^+ + \, 1.001 \left(\frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \right)^- + \, \frac{10^{-5}}{x_j^{\max} - x_j^{\min}} \right), \tag{2.4}$$

$$r_i^{(k)} = f_i(\mathbf{x}^{(k)}) - \sum_{j=1}^{n} \left(\frac{p_{ij}^{(k)}}{u_j^{(k)} - x_j^{(k)}} + \frac{q_{ij}^{(k)}}{x_j^{(k)} - l_j^{(k)}} \right). \tag{2.5}$$

Here, $\left(\dfrac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \right)^+$ denotes the largest of the two numbers $\dfrac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)})$ and 0,

while $\left(\dfrac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \right)^-$ denotes the largest of the two numbers $-\dfrac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)})$ and 0.

It follows from the formulas (2.2)–(2.5) that the functions $\tilde{f}_i^{(k)}$ are always first order approximations of the original functions f_i at the current iteration point, i.e.

$$\tilde{f}_i^{(k)}(\mathbf{x}^{(k)}) = f_i(\mathbf{x}^{(k)}) \quad \text{and} \quad \frac{\partial \tilde{f}_i^{(k)}}{\partial x_j}(\mathbf{x}^{(k)}) = \frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}). \tag{2.6}$$

Moreover, the approximating functions $\tilde{f}_0^{(k)}, \ldots, \tilde{f}_m^{(k)}$ are strictly convex. Based on this, it can be shown that there is always a unique optimal solution to the MMA subproblem.

An alternative, but equivalent, formulation of the MMA subproblem (2.1) is as follows:

$$\begin{aligned}
&\text{minimize} && \sum_{j=1}^{n}\left(\frac{p_{0j}^{(k)}}{u_j^{(k)} - x_j} + \frac{q_{0j}^{(k)}}{x_j - l_j^{(k)}}\right) + r_0^{(k)} + z + \tfrac{1}{2}d_0 z^2 + \sum_{i=1}^{m}(c_i y_i + \tfrac{1}{2}y_i^2)\\
&\text{subject to} && \sum_{j=1}^{n}\left(\frac{p_{ij}^{(k)}}{u_j^{(k)} - x_j} + \frac{q_{ij}^{(k)}}{x_j - l_j^{(k)}}\right) - a_i z - y_i \le b_i^{(k)}, \quad i = 1, \ldots, m\\
& && \alpha_j^{(k)} \le x_j \le \beta_j^{(k)}, \quad j = 1, \ldots, n,\\
& && y_i \ge 0, \quad i = 1, \ldots, m\\
& && z \ge 0,
\end{aligned} \tag{2.7}$$

where $b_i^{(k)} = f_i^{\max} - r_i^{(k)}$ for $i = 1, \ldots, m$.

The bounds $\alpha_j^{(k)}$ and $\beta_j^{(k)}$ in (2.1) and (2.7) are chosen as

$$\alpha_j^{(k)} = \max\{\, x_j^{\min},\ \ l_j^{(k)} + 0.1(x_j^{(k)} - l_j^{(k)}),\ \ x_j^{(k)} - 0.5(x_j^{\max} - x_j^{\min})\,\}, \tag{2.8}$$

$$\beta_j^{(k)} = \min\{\, x_j^{\max},\ \ u_j^{(k)} - 0.1(u_j^{(k)} - x_j^{(k)}),\ \ x_j^{(k)} + 0.5(x_j^{\max} - x_j^{\min})\,\}, \tag{2.9}$$

which means that the constraints $\alpha_j^{(k)} \le x_j \le \beta_j^{(k)}$ are equivalent to the following three sets of constraints:

$$x_j^{\min} \le x_j \le x_j^{\max}, \tag{2.10}$$

$$-0.9(x_j^{(k)} - l_j^{(k)}) \le x_j - x_j^{(k)} \le 0.9(u_j^{(k)} - x_j^{(k)}), \tag{2.11}$$

$$-0.5(x_j^{\max} - x_j^{\min}) \le x_j - x_j^{(k)} \le 0.5(x_j^{\max} - x_j^{\min}). \tag{2.12}$$

The default rules for updating the lower asymptotes $l_j^{(k)}$ and the upper asymptotes $u_j^{(k)}$ are as follows. The first two iterations, when $k = 1$ and $k = 2$,

$$\begin{aligned}
l_j^{(k)} &= x_j^{(k)} - 0.5(x_j^{\max} - x_j^{\min}),\\
u_j^{(k)} &= x_j^{(k)} + 0.5(x_j^{\max} - x_j^{\min}).
\end{aligned} \tag{2.13}$$

In later iterations, when $k \geq 3$,

$$\begin{array}{rcl} l_j^{(k)} & = & x_j^{(k)} - \gamma_j^{(k)}(x_j^{(k-1)} - l_j^{(k-1)}), \\ u_j^{(k)} & = & x_j^{(k)} + \gamma_j^{(k)}(u_j^{(k-1)} - x_j^{(k-1)}), \end{array} \tag{2.14}$$

where

$$\gamma_j^{(k)} = \begin{cases} 0.7 & \text{if} \quad (x_j^{(k)} - x_j^{(k-1)})(x_j^{(k-1)} - x_j^{(k-2)}) < 0, \\ 1.2 & \text{if} \quad (x_j^{(k)} - x_j^{(k-1)})(x_j^{(k-1)} - x_j^{(k-2)}) > 0, \\ 1 & \text{if} \quad (x_j^{(k)} - x_j^{(k-1)})(x_j^{(k-1)} - x_j^{(k-2)}) = 0, \end{cases} \tag{2.15}$$

provided that this leads to values that satisfy

$$\begin{array}{rcl} l_j^{(k)} & \leq & x_j^{(k)} - 0.01(x_j^{\max} - x_j^{\min}), \\ l_j^{(k)} & \geq & x_j^{(k)} - 10(x_j^{\max} - x_j^{\min}), \\ u_j^{(k)} & \geq & x_j^{(k)} + 0.01(x_j^{\max} - x_j^{\min}), \\ u_j^{(k)} & \leq & x_j^{(k)} + 10(x_j^{\max} - x_j^{\min}). \end{array} \tag{2.16}$$

If any of these bounds is violated, the corresponding $l_j^{(k)}$ or $u_j^{(k)}$ is put to the right hand side of the violated inequality.

Note that most of the explicit numbers in the above expressions are just default values of different parameters in the fortran code. More precisely:

The number 10^{-5} in (2.3) and (2.4) is the default value of the parameter `RAAI`.
The number 0.1 in (2.8) and (2.9) is the default value of the parameter `ALBEFA`.
The number 0.5 in (2.8) and (2.9) is the default value of the parameter `XXMOVE`.
The number 0.5 in (2.13) is the default value of the parameter `GHINIT`.
The number 0.7 in (2.15) is the default value of the parameter `GHDECR`.
The number 1.2 in (2.15) is the default value of the parameter `GHINCR`.
The number 0.01 in (2.16) is the default value of the parameter `ASYMIN`.
The number 10.0 in (2.16) is the default value of the parameter `ASYMAX`.

All these values can be carefully changed by the user. As an example, a more conservative method is obtain by decreasing `XXMOVE` and/or `ASYMAX` and/or `GHINIT` and/or `GHINCR`.

3 GCMMA – the globally convergent version of MMA

The globally convergent version of MMA, from now on called GCMMA, for solving problems of the form (1.1) consists of "outer" and "inner" iterations. The index k is used to denote the outer iteration number, while the index ν is used to denote the inner iteration number. Within each outer iteration, there may be zero, one, or several inner iterations. The double index (k,ν) is used to denote the ν:th inner iteration within the k:th outer iteration.

The first iteration point is obtained by first chosing $\mathbf{x}^{(1)} \in X$ and then chosing $\mathbf{y}^{(1)}$ and $z^{(1)}$ such that $(\mathbf{x}^{(1)},\mathbf{y}^{(1)},z^{(1)})$ becomes a feasible solution of (1.1). This is easy. An outer iteration of the method, going from the k:th iteration point $(\mathbf{x}^{(k)},\mathbf{y}^{(k)},z^{(k)})$ to the $(k+1)$:th iteration point $(\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)},z^{(k+1)})$, can be described as follows:

Given $(\mathbf{x}^{(k)},\mathbf{y}^{(k)},z^{(k)})$, an approximating subproblem is generated and solved. In this subproblem, the functions $f_i(\mathbf{x})$ are replaced by certain convex functions $\tilde{f}_i^{(k,0)}(\mathbf{x})$. The optimal solution of this subproblem is denoted $(\hat{\mathbf{x}}^{(k,0)},\hat{\mathbf{y}}^{(k,0)},\hat{z}^{(k,0)})$. If $\tilde{f}_i^{(k,0)}(\hat{\mathbf{x}}^{(k,0)}) \geq f_i(\hat{\mathbf{x}}^{(k,0)})$, for all $i=0,1,\ldots,m$, the next iteration point becomes $(\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)},z^{(k+1)}) = (\hat{\mathbf{x}}^{(k,0)},\hat{\mathbf{y}}^{(k,0)},\hat{z}^{(k,0)})$, and the outer iteration is completed (without any inner iterations needed). Otherwise, an inner iteration is made, which means that a new subproblem is generated and solved at $\mathbf{x}^{(k)}$, with new approximating functions $\tilde{f}_i^{(k,1)}(\mathbf{x})$ which are more conservative than $\tilde{f}_i^{(k,0)}(\mathbf{x})$ for those indices i for which the above inequality was violated. The optimal solution of this new subproblem is denoted $(\hat{\mathbf{x}}^{(k,1)},\hat{\mathbf{y}}^{(k,1)},\hat{z}^{(k,1)})$. If $\tilde{f}_i^{(k,1)}(\hat{\mathbf{x}}^{(k,1)}) \geq f_i(\hat{\mathbf{x}}^{(k,1)})$, for all $i = 0,1,\ldots,m$, the next iteration point becomes $(\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)},z^{(k+1)}) = (\hat{\mathbf{x}}^{(k,1)},\hat{\mathbf{y}}^{(k,1)},\hat{z}^{(k,1)})$, and the outer iteration is completed (with one inner iterations needed). Otherwise, another inner iteration is made, which means that a new subproblem is generated and solved at $\mathbf{x}^{(k)}$, with new approximating functions $\tilde{f}_i^{(k,2)}(\mathbf{x})$, etc. These inner iterations are repeated until $\tilde{f}_i^{(k,\nu)}(\hat{\mathbf{x}}^{(k,\nu)}) \geq f_i(\hat{\mathbf{x}}^{(k,\nu)})$ for all $i=0,1,\ldots,m$, which always happens after a finite (usually small) number of inner iterations. Then the next iteration point becomes $(\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)},z^{(k+1)}) = (\hat{\mathbf{x}}^{(k,\nu)},\hat{\mathbf{y}}^{(k,\nu)},\hat{z}^{(k,\nu)})$, and the outer iteration is completed (with ν inner iterations needed).

