Fluid Mechanics Research Laboratory
Department of Mechanical Engineering
FAMU/FSU College of Engineering
Florida State University

Least-Maximum Solution Of Underdetermined Linear Systems

Leon Van Dommelen

Report FMRL TR–4 **August 1989**

Least-Maximum Solution Of Underdetermined Linear Systems

Leon Van Dommelen

Department of Mechanical Engineering
FAMU/FSU College of Engineering
The Florida State University
P.O. Box 2175
Tallahassee, Florida 32316-2175

Report FMRL TR–4 August 1989

Abstract

This paper proposes an algorithm to find the solution of an underdetermined linear system of equation with the smallest maximum coefficient. The algorithm provides a generalized triangularization of the matrix.

1. Introduction.

This note concerns itself with an algorithm to find the least maximum solution to an underdetermined linear system of equations,

$$A\mathbf{x} = \mathbf{b} \tag{1.1}$$

where A is an m by n matrix with $m \leq n$. The least–maximum solution $\mathbf{x} = \mathbf{e}$ is the solution with the smallest maximum norm,

$$A\mathbf{e} = \mathbf{b} \qquad \epsilon = ||\mathbf{e}||_\infty \tag{1.2a,b}$$

where

$$||\mathbf{x}||_\infty \geq \epsilon \quad \text{when} \quad A\mathbf{x} = \mathbf{b} \tag{1.3}$$

A problem of this type arises in the generation of discretization formulae for random distributions of points in a mesh-free environment. Constraints of accuracy lead to a system of linear equations for the discretization, while stability constraints put a limit on the allowed magnitude of the unknowns. Since an analytical solution is not feasible in the mesh-free environment, the procedure is to add unknowns until the least–maximum solution is within the prescribed bounds.

The least–maximum solution for a single equation, $m = 1$, is simple. By substitution it can be verified that the following is a solution:

$$e_i = \text{sign}(b_m)\text{sign}(a_{mi})\epsilon_m \quad (a_{mi} \neq 0) \tag{1.4a}$$

$$|e_i| \leq \epsilon_m \qquad (a_{mi} = 0) \tag{1.4b}$$

if

$$\epsilon_m = \frac{|b_m|}{a_m} \tag{1.5a}$$

where a_m is the L_1–norm of the single row of matrix A,

$$a_m = ||\mathbf{a}_m^T||_1 = \sum_{j=1}^{n} |a_{mj}| \tag{1.5b}$$

According to Holder's inequality,

$$||\mathbf{a}_m^T||_1 ||\mathbf{x}||_\infty \geq |b_m| \tag{1.6}$$

for any solution $\mathbf{x}$, so that (1.4), (1.5) is the desired least–maximum solution with $\epsilon_m = \epsilon$.

When there is more than one equation, the solution is considerably more complicated. The correct maximum solution $\mathbf{e}$ must still satisfy each equation, including the m-th. Thus the Holder inequality (1.6) shows that its maximum norm ϵ is at least equal to ϵ_m. But, (1.4) is not a solution to the full system;

The present procedure iteratively increases the value of ϵ_m to approach ϵ from below, by forming new linear combinations of equations. Eventually, this results in the correct

value of the maximum norm of the least maximum solution. In addition, at convergence the m-solution gives the correct least–maximum solution for the unknowns determined in (1.4a), while similar least–maximum solutions to the other equations give the unknowns not determined by (1.4a).

2. Increase of the lower bound.

Our purpose in this section is to attempt to increase the lower bound to the least–maximum ϵ by combination of any two existing equations. Two special cases will be excluded for now:

(a) If the L_1–norm a_i of any equation vanishes while the right hand side b_i is non-