It should be noted that in each inner iteration, there is no need to recalculate the gradients $\nabla f_i(\mathbf{x}^{(k)})$, since $\mathbf{x}^{(k)}$ has not changed. Gradients of the original functions f_i are calculated only once in each outer iteration. This is an important note since the calculation of gradients is typically the most time consuming part in structural optimization.

The GCMMA subproblem looks as follows, for $k \in \{1,2,3,\ldots\}$ and $\nu \in \{0,1,2,\ldots\}$:

$$\begin{aligned}
&\text{minimize} && \tilde{f}_0^{(k,\nu)}(\mathbf{x}) + z + \tfrac{1}{2}d_0 z^2 + \sum_{i=1}^{m}(c_i y_i + \tfrac{1}{2}y_i^2) && \\
&\text{subject to} && \tilde{f}_i^{(k,\nu)}(\mathbf{x}) - a_i z - y_i \leq f_i^{\max}, && i=1,\ldots,m \\
& && \alpha_j^{(k)} \leq x_j \leq \beta_j^{(k)}, && j=1,\ldots,n, \\
& && y_i \geq 0, && i=1,\ldots,m \\
& && z \geq 0, &&
\end{aligned} \tag{3.1}$$

where the approximating functions $\tilde{f}_i^{(k,\nu)}(\mathbf{x})$ are chosen as

$$\tilde{f}_i^{(k,\nu)}(\mathbf{x}) = \sum_{j=1}^{n} \left(\frac{p_{ij}^{(k,\nu)}}{u_j^{(k)} - x_j} + \frac{q_{ij}^{(k,\nu)}}{x_j - l_j^{(k)}} \right) + r_i^{(k,\nu)}, \quad i = 0, 1, \ldots, m\,. \tag{3.2}$$

Here,

$$p_{ij}^{(k,\nu)} = (u_j^{(k)} - x_j^{(k)})^2 \left(1.001 \Big(\frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \Big)^+ + \; 0.001 \Big(\frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \Big)^- + \; \frac{\rho_i^{(k,\nu)}}{x_j^{\max} - x_j^{\min}} \right), \tag{3.3}$$

$$q_{ij}^{(k,\nu)} = (x_j^{(k)} - l_j^{(k)})^2 \left(0.001 \Big(\frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \Big)^+ + \; 1.001 \Big(\frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \Big)^- + \; \frac{\rho_i^{(k,\nu)}}{x_j^{\max} - x_j^{\min}} \right), \tag{3.4}$$

$$r_i^{(k,\nu)} = f_i(\mathbf{x}^{(k)}) - \sum_{j=1}^{n} \left(\frac{p_{ij}^{(k,\nu)}}{u_j^{(k)} - x_j^{(k)}} + \frac{q_{ij}^{(k,\nu)}}{x_j^{(k)} - l_j^{(k)}} \right). \tag{3.5}$$

Between each outer iteration, the bounds $\alpha_j^{(k)}$ and $\beta_j^{(k)}$ and the asymptotes $l_j^{(k)}$ and $u_j^{(k)}$ are updated as in the original MMA, the formulas (2.8)–(2.16) still hold.
The parameters $\rho_i^{(k,\nu)}$ in (3.3) and (3.4) are strictly positive and updated as follows: Within a given outer iteration k, the only differences between two inner iterations are the values of some of these parameters. In the beginning of each outer iteration, when $\nu = 0$, the following values are used:

$$\rho_i^{(k,0)} = \max\{ \, \rho^{\min}, \; \frac{0.1}{n} \sum_{j=1}^{n} \left| \frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}) \right| (x_j^{\max} - x_j^{\min}) \, \} \,, \quad \text{for } \; i = 0, 1, .., m, \tag{3.6}$$

where the default value for the parameter $\rho^{\min}$, called `RAAMIN` in the fortran code, is 10^{-6}.

In each new inner iteration, the updating of $\rho_i^{(k,\nu)}$ is based on the solution of the most recent subproblem. Note that $\tilde{f}_i^{(k,\nu)}(\mathbf{x})$ may be written on the form:

$$\tilde{f}_i^{(k,\nu)}(\mathbf{x}) = h_i^{(k)}(\mathbf{x}) + \rho_i^{(k,\nu)} d^{(k)}(\mathbf{x}),$$

where $h_i^{(k)}(\mathbf{x})$ and $d^{(k)}(\mathbf{x})$ do not depend on $\rho_i^{(k,\nu)}$. Some calculations give that

$$d^{(k)}(\mathbf{x}) = \sum_{j=1}^{n} \frac{(u_j^{(k)} - l_j^{(k)})(x_j - x_j^{(k)})^2}{(u_j^{(k)} - x_j)(x_j - l_j^{(k)})(x_j^{\max} - x_j^{\min})} \,. \tag{3.7}$$

Now, let

$$\delta_i^{(k,\nu)} = \frac{f_i(\hat{\mathbf{x}}^{(k,\nu)}) - \tilde{f}_i^{(k,\nu)}(\hat{\mathbf{x}}^{(k,\nu)})}{d^{(k)}(\hat{\mathbf{x}}^{(k,\nu)})} \,. \tag{3.8}$$

Then $h_i^{(k)}(\hat{\mathbf{x}}^{(k,\nu)}) + (\rho_i^{(k,\nu)} + \delta_i^{(k,\nu)}) d^{(k)}(\hat{\mathbf{x}}^{(k,\nu)}) = f_i(\hat{\mathbf{x}}^{(k,\nu)})$, which shows that $\rho_i^{(k,\nu)} + \delta_i^{(k,\nu)}$ might be a natural value of $\rho_i^{(k,\nu+1)}$. In order to get a globally convergent method, this natural value is modified as follows.

$$\begin{aligned} \rho_i^{(k,\nu+1)} &= \min\{ \, 1.1 \, (\rho_i^{(k,\nu)} + \delta_i^{(k,\nu)}) \; , \; 10 \rho_i^{(k,\nu)} \} & \text{if } \delta_i^{(k,\nu)} > 0, \\ \rho_i^{(k,\nu+1)} &= \rho_i^{(k,\nu)} & \text{if } \delta_i^{(k,\nu)} \leq 0. \end{aligned} \tag{3.9}$$

It follows from the formulas (3.2)–(3.5) that the functions $\tilde{f}_i^{(k,\nu)}$ are always first order approximations of the original functions f_i at the current iteration point, i.e.

$$\tilde{f}_i^{(k,\nu)}(\mathbf{x}^{(k)}) = f_i(\mathbf{x}^{(k)}) \;\text{ and }\; \frac{\partial \tilde{f}_i^{(k,\nu)}}{\partial x_j}(\mathbf{x}^{(k)}) = \frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)}). \tag{3.10}$$

Since the parameters $\rho_i^{(k,\nu)}$ are always strictly positive, the functions $\tilde{f}_i^{(k,\nu)}$ are strictly convex. Based on this, it can be shown that there is always a unique optimal solution to the GCMMA subproblem.

4 A dual method for solving the subproblems

In the fortran implementation, a dual approach based on Lagrangean relaxation is used for solving the subproblems in both MMA and GCMMA. The dual problem, which is a maximization problem with a concave objective function and no other constraints than non-negativity requirements on the (dual) variables, is solved by a modified Newton method, combined with an "active set strategy" to handle the non-negativity constraints, whereafter the optimal dual solution is translated to a corresponding optimal solution of the primal MMA subproblem. The purpose of this section is to describe this approach in more details.

The MMA and GCMMA subproblems are in this section written as

$$\begin{aligned}
&\text{minimize} && \tilde{f}_0(\mathbf{x}) + z + \tfrac{1}{2}d_0 z^2 + \sum_{i=1}^{m}(c_i y_i + \tfrac{1}{2}y_i^2) && \\
&\text{subject to} && \tilde{f}_i(\mathbf{x}) - a_i z - y_i \leq b_i\,, && i = 1,\ldots,m \\
& && \alpha_j \leq x_j \leq \beta_j\,, && j = 1,\ldots,n \\
& && z \geq 0, \quad y_i \geq 0, && i = 1,\ldots,m
\end{aligned} \tag{4.1}$$

where

$$\tilde{f}_i(\mathbf{x}) = \sum_{j=1}^{n} \tilde{f}_{ij}(x_j) = \sum_{j=1}^{n} \left(\frac{p_{ij}}{u_j - x_j} + \frac{q_{ij}}{x_j - l_j} \right), \quad \text{for } i = 0,1,\ldots,m. \tag{4.2}$$

The constraints $\tilde{f}_i(\mathbf{x}) - a_i z - y_i \leq b_i$ will be called the *explicit* constraints, while the constraints $\alpha_j \leq x_j \leq \beta_j$, $z \geq 0$ and $y_i \geq 0$ will be called the *implicit* constraints.

4.1 Lagrangean relaxation and the dual subproblem

A Lagrangean relaxation of the problem, with respect to the *explicit* constraints only, gives rise to the following Lagrange function:

$$L(\mathbf{x},\mathbf{y},z,\boldsymbol{\lambda}) = \tilde{f}_0(\mathbf{x}) + z + \tfrac{1}{2}d_0 z^2 + \sum_{i=1}^{m}(c_i y_i + \tfrac{1}{2}y_i^2) + \sum_{i=1}^{m}\lambda_i(\tilde{f}_i(\mathbf{x}) - a_i z - y_i - b_i), \tag{4.3}$$

where $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_m)^{\mathsf{T}}$ is the vector of *non-negative* Lagrange multipliers λ_i for the explicit constraints. This Lagrange function can be written

$$L(\mathbf{x},\mathbf{y},z,\boldsymbol{\lambda}) = L^x(\mathbf{x},\boldsymbol{\lambda}) + L^y(\mathbf{y},\boldsymbol{\lambda}) + L^z(z,\boldsymbol{\lambda}) - \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{b}, \quad \text{where} \tag{4.4}$$

$$L^x(\mathbf{x},\boldsymbol{\lambda}) = \tilde{f}_0(\mathbf{x}) + \sum_{i=1}^{m}\lambda_i\tilde{f}_i(\mathbf{x}) = \sum_{j=1}^{n}\left(\frac{p_j(\boldsymbol{\lambda})}{u_j - x_j} + \frac{q_j(\boldsymbol{\lambda})}{x_j - l_j}\right) = \sum_{j=1}^{n} L_j^x(x_j,\boldsymbol{\lambda}), \tag{4.5}$$

$$L^y(\mathbf{y},\boldsymbol{\lambda}) = \sum_{i=1}^{m}(c_i y_i + \tfrac{1}{2}y_i^2 - \lambda_i y_i)\,, \tag{4.6}$$

$$L^z(z,\boldsymbol{\lambda}) = (1 - \boldsymbol{\lambda}^{\mathsf{T}}a)\, z + \tfrac{1}{2}d_0 z^2. \tag{4.7}$$

Here,

$$p_j(\boldsymbol{\lambda}) = p_{0j} + \sum_i^m \lambda_i p_{ij} > 0 \quad \text{and} \quad q_j(\boldsymbol{\lambda}) = q_{0j} + \sum_i^m \lambda_i q_{ij} > 0. \tag{4.8}$$

The dual function value $\varphi(\boldsymbol{\lambda})$ at the point $\boldsymbol{\lambda} \geq \mathbf{0}$ is by definition obtained as the optimal value of the problem of minimizing the Lagrange function with respect to the "primal variables" $\mathbf{x}$, $\mathbf{y}$ and z, subject to the *implicit* constraints only, i.e. as the optimal value of the problem

$$\begin{aligned} &\text{minimize} && L(\mathbf{x},\mathbf{y},z,\boldsymbol{\lambda}) \\ &\text{subject to} && \alpha_j \leq x_j \leq \beta_j\,, \quad j = 1,\ldots,n, \\ & && z \geq 0, \quad y_i \geq 0, \quad i = 1,\ldots,m. \end{aligned} \tag{4.9}$$

Note that, in this problem (4.9), the vector $\boldsymbol{\lambda}$ is held fixed.
Due to (4.4), problem (4.9) can be divided into three separate problems which can be solved independently of each other. First one problem in $\mathbf{x} \in I\!\!R^n$, namely

$$\begin{aligned} &\text{minimize} && L^x(\mathbf{x},\boldsymbol{\lambda}) = \sum_{j=1}^n L_j^x(x_j,\boldsymbol{\lambda}) \\ &\text{subject to} && \alpha_j \leq x_j \leq \beta_j\,, \quad j = 1,\ldots,n, \end{aligned} \tag{4.10}$$

which in turn can be divided into n separate problems, each involving only *one* variable x_j :

$$\text{minimize} \;\; L_j^x(x_j,\boldsymbol{\lambda}) = \frac{p_j(\boldsymbol{\lambda})}{u_j - x_j} + \frac{q_j(\boldsymbol{\lambda})}{x_j - l_j} \;\; \text{subject to} \;\; \alpha_j \leq x_j \leq \beta_j\,. \tag{4.11}$$

Then one problem in $\mathbf{y} \in I\!\!R^m$, namely

$$\begin{aligned} &\text{minimize} && L^y(\mathbf{y},\boldsymbol{\lambda}) \\ &\text{subject to} && y_i \geq 0, \quad i = 1,\ldots,m, \end{aligned} \tag{4.12}$$

which in turn can be divided into m separate problems, each involving only *one* variable y_i :

$$\text{minimize} \;\; (c_i - \lambda_i)\, y_i + \tfrac{1}{2} y_i^2 \;\; \text{subject to} \;\; y_i \geq 0. \tag{4.13}$$

And finally one problem in $z \in I\!\!R$, namely

$$\text{minimize} \;\; L^z(z,\boldsymbol{\lambda}) = (1 - \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a})\, z + \tfrac{1}{2} d_0 z^2 \;\; \text{subject to} \;\; z \geq 0. \tag{4.14}$$

Each of the three problems (4.11), (4.13) and (4.14) can easily be solved analytically.
The optimal solutions $\hat{y}_i(\boldsymbol{\lambda})$ and $\hat{z}(\boldsymbol{\lambda})$ of (4.13) and (4.14), respectively, are given by

$$\hat{y}_i(\boldsymbol{\lambda}) = (\lambda_i - c_i)^+ \;\; \text{and} \;\; \hat{z}(\boldsymbol{\lambda}) = \frac{(\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a} - 1)^+}{d_0}\,, \tag{4.15}$$

where $(\lambda_i - c_i)^+$ denotes the largest of the two numbers $\lambda_i - c_i$ and 0,
and $(\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a} - 1)^+$ denotes the largest of the two numbers $\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a} - 1$ and 0.

The optimal values of the problems (4.12) and (4.14), respectively, are then given by

$$L^y(\hat{\mathbf{y}}(\boldsymbol{\lambda}),\boldsymbol{\lambda}) = -\tfrac{1}{2}\sum_{i=1}^{m}((\lambda_i - c_i)^+)^2 = -\tfrac{1}{2}\sum_{i=1}^{m}\hat{y}_i(\boldsymbol{\lambda})^2 \tag{4.16}$$

and

$$L^z(\hat{z}(\boldsymbol{\lambda}),\boldsymbol{\lambda}) = -\frac{((\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a}-1)^+)^2}{2d_0} = -\tfrac{1}{2}d_0\hat{z}(\boldsymbol{\lambda})^2\ . \tag{4.17}$$

Let $\xi_j(\boldsymbol{\lambda})$ denote the unique number $x_j \in (l_j, u_j)$ which satisfies the equation

$$\frac{\partial L_j^x}{\partial x_j}(x_j,\boldsymbol{\lambda}) = 0, \tag{4.18}$$

where $L_j^x(x_j,\boldsymbol{\lambda})$ is defined in (4.11). Straigh-forward calculations give that

$$\xi_j(\boldsymbol{\lambda}) \;=\; \frac{l_j\sqrt{p_j(\boldsymbol{\lambda})}\;+\;u_j\sqrt{q_j(\boldsymbol{\lambda})}}{\sqrt{p_j(\boldsymbol{\lambda})}\;+\;\sqrt{q_j(\boldsymbol{\lambda})}}. \tag{4.19}$$

Then the optimal solution $\hat{x}_j(\boldsymbol{\lambda})$ of (4.11) is given by

$$\hat{x}_j(\boldsymbol{\lambda}) = \begin{cases} \alpha_j & \text{if } \ \xi_j(\boldsymbol{\lambda}) \le \alpha_j, \\ \xi_j(\boldsymbol{\lambda}) & \text{if } \ \alpha_j < \xi_j(\boldsymbol{\lambda}) < \beta_j, \\ \beta_j & \text{if } \ \xi_j(\boldsymbol{\lambda}) \ge \beta_j, \end{cases} \tag{4.20}$$

The optimal value of the problems (4.10) is then given by

$$L^x(\hat{\mathbf{x}}(\boldsymbol{\lambda}),\boldsymbol{\lambda}) = \sum_{j=1}^{n} L_j^x(\hat{x}_j(\boldsymbol{\lambda}),\boldsymbol{\lambda}) = \sum_{j=1}^{n}\left(\frac{p_j(\boldsymbol{\lambda})}{u_j - \hat{x}_j(\boldsymbol{\lambda})} + \frac{q_j(\boldsymbol{\lambda})}{\hat{x}_j(\boldsymbol{\lambda}) - l_j}\right). \tag{4.21}$$

Finally, the optimal value of the problem (4.9) is given by

$$L(\hat{\mathbf{x}}(\boldsymbol{\lambda}),\hat{\mathbf{y}}(\boldsymbol{\lambda}),\hat{z}(\boldsymbol{\lambda}),\boldsymbol{\lambda}) = L^x(\hat{\mathbf{x}}(\boldsymbol{\lambda}),\boldsymbol{\lambda}) + L^y(\hat{\mathbf{y}}(\boldsymbol{\lambda}),\boldsymbol{\lambda}) + L^z(\hat{z}(\boldsymbol{\lambda}),\boldsymbol{\lambda}) - \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{b}, \tag{4.22}$$

which by definition is the dual function value $\varphi(\boldsymbol{\lambda})$ at the point $\boldsymbol{\lambda}$. Thus,

$$\varphi(\boldsymbol{\lambda}) = \sum_{j=1}^{n}\left(\frac{p_j(\boldsymbol{\lambda})}{u_j-\hat{x}_j(\boldsymbol{\lambda})} + \frac{q_j(\boldsymbol{\lambda})}{\hat{x}_j(\boldsymbol{\lambda})-l_j}\right) - \tfrac{1}{2}\sum_{i=1}^{m}((\lambda_i-c_i)^+)^2 - \frac{((\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a}-1)^+)^2}{2d_0} - \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{b}\,. \tag{4.23}$$

The *dual problem* D is by definition the following maximization problem.

$$\text{D:}\ \text{ maximize } \varphi(\boldsymbol{\lambda}) \ \text{ subject to } \boldsymbol{\lambda} \ge \mathbf{0}. \tag{4.24}$$

The following theorem, which is proved later, provides the main motivation for considering this dual problem D.

Theorem 1: Assume that $\hat{\boldsymbol{\lambda}}$ is an optimal solution to the dual problem (4.24), and let $(\hat{\mathbf{x}},\ \hat{\mathbf{y}},\ \hat{z})$ be the corresponding primal solution obtained from the formulas (4.20) and (4.15) with $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$.
Then $(\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{z})$ is the unique optimal solution to the primal subproblem (4.1).
Moreover, the optimal values of the problems (4.1) and (4.24) are equal.

4.2 Gradient and Hessian of the dual function

In this section, explicit expressions for the gradient and the Hessian of the dual function φ are derived. These are needed when a modified Newton method is used for solving the dual problem. It turns out that the first order derivatives of φ are continuous everywhere, while the second order derivatives are not.

First, some simplifying notations are introduced. Let

$\mathbf{w} = (\mathbf{x}^\mathsf{T}, \mathbf{y}^\mathsf{T}, z)^\mathsf{T}$,

$W = \{\, \mathbf{w} = (\mathbf{x}^\mathsf{T}, \mathbf{y}^\mathsf{T}, z)^\mathsf{T} \mid \alpha_j \leq x_j \leq \beta_j, \forall j,\ y_i \geq 0, \forall i,\ z \geq 0 \,\}$,

$g_0(\mathbf{w}) = \tilde{f}_0(\mathbf{x}) + z + \frac{1}{2} d_0 z^2 + \sum_i (c_i y_i + \frac{1}{2} y_i^2)$,

$g_i(\mathbf{w}) = \tilde{f}_i(\mathbf{x}) - a_i z - y_i - b_i$,

$\mathbf{g}(\mathbf{w}) = (g_1(\mathbf{w}), \ldots, g_m(\mathbf{w}))^\mathsf{T}$.

Then the primal subproblem (4.1) may be written

$$\begin{aligned} \mathrm{P}: \quad & \text{minimize} && g_0(\mathbf{w}) \\ & \text{subject to} && g_i(\mathbf{w}) \leq 0, \quad i = 1, \ldots, m \\ & && \mathbf{w} \in W. \end{aligned} \tag{4.25}$$

The corresponding Lagrange function, which is the same Lagrange function as in (4.3), is

$$L(\mathbf{w}, \boldsymbol{\lambda}) = g_0(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) = g_0(\mathbf{w}) + \boldsymbol{\lambda}^\mathsf{T} \mathbf{g}(\mathbf{w}), \tag{4.26}$$

while the corresponding dual function, which is the same dual function as in (4.23), is

$$\varphi(\boldsymbol{\lambda}) = \min_{\mathbf{w} \in W} L(\mathbf{w}, \boldsymbol{\lambda}) = L(\hat{\mathbf{w}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = g_0(\hat{\mathbf{w}}(\boldsymbol{\lambda})) + \boldsymbol{\lambda}^\mathsf{T} \mathbf{g}(\hat{\mathbf{w}}(\boldsymbol{\lambda})), \tag{4.27}$$

where $\hat{\mathbf{w}}(\boldsymbol{\lambda}) = (\hat{\mathbf{x}}(\boldsymbol{\lambda})^\mathsf{T}, \hat{\mathbf{y}}(\boldsymbol{\lambda})^\mathsf{T}, \hat{z}(\boldsymbol{\lambda}))^\mathsf{T}$ is the unique $\mathbf{w} \in W$ which minimizes $L(\mathbf{w}, \boldsymbol{\lambda})$ on W. The different components in the vector $\hat{\mathbf{w}}(\boldsymbol{\lambda})$ are given by the formulas (4.20) and (4.15).

Lemma 4.2.1: If $\boldsymbol{\lambda} \in \Lambda$ and $\tilde{\boldsymbol{\lambda}} \in \Lambda$ then

$$\varphi(\boldsymbol{\lambda}) - \varphi(\tilde{\boldsymbol{\lambda}}) \leq (\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}})^\mathsf{T} \mathbf{g}(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}})), \tag{4.28}$$

$$\varphi(\tilde{\boldsymbol{\lambda}}) - \varphi(\boldsymbol{\lambda}) \leq (\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda})^\mathsf{T} \mathbf{g}(\hat{\mathbf{w}}(\boldsymbol{\lambda})). \tag{4.29}$$

Proof: First, (4.28) follows from

$\varphi(\boldsymbol{\lambda}) - \varphi(\tilde{\boldsymbol{\lambda}}) = L(\, \hat{\mathbf{w}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) - L(\, \hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) \leq L(\, \hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}}), \boldsymbol{\lambda}) - L(\, \hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}}), \tilde{\boldsymbol{\lambda}}) = (\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}})^\mathsf{T} \mathbf{g}(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}}))$.

Then (4.29) follows by simply letting $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\lambda}}$ change place in (4.28).

Lemma 4.2.2: If $\tilde{\boldsymbol{\lambda}} \in \Lambda$ and $\mathbf{d} \in I\!R^m$ then

$$\lim_{t \to 0} \frac{\varphi(\tilde{\boldsymbol{\lambda}} + t\,\mathbf{d}) - \varphi(\tilde{\boldsymbol{\lambda}})}{t} = \mathbf{g}(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}}))^{\mathsf{T}}\mathbf{d}. \tag{4.30}$$

Proof: Let $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}} + t\,\mathbf{d}$. For $t > 0$, the inequalitites (4.28) and (4.29) imply that:

$$\mathbf{g}(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}} + t\,\mathbf{d}))^{\mathsf{T}}\mathbf{d} \leq \frac{\varphi(\tilde{\boldsymbol{\lambda}} + t\,\mathbf{d}) - \varphi(\tilde{\boldsymbol{\lambda}})}{t} \leq \mathbf{g}(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}}))^{\mathsf{T}}\mathbf{d},$$

while for $t < 0$, they imply that:

$$\mathbf{g}(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}}))^{\mathsf{T}}\mathbf{d} \leq \frac{\varphi(\tilde{\boldsymbol{\lambda}} + t\,\mathbf{d}) - \varphi(\tilde{\boldsymbol{\lambda}})}{t} \leq \mathbf{g}(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}} + t\,\mathbf{d}))^{\mathsf{T}}\mathbf{d}.$$

The lemma now follows from the fact that $\mathbf{g}(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}} + t\,\mathbf{d})) \to \mathbf{g}(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}}))$ when $t \to 0$.

Lemma 4.2.3: The partial derivatives of the dual function φ, at a given point $\boldsymbol{\lambda} \in \Lambda$, are given by

$$\frac{\partial \varphi}{\partial \lambda_i}(\boldsymbol{\lambda}) = g_i(\hat{\mathbf{w}}(\boldsymbol{\lambda})),\ i = 1, \ldots, m. \tag{4.31}$$

Proof: Follows from Lemma 4.2.2 by letting $\mathbf{d} = \mathbf{e}_i$.

Lemma 4.2.4: The partial derivatives of the dual function φ, at a given point $\boldsymbol{\lambda} \in \Lambda$, are given by

$$\frac{\partial \varphi}{\partial \lambda_i}(\boldsymbol{\lambda}) = \sum_{j=1}^{n} \left(\frac{p_{ij}}{u_j - \hat{x}_j(\boldsymbol{\lambda})} + \frac{q_{ij}}{\hat{x}_j(\boldsymbol{\lambda}) - l_j} \right) - \frac{a_i}{d_0}(\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a} - 1)^+ - (\lambda_i - c_i)^+ - b_i\,. \tag{4.32}$$

where $\hat{x}_j(\boldsymbol{\lambda})$ is defined by the formula (4.20).

Proof:
According to Lemma 4.2.3, and the definition of $g_i(\mathbf{w})$, the partial derivative of the dual function are given by

$$\frac{\partial \varphi}{\partial \lambda_i}(\boldsymbol{\lambda}) = \tilde{f}_i(\hat{\mathbf{x}}(\boldsymbol{\lambda})) - a_i \hat{z}(\boldsymbol{\lambda}) - \hat{y}_i(\boldsymbol{\lambda}) - b_i\,. \tag{4.33}$$

The lemma now follows from the definition of $\tilde{f}_i(\mathbf{x})$ and the formulas (4.15).

Since these partial derivatives of φ are continuous, the dual function φ is a continuously differentiable function.

Now to the Hessian matrix. The dual function can be written

$$\varphi(\boldsymbol{\lambda}) = \varphi^x(\boldsymbol{\lambda}) + \varphi^y(\boldsymbol{\lambda}) + \varphi^z(\boldsymbol{\lambda}) - \boldsymbol{\lambda}^{\mathsf{T}} b, \quad \text{where} \tag{4.34}$$

$$\varphi^x(\boldsymbol{\lambda}) = \sum_{j=1}^{n} \left(\frac{p_j(\boldsymbol{\lambda})}{u_j - \hat{x}_j(\boldsymbol{\lambda})} + \frac{q_j(\boldsymbol{\lambda})}{\hat{x}_j(\boldsymbol{\lambda}) - l_j} \right) = \sum_{j=1}^{n} \varphi_j^x(\boldsymbol{\lambda}), \tag{4.35}$$

$$\varphi^y(\boldsymbol{\lambda}) = -\tfrac{1}{2} \sum_{i=1}^{m} ((\lambda_i - c_i)^+)^2, \tag{4.36}$$

$$\varphi^z(\boldsymbol{\lambda}) = -\frac{((\boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a} - 1)^+)^2}{2d_0}. \tag{4.37}$$

The Hessian matrix of φ is a corresponding sum of the Hessian matrices of φ^x, φ^y and φ^z.

Lemma 4.2.5: The $m \times m$ Hessian matrix $\Phi(\boldsymbol{\lambda})$ of the dual function φ is given by

$$\Phi(\boldsymbol{\lambda}) = \Phi^y(\boldsymbol{\lambda}) + \Phi^z(\boldsymbol{\lambda}) + \sum_{j=1}^{n} \Phi_j^x(\boldsymbol{\lambda}), \tag{4.38}$$

where $\Phi^y(\boldsymbol{\lambda})$ is a diagonal matrix with diagonal elements

$$[\Phi^y(\boldsymbol{\lambda})]_{ii} = \begin{cases} 0 & \text{if} \quad \lambda_i < c_i, \\ \text{not defined} & \text{if} \quad \lambda_i = c_i, \\ -1 & \text{if} \quad \lambda_i > c_i, \end{cases} \tag{4.39}$$

the elements of the matrix $\Phi^z(\boldsymbol{\lambda})$ are given by

$$[\Phi^z(\boldsymbol{\lambda})]_{ik} = \begin{cases} 0 & \text{if} \quad \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a} < 1, \\ \text{not defined} & \text{if} \quad \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a} = 1, \\ -\dfrac{a_i a_k}{d_0} & \text{if} \quad \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{a} > 1, \end{cases} \tag{4.40}$$

and the elements of the matrix $\Phi_j^x(\boldsymbol{\lambda})$ are given by

$$\left[\Phi_j^x(\boldsymbol{\lambda})\right]_{ik} = \begin{cases} 0 & \text{if} \quad \xi_j(\boldsymbol{\lambda}) < \alpha_j \;\; \text{or} \;\; \xi_j(\boldsymbol{\lambda}) > \beta_j, \\ \text{typically not defined} & \text{if} \quad \xi_j(\boldsymbol{\lambda}) = \alpha_j \;\; \text{or} \;\; \xi_j(\boldsymbol{\lambda}) = \beta_j, \\ -h_j(\boldsymbol{\lambda})\, v_{ij}(\boldsymbol{\lambda})\, v_{kj}(\boldsymbol{\lambda}) & \text{if} \quad \alpha_j < \xi_j(\boldsymbol{\lambda}) < \beta_j, \end{cases} \tag{4.41}$$

where

$$v_{ij}(\boldsymbol{\lambda}) = \frac{p_{ij}}{p_j(\boldsymbol{\lambda})} - \frac{q_{ij}}{q_j(\boldsymbol{\lambda})} \quad \text{and} \quad h_j(\boldsymbol{\lambda}) = \frac{\sqrt{p_j(\boldsymbol{\lambda}) q_j(\boldsymbol{\lambda})}}{2(u_j - l_j)}. \tag{4.42}$$

Proof:
Straight-forward analytical calculations of the second order derivatives of $\varphi^y(\boldsymbol{\lambda})$ in (4.36) and $\varphi^z(\boldsymbol{\lambda})$ in (4.37) directly give the expressions for $\Phi^y(\boldsymbol{\lambda})$ in (4.39) and $\Phi^z(\boldsymbol{\lambda})$ in (4.40).

If $\alpha_j < \xi_j(\boldsymbol{\lambda}) < \beta_j$ then $\hat{x}_j(\boldsymbol{\lambda}) = \xi_j(\boldsymbol{\lambda})$ so that

$$\frac{\partial \varphi_j^x}{\partial \lambda_i}(\boldsymbol{\lambda}) = \tilde{f}_{ij}(\hat{x}_j(\boldsymbol{\lambda})) = \tilde{f}_{ij}(\xi_j(\boldsymbol{\lambda})) = \frac{p_{ij}}{u_j - \xi_j(\boldsymbol{\lambda})} + \frac{q_{ij}}{\xi_j(\boldsymbol{\lambda}) - l_j}. \tag{4.43}$$

Note that this holds not only at $\boldsymbol{\lambda}$ but also in a neighbourhood of $\boldsymbol{\lambda}$.
The expression (4.19) for $\xi_j(\boldsymbol{\lambda})$ now gives that

$$\frac{\partial \varphi_j^x}{\partial \lambda_i}(\boldsymbol{\lambda}) = \frac{\sqrt{p_j(\boldsymbol{\lambda})} + \sqrt{q_j(\boldsymbol{\lambda})}}{u_j - l_j} \left(\frac{p_{ij}}{\sqrt{p_j(\boldsymbol{\lambda})}} + \frac{q_{ij}}{\sqrt{q_j(\boldsymbol{\lambda})}} \right), \tag{4.44}$$

and then, by taking the derivative of this with respect to λ_k,

$$\frac{\partial^2 \varphi_j^x}{\partial \lambda_k \partial \lambda_i}(\boldsymbol{\lambda}) = -\frac{\sqrt{p_j(\boldsymbol{\lambda}) q_j(\boldsymbol{\lambda})}}{2(u_j - l_j)} \left(\frac{p_{ij}}{p_j(\boldsymbol{\lambda})} - \frac{q_{ij}}{q_j(\boldsymbol{\lambda})} \right) \left(\frac{p_{kj}}{p_j(\boldsymbol{\lambda})} - \frac{q_{kj}}{q_j(\boldsymbol{\lambda})} \right), \tag{4.45}$$

which proves the last line in (4.41).

If $\xi_j(\boldsymbol{\lambda}) < \alpha_j$ or $\xi_j(\boldsymbol{\lambda}) > \beta_j$ then $\hat{x}_j(\boldsymbol{\lambda}) = \alpha_j$ or β_j, and then

$$\frac{\partial \varphi_j^x}{\partial \lambda_i}(\boldsymbol{\lambda}) = \tilde{f}_{ij}(\hat{x}_j(\boldsymbol{\lambda})) = \frac{p_{ij}}{u_j - \alpha_j} + \frac{q_{ij}}{\alpha_j - l_j} \quad \text{or} \quad \frac{p_{ij}}{u_j - \beta_j} + \frac{q_{ij}}{\beta_j - l_j}. \tag{4.46}$$

Thus, the first order derivatives of φ_j^x are constant in a neighbourhood of $\boldsymbol{\lambda}$, which implies that all the second order derivates of φ_j^x are zero at $\boldsymbol{\lambda}$. This proves the first line in (4.41).

On the hyperplanes $\xi_j(\boldsymbol{\lambda}) = \alpha_j$ and $\xi_j(\boldsymbol{\lambda}) = \beta_j$, the second order derivatives of φ_j^x typically don't exist since they make a sudden jump when $\boldsymbol{\lambda}$ passes through any of these hyperplanes. (It is typically not the case that the right hand side in (4.45) is zero when $\xi_j(\boldsymbol{\lambda}) = \alpha_j$ or β_j.)

4.3 Some theoretical results for the dual problem

Lemma 4.3.1: The point $\hat{\boldsymbol{\lambda}}$ is an optimal solution to the dual problem (4.24) if and only if $\hat{\boldsymbol{\lambda}} \geq \mathbf{0}$, $\mathbf{g}(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) \leq \mathbf{0}$ and $\hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) = 0$.

Proof:
First, assume that $\hat{\boldsymbol{\lambda}} \geq \mathbf{0}$, $\mathbf{g}(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) \leq \mathbf{0}$ and $\hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) = 0$.
Then, for any $\boldsymbol{\lambda} \geq \mathbf{0}$, $(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}})^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) = \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) \leq 0$, and then the inequality (4.28) implies that $\varphi(\boldsymbol{\lambda}) - \varphi(\hat{\boldsymbol{\lambda}}) \leq 0$, so that $\hat{\boldsymbol{\lambda}}$ is an optimal solution to the dual problem (4.24).

In the rest of the proof, assume that $\hat{\boldsymbol{\lambda}}$ is an optimal solution to the dual problem (4.24). Then $\hat{\boldsymbol{\lambda}} \geq \mathbf{0}$ and $\varphi(\hat{\boldsymbol{\lambda}}) - \varphi(\boldsymbol{\lambda}) \geq 0$ for every $\boldsymbol{\lambda} \geq \mathbf{0}$, and then (4.29) implies that

$$(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda})^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}(\boldsymbol{\lambda})) \geq 0 \text{ for every } \boldsymbol{\lambda} \geq \mathbf{0}. \tag{4.47}$$

First, we show that $\mathbf{g}(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) \leq \mathbf{0}$. Assume that $g_k(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) > 0$ for some k.
Then let $\lambda_i = \hat{\lambda}_i$ for all $i \neq k$ and $\lambda_k = \hat{\lambda}_k + \delta_k$ where $\delta_k > 0$.
If δ_k is chosen sufficiently small then $g_k(\hat{\mathbf{w}}(\boldsymbol{\lambda})) > 0$.
But then $\boldsymbol{\lambda} \geq \mathbf{0}$ and $(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda})^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}(\boldsymbol{\lambda})) = -\delta_k g_k(\hat{\mathbf{w}}(\boldsymbol{\lambda})) < 0$, which contradicts (4.47).

Next, we show that $\hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) = 0$. Assume that $\hat{\lambda}_k g_k(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) \neq 0$ for some k.
Since $\hat{\lambda}_k \geq 0$ and $g_k(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) \leq 0$, it follows that $\hat{\lambda}_k > 0$ and $g_k(\hat{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) < 0$ for this k.
Then let $\lambda_i = \hat{\lambda}_i$ for all $i \neq k$ and $\lambda_k = \hat{\lambda}_k - \delta_k$ where $0 < \delta_k < \hat{\lambda}_k$.
If δ_k is chosen sufficiently small then $g_k(\hat{\mathbf{w}}(\boldsymbol{\lambda})) < 0$.
But then $\boldsymbol{\lambda} \geq \mathbf{0}$ and $(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda})^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}(\boldsymbol{\lambda})) = \delta_k g_k(\hat{\mathbf{w}}(\boldsymbol{\lambda})) < 0$, which contradicts (4.47).

The preparations are now made for proving Theorem 1 from section 4.1.
Before the proof, the theorem is reformulated with the simplified notations.

Theorem 1: If $\hat{\boldsymbol{\lambda}}$ is an optimal solution of the dual problem (4.24) then $\tilde{\mathbf{w}}(\hat{\boldsymbol{\lambda}})$ is the unique globally optimal solution of the primal subproblem (4.1). Moreover, $g_0(\tilde{\mathbf{w}}(\hat{\boldsymbol{\lambda}})) = \varphi(\hat{\boldsymbol{\lambda}})$, so that the optimal values of the primal problem and the dual problem are equal.

Proof:
Assume that $\hat{\boldsymbol{\lambda}}$ is an optimal solution to the dual problem and let $\hat{\mathbf{w}} = \tilde{\mathbf{w}}(\hat{\boldsymbol{\lambda}})$.
Then, from Lemma 4.3.1, $\mathbf{g}(\hat{\mathbf{w}}) \leq \mathbf{0}$ and $\hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}) = 0$.
This implies, in particular, that $\hat{\mathbf{w}}$ is a feasible solution to the primal problem.
Now, assume that $\mathbf{w}$ is another feasible solution to the primal problem, i.e. $\mathbf{w} \in W$, $\mathbf{g}(\mathbf{w}) \leq \mathbf{0}$ and $\mathbf{w} \neq \hat{\mathbf{w}}$. Then

$$g_0(\hat{\mathbf{w}}) = g_0(\hat{\mathbf{w}}) + \hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}) < g_0(\mathbf{w}) + \hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{g}(\mathbf{w}) \leq g_0(\mathbf{w}), \tag{4.48}$$

where the equality follows since $\hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{g}(\hat{\mathbf{w}}) = 0$, the strict inequality follows since $\hat{\mathbf{w}}$ is the unique minimizer of the Lagrange function, and the final inequality follows since $\mathbf{g}(\mathbf{w}) \leq \mathbf{0}$ and $\hat{\boldsymbol{\lambda}} \geq \mathbf{0}$. The equality in (4.48) also implies that $g_0(\hat{\mathbf{w}}) = L(\hat{\mathbf{w}}, \hat{\boldsymbol{\lambda}}) = L(\tilde{\mathbf{w}}(\hat{\boldsymbol{\lambda}}), \hat{\boldsymbol{\lambda}}) = \varphi(\hat{\boldsymbol{\lambda}})$.

4.4 ε_g–optimal dual solutions and sufficiently good primal solutions

We repeat that the dual problem reads

$$\text{D:}\;\; \text{maximize } \varphi(\boldsymbol{\lambda}) \;\; \text{subject to } \boldsymbol{\lambda} \geq \mathbf{0}. \tag{4.49}$$

Since the dual objective function φ is concave and continuously differentiable, a given point $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_m)^{\mathsf{T}} \in I\!\!R^m$ is an optimal solution to this dual problem if and only if

$$\begin{aligned}
&(1) \quad && \hat{\lambda}_i \geq 0 && \text{for all } i = 1, \ldots, m, \\
&(2) && \frac{\partial \varphi}{\partial \lambda_i}(\hat{\boldsymbol{\lambda}}) \leq 0 && \text{for all } i \in I_0(\hat{\boldsymbol{\lambda}}), \\
&(3) && \frac{\partial \varphi}{\partial \lambda_i}(\hat{\boldsymbol{\lambda}}) = 0 && \text{for all } i \in I_1(\hat{\boldsymbol{\lambda}}),
\end{aligned} \tag{4.50}$$

where $I_0(\hat{\boldsymbol{\lambda}}) = \{\, i \mid \hat{\lambda}_i = 0 \,\}$ and $I_1(\hat{\boldsymbol{\lambda}}) = \{\, i \mid \hat{\lambda}_i > 0 \,\}$.

Then the unique optimal solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{z})$ to the primal subproblem (4.1) is obtained from the formulas (4.20) and (4.15), with $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$.

In the fortran implementation, the optimality conditions (2) and (3) are relaxed slightly.

Def: A dual vector $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m)^{\mathsf{T}}$ is said to be a "ε_g–optimal solution" to D if

$$\begin{aligned}
&(1) \quad && \tilde{\lambda}_i \geq 0 && \text{for all } i = 1, \ldots, m, \\
&(2) && \frac{\partial \varphi}{\partial \lambda_i}(\tilde{\boldsymbol{\lambda}}) \leq \varepsilon_g && \text{for all } i \in I_0(\tilde{\boldsymbol{\lambda}}), \\
&(3) && \left| \frac{\partial \varphi}{\partial \lambda_i}(\tilde{\boldsymbol{\lambda}}) \right| \leq \varepsilon_g && \text{for all } i \in I_1(\tilde{\boldsymbol{\lambda}}),
\end{aligned} \tag{4.51}$$

The default value of the tolerance parameter ε_g, called `GEPS` in the fortran code, is 10^{-5}. The corresponding primal solution $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{z})$, obtained from (4.20) and (4.15) with $\boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}$, is then considered to be a "sufficiently good" solution to the primal subproblem (4.1).

In order to analyze *how* good a "sufficiently good" primal solution is, the following problem is considered, where the parameter δ is a (typically small) real number, positive or negative or zero.

$$\begin{aligned}
\text{P}(\delta): \quad & \text{minimize} && g_0(\mathbf{w}) \\
& \text{subject to} && g_i(\mathbf{w}) \leq \delta\,, \quad i = 1, \ldots, m \\
& && \mathbf{w} \in W.
\end{aligned} \tag{4.52}$$

Note that if $\delta = 0$ then the primal subproblem (4.1) is obtained.

The Lagrange function corresponding to $\text{P}(\delta)$ is

$$L_\delta(\mathbf{w}, \boldsymbol{\lambda}) = g_0(\mathbf{w}) + \sum_i \lambda_i (g_i(\mathbf{w}) - \delta) = L_0(\mathbf{w}, \boldsymbol{\lambda}) - \delta \sum_i \lambda_i, \tag{4.53}$$

where $L_0(\mathbf{w}, \boldsymbol{\lambda})$ is the Lagrange function in (4.3).

The corresponding dual objective function is

$$\varphi_\delta(\boldsymbol{\lambda}) = \min_{\mathbf{w}\in W} L_\delta(\mathbf{w},\boldsymbol{\lambda}) = \min_{\mathbf{w}\in W} L_0(\mathbf{w},\boldsymbol{\lambda}) - \delta\sum_i \lambda_i = L_0(\hat{\mathbf{w}}(\boldsymbol{\lambda}),\boldsymbol{\lambda}) - \delta\sum_i \lambda_i, \tag{4.54}$$

where $\hat{\mathbf{w}}(\boldsymbol{\lambda})$ is the unique $\mathbf{w} \in W$ which minimizes $L_0(\mathbf{w},\boldsymbol{\lambda})$ on W (for given $\boldsymbol{\lambda} \geq \mathbf{0}$).

It may be noted that

$$\varphi_\delta(\boldsymbol{\lambda}) = \varphi_0(\boldsymbol{\lambda}) - \delta\sum_i \lambda_i, \tag{4.55}$$

where $\varphi_0(\boldsymbol{\lambda})$ is the dual function in (4.23).

The dual problem D(δ) corresponding to the primal problem P(δ) is

$$\text{D}(\delta):\quad \text{maximize } \ \varphi_\delta(\boldsymbol{\lambda}) \ \text{ subject to } \boldsymbol{\lambda} \geq \mathbf{0}, \tag{4.56}$$

which equivalently may be written

$$\text{D}(\delta):\quad \text{maximize } \ \varphi_0(\boldsymbol{\lambda}) - \delta\sum_i \lambda_i \ \text{ subject to } \boldsymbol{\lambda} \geq \mathbf{0}. \tag{4.57}$$

Lemma 4.4.1: Assume that $\tilde{\boldsymbol{\lambda}}$ is an ε_g-optimal solution to D(δ), and let $\tilde{\mathbf{w}} = \hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}})$. Then $\tilde{\mathbf{w}}$ is a feasible solution to P($\delta{+}\varepsilon_g$), and $g_0(\tilde{\mathbf{w}}) \leq g_0(\mathbf{w})$ for every feasible solution $\mathbf{w}$ to P($\delta{-}\varepsilon_g$).

Proof: First, it follows from (4.31) that $\dfrac{\partial\varphi_\delta}{\partial\lambda_i}(\tilde{\boldsymbol{\lambda}}) = g_i(\hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}}))-\delta = g_i(\tilde{\mathbf{w}})-\delta$.
Therefore, it follows from the definition (4.51) that

$$\begin{aligned} &(2) && g_i(\tilde{\mathbf{w}}) \leq \delta{+}\varepsilon_g \quad \text{if } \ \tilde{\lambda}_i = 0,\\ &(3) && \delta{-}\varepsilon_g \leq g_i(\tilde{\mathbf{w}}) \leq \delta{+}\varepsilon_g \quad \text{if } \ \tilde{\lambda}_i > 0. \end{aligned} \tag{4.58}$$

In particular, this implies that $\tilde{\mathbf{w}}$ is a feasible solution to P($\delta{+}\varepsilon_g$).

Next, assume that $\mathbf{w}$ is a feasible solution to P($\delta{-}\varepsilon_g$). Then

$$\begin{aligned} g_0(\tilde{\mathbf{w}}) &= L_0(\tilde{\mathbf{w}},\tilde{\boldsymbol{\lambda}}) - \sum_{\tilde{\lambda}_i>0}\tilde{\lambda}_i g_i(\tilde{\mathbf{w}})\\ &\leq L_0(\mathbf{w},\tilde{\boldsymbol{\lambda}}) - \sum_{\tilde{\lambda}_i>0}\tilde{\lambda}_i g_i(\tilde{\mathbf{w}})\\ &= g_0(\mathbf{w}) + \sum_{\tilde{\lambda}_i>0}\tilde{\lambda}_i g_i(\mathbf{w}) - \sum_{\tilde{\lambda}_i>0}\tilde{\lambda}_i g_i(\tilde{\mathbf{w}})\\ &\leq g_0(\mathbf{w}) + \sum_{\tilde{\lambda}_i>0}\tilde{\lambda}_i(\delta{-}\varepsilon_g) - \sum_{\tilde{\lambda}_i>0}\tilde{\lambda}_i(\delta{-}\varepsilon_g)\\ &= g_0(\mathbf{w}), \end{aligned} \tag{4.59}$$

and the proof is complete.

Three special cases may be pointed out: $\delta = -\varepsilon_g,\ \delta = \varepsilon_g$ and $\delta = 0$.

The corresponding statements read:

Corollary 4.4.1:

Assume that $\tilde{\boldsymbol{\lambda}}$ is an ε_g-optimal solution to $\mathrm{D}(-\varepsilon_g)$, and let $\tilde{\mathbf{w}} = \hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}})$.
Then $\tilde{\mathbf{w}}$ is a feasible solution to P(0) and $g_0(\tilde{\mathbf{w}}) \leq$ the optimal value of $\mathrm{P}(-2\varepsilon_g)$.

Assume that $\tilde{\boldsymbol{\lambda}}$ is an ε_g-optimal solution to $\mathrm{D}(\varepsilon_g)$, and let $\tilde{\mathbf{w}} = \hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}})$.
Then $\tilde{\mathbf{w}}$ is a feasible solution to $\mathrm{P}(2\varepsilon_g)$ and $g_0(\tilde{\mathbf{w}}) \leq$ the optimal value of P(0).

Assume that $\tilde{\boldsymbol{\lambda}}$ is an ε_g-optimal solution to D(0), and let $\tilde{\mathbf{w}} = \hat{\mathbf{w}}(\tilde{\boldsymbol{\lambda}})$.
Then $\tilde{\mathbf{w}}$ is a feasible solution to $\mathrm{P}(\varepsilon_g)$ and $g_0(\tilde{\mathbf{w}}) \leq$ the optimal value of $\mathrm{P}(-\varepsilon_g)$.

4.5 An active set method for the dual problem

Consider again the dual problem

$$\mathrm{D}: \quad \text{maximize } \varphi(\boldsymbol{\lambda}) \text{ subject to } \lambda_i \geq 0 \text{ for } i = 1, \ldots, m. \tag{4.60}$$

In this section, the gradient vector of the dual objective function is denoted $\mathbf{g}(\boldsymbol{\lambda})$, i.e.

$$\mathbf{g}(\boldsymbol{\lambda}) = (g_1(\boldsymbol{\lambda}), \ldots, g_m(\boldsymbol{\lambda}))^{\mathsf{T}} = \left(\frac{\partial \varphi}{\partial \lambda_1}(\boldsymbol{\lambda}), \ldots, \frac{\partial \varphi}{\partial \lambda_m}(\boldsymbol{\lambda}) \right)^{\mathsf{T}}. \tag{4.61}$$

Let $\mathcal{M} = \{1, \ldots, m\}$ and let $\mathcal{A}$ be a given subset of $\mathcal{M}$, i.e. $\mathcal{A} \subseteq \mathcal{M}$.

Def: The *equality-constrained subproblem* $\mathrm{D}_{\mathcal{A}}$ corresponding to $\mathcal{A}$ is defined as

$$\mathrm{D}_{\mathcal{A}}: \quad \text{maximize } \varphi(\boldsymbol{\lambda}) \text{ subject to } \lambda_i = 0 \text{ for all } i \in \mathcal{A}. \tag{4.62}$$

Lemma 4.5.1: Assume that the point $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_m)^{\mathsf{T}}$ and the index set $\mathcal{A}$ satisfy that

$$\begin{array}{lrl} (1) & \hat{\lambda}_i = 0 & \text{for all } i \in \mathcal{A}, \\ (2) & g_i(\hat{\boldsymbol{\lambda}}) = 0 & \text{for all } i \in \mathcal{M} \setminus \mathcal{A}, \\ (3) & \hat{\lambda}_i \geq 0 & \text{for all } i \in \mathcal{M} \setminus \mathcal{A}, \\ (4) & g_i(\hat{\boldsymbol{\lambda}}) \leq 0 & \text{for all } i \in \mathcal{A}. \end{array} \tag{4.63}$$

Then $\hat{\boldsymbol{\lambda}}$ is an optimal solution to the problem D.
Conversely, If $\hat{\boldsymbol{\lambda}}$ is an optimal solution to the problem D, then there is an index set $\mathcal{A} \subseteq \mathcal{M}$ such that (1)–(4) are satisfied (namely every index set $\mathcal{A}$ such that $\{\, i \in \mathcal{M} \mid g_i(\hat{\boldsymbol{\lambda}}) < 0\} \subseteq \mathcal{A} \subseteq \{\, i \in \mathcal{M} \mid \hat{\lambda}_i = 0\}$).

Lemma 4.5.2: The point $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^{\mathsf{T}}$ is an optimal solution to the equality-constrained subproblem $\mathrm{D}_{\mathcal{A}}$ if and only if

$$\begin{array}{lrl} (1) & \lambda_i = 0 & \text{for all } i \in \mathcal{A}, \\ (2) & g_i(\boldsymbol{\lambda}) = 0 & \text{for all } i \in \mathcal{M} \setminus \mathcal{A}. \end{array} \tag{4.64}$$

Lemma 4.5.3: Assume that the point $\hat{\boldsymbol{\lambda}}$ and the index set $\mathcal{A}$ satisfy (1)–(4) in Lemma 4.5.1. Assume further that the same index set $\mathcal{A}$ and the possibly new point $\boldsymbol{\lambda}$ satisfy (1)–(3) in Lemma 4.5.1, i.e.

$$\begin{array}{lrl} (1) & \lambda_i = 0 & \text{for all } i \in \mathcal{A}, \\ (2) & g_i(\boldsymbol{\lambda}) = 0 & \text{for all } i \in \mathcal{M} \setminus \mathcal{A}, \\ (3) & \lambda_i \geq 0 & \text{for all } i \in \mathcal{M} \setminus \mathcal{A}. \end{array} \tag{4.65}$$

Then $\boldsymbol{\lambda}$ and $\mathcal{A}$ also satisfy (4) in Lemma 4.5.1, i.e.

$$(4) \quad g_i(\boldsymbol{\lambda}) \leq 0 \quad \text{for all } i \in \mathcal{A}, \tag{4.66}$$

and $\boldsymbol{\lambda}$ is an optimal solution both to $\mathrm{D}_{\mathcal{A}}$ and to D.

Thus, an optimal solution to the inequality-constrained dual problem D might be obtained by solving the (typically much easier) equality-constrained subproblem $\mathrm{D}_{\mathcal{A}}$, while ensuring that $\boldsymbol{\lambda} \geq \mathbf{0}$ is satisfied. The difficulty, of course, is that the "optimal" index set $\mathcal{A}$ is not known beforehand.

The main idea with an "active set method" is to make a good guess of the "optimal" index set $\mathcal{A}$, a guess which is iteratively updated during the solution process.

When "optimal solution" is replaced by "ε_g-optimal solution", the corresponding result is as follows, where (4.67) is essentially a repetition of (4.51).

Def: A point $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m)^{\mathsf{T}}$ is an ε_g-optimal solution to the problem D if there is an index set $\mathcal{A} \subseteq \mathcal{M}$ such that the following conditions hold:

$$\begin{array}{rl} \tilde{\lambda}_i = 0 & \text{for all } i \in \mathcal{A}, \\ |g_i(\tilde{\boldsymbol{\lambda}})| \leq \varepsilon_g & \text{for all } i \in \mathcal{M} \setminus \mathcal{A}. \\ \tilde{\lambda}_i \geq 0 & \text{for all } i \in \mathcal{M} \setminus \mathcal{A}. \\ g_i(\tilde{\boldsymbol{\lambda}}) \leq \varepsilon_g & \text{for all } i \in \mathcal{A}. \end{array} \tag{4.67}$$

Def: A point $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^{\mathsf{T}}$ is an ε_g-optimal solution to the problem $\mathrm{D}_{\mathcal{A}}$ if the following conditions hold:

$$\begin{array}{rl} \lambda_i = 0 & \text{for all } i \in \mathcal{A}, \\ |g_i(\boldsymbol{\lambda})| \leq \varepsilon_g & \text{for all } i \in \mathcal{M} \setminus \mathcal{A}. \end{array} \tag{4.68}$$

Lemma 4.5.4: Assume that $\tilde{\boldsymbol{\lambda}}$ is an ε_g-optimal solution to D, and $\mathcal{A} = \{\, i \in \mathcal{M} \mid \tilde{\lambda}_i = 0 \,\}$. Then $\tilde{\boldsymbol{\lambda}}$ is an ε_g-optimal solution also to $\mathrm{D}_{\mathcal{A}}$.
Further, each ε_g-optimal solution $\boldsymbol{\lambda}$ to $\mathrm{D}_{\mathcal{A}}$ which satisfies $\boldsymbol{\lambda} \geq \mathbf{0}$ and $g_i(\boldsymbol{\lambda}) \leq \varepsilon_g$ for all $i \in \mathcal{A}$ is an ε_g-optimal solution to D.

4.6 Active set algorithm for finding an ε_g−optimal solution

The active set algorithm, used in the fortran implementation for finding an ε_g−optimal solution to the dual problem D, is now described.
At iteration k, the current index set guess is $\mathcal{A}^{(k)}$ and the current iteration point is $\boldsymbol{\lambda}^{(k)}$.
It might be noted that if $\boldsymbol{\lambda}^{(k+1)} \neq \boldsymbol{\lambda}^{(k)}$ then $\varphi(\boldsymbol{\lambda}^{(k+1)}) > \varphi(\boldsymbol{\lambda}^{(k)})$.

STEP 0: The index counter is set to $k = 1$, the first iteration point is set to $\boldsymbol{\lambda}^{(1)} = \mathbf{0}$, and the first index guess is set to $\mathcal{A}^{(1)} = \{\ i \in \mathcal{M} \mid g_i(\boldsymbol{\lambda}^{(1)}) \leq \varepsilon_g\ \}$.
If $\mathcal{A}^{(1)} = \mathcal{M}$, then $\boldsymbol{\lambda}^{(1)} = \mathbf{0}$ is an ε_g−optimal solution to the dual problem D and the algorithm is stopped! Otherwise, the algorithm continues with STEP 1.

STEP 1: Here, the current index set $\mathcal{A}^{(k)} \subset \mathcal{M}$ is given, together with the current iteration point $\boldsymbol{\lambda}^{(k)} \in I\!R^m$ which satisfies $\lambda_i^{(k)} = 0$ for $i \in \mathcal{A}^{(k)}$ and $\lambda_i^{(k)} \geq 0$ for $i \in \mathcal{M} \setminus \mathcal{A}^{(k)}$.
If $|\, g_i(\boldsymbol{\lambda}^{(k)})| \leq \varepsilon_g$ for all $i \in \mathcal{M} \setminus \mathcal{A}^{(k)}$, the algorithm continues with STEP 2.
Otherwise, the algorithm continues with STEP 3.

STEP 2: Here, the current iteration point $\boldsymbol{\lambda}^{(k)}$ is an ε_g−optimal solution to $\mathrm{D}_{\mathcal{A}^{(k)}}$.
If $g_i(\boldsymbol{\lambda}^{(k)}) \leq \varepsilon_g$ for all $i \in \mathcal{A}^{(k)}$ then $\boldsymbol{\lambda}^{(k)}$ is an ε_g−optimal solution to D and the algorithm is stopped!
Otherwise, let $p \in \mathcal{A}^{(k)}$ be an index such that $g_p(\boldsymbol{\lambda}^{(k)}) = \max_i\{\, g_i(\boldsymbol{\lambda}^{(k)}) \mid i \in \mathcal{A}^{(k)}\}\ (> \varepsilon_g)$.
Then the index set, the iteration point, and the iteration counter are updated as follows:
$\mathcal{A}^{(k+1)} = \mathcal{A}^{(k)} \setminus \{p\}, \quad \boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)}, \quad k \leftarrow k+1,$
whereafter the algorithm continues with STEP 1.

STEP 3: Here, the current iteration point $\boldsymbol{\lambda}^{(k)}$ is *not* an ε_g−optimal solution to $\mathrm{D}_{\mathcal{A}^{(k)}}$.
Then a search direction $\mathbf{d}^{(k)} \in I\!R^m$ such that $d_i^{(k)} = 0$ for $i \in \mathcal{A}^{(k)}$ and $\mathbf{g}(\boldsymbol{\lambda}^{(k)})^{\mathsf{T}}\mathbf{d}^{(k)} > 0$ is calculated. If $\mathbf{d}^{(k)} \geq \mathbf{0}$ the algorithm continues with STEP 4.
Otherwise, a maximal steplength is calculated through $t_{\max} = \min\limits_i \left\{ \dfrac{\lambda_i^{(k)}}{-d_i^{(k)}} \mid d_i^{(k)} < 0 \right\}$.
If $\mathbf{g}(\boldsymbol{\lambda}^{(k)} + t_{\max}\mathbf{d}^{(k)})^{\mathsf{T}}\mathbf{d}^{(k)} < 0$ the algorithm continues with STEP 4.
Otherwise, let $q \in \mathcal{M} \setminus \mathcal{A}^{(k)}$ be an index such that $\lambda_q^{(k)} + t_{\max} d_q^{(k)} = 0$.
Then the index set, the iteration point, and the iteration counter are updated as follows:
$\mathcal{A}^{(k+1)} = \mathcal{A}^{(k)} \cup \{q\}, \quad \boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} + t_{\max}\mathbf{d}^{(k)}, \quad k \leftarrow k+1,$
whereafter the algorithm continues with STEP 1.

STEP 4: Here, a line search from the point $\boldsymbol{\lambda}^{(k)}$ in the search direction $\mathbf{d}^{(k)}$ is made.
A steplength t_k such that $\mathbf{g}(\boldsymbol{\lambda}^{(k)} + t_k\mathbf{d}^{(k)})^{\mathsf{T}}\mathbf{d}^{(k)} > 0 \geq \mathbf{g}(\boldsymbol{\lambda}^{(k)} + (1+\varepsilon)t_k\mathbf{d}^{(k)})^{\mathsf{T}}\mathbf{d}^{(k)}$
is calculated. (A default value of the parameter ε is 0.03.)
Then the index set, the iteration point, and the iteration counter are updated as follows:
$\mathcal{A}^{(k+1)} = \mathcal{A}^{(k)}, \quad \boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} + t_k\mathbf{d}^{(k)}, \quad k \leftarrow k+1,$
whereafter the algorithm continues with STEP 1.

5 A small test problem

Consider the following problem in the variables $\mathbf{x} = (x_1, x_2, x_3)^\mathsf{T}$:

$$\begin{aligned} \text{minimize} \quad & x_1^2 + x_2^2 + x_3^2 \\ \text{subject to} \quad & (x_1 - 5)^2 + (x_2 - 2)^2 + (x_3 - 1)^2 \leq 9, \\ & (x_1 - 3)^2 + (x_2 - 4)^2 + (x_3 - 3)^2 \leq 9, \\ & 0 \leq x_j \leq 5, \quad j = 1, 2, 3. \end{aligned} \tag{5.1}$$

With $c_1 = c_2 = 1000$, $a_1 = a_2 = 0$, the starting point $(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}) = (4, 3, 2)$, and the parameter GEPS $= 10^{-7}$, it turned out that $y_1^{(k)} = y_2^{(k)} = z^{(k)} = 0$ for all $k \geq 2$, while $\mathbf{x}^{(k)}$, $f_0(\mathbf{x}^{(k)})$, $f_1(\mathbf{x}^{(k)})$ and $f_2(\mathbf{x}^{(k)})$, for $k = 1, \ldots, 7$, are shown in the following two tables:

MMA:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$f_0(\mathbf{x}^{(k)})$	$f_1(\mathbf{x}^{(k)})$	$f_2(\mathbf{x}^{(k)})$
1	4.000000	3.000000	2.000000	29.000000	3.000000	3.000000
2	2.390298	1.805719	0.992865	9.959929	6.848340	9.215195
3	2.038452	1.762359	1.241707	8.803031	8.885662	9.023207
4	2.017793	1.778557	1.239183	8.770329	8.999802	9.000017
5	2.017626	1.779369	1.238257	8.770249	9.000001	8.999998
6	2.017554	1.779796	1.237758	8.770246	9.000000	9.000000
7	2.017526	1.779968	1.237558	8.770246	9.000000	9.000000

GCMMA:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$f_0(\mathbf{x}^{(k)})$	$f_1(\mathbf{x}^{(k)})$	$f_2(\mathbf{x}^{(k)})$
1	4.000000	3.000000	2.000000	29.000000	3.000000	3.000000
2	2.555037	1.890622	1.076547	11.261620	5.995666	8.347138
3	2.072173	1.795876	1.191027	8.937619	8.650326	8.991408
4	2.016184	1.791365	1.224353	8.773025	8.997020	8.998887
5	2.016950	1.783479	1.233496	8.770396	8.999988	8.999891
6	2.017408	1.780681	1.236728	8.770255	8.999998	8.999992
7	2.017508	1.780073	1.237436	8.770246	9.000000	9.000000

Note that, on this particular problem, MMA is slightly faster than GCMMA. Also note that some of the iteration points generated by MMA are slightly infeasible, while all the iteration points generated by GCMMA are feasible (within the tolerances given by GEPS).

6 References

[1] K. Svanberg, The method of moving asymptotes – a new method for structural optimization, *International Journal for Numerical Methods in Engineering*, 1987, 24, 359-373.

[2] K. Svanberg, A class of globally convergent optimization methods based on conservative convex separable approximations, *SIAM Journal of Optimization*, 2002, 12, 555-573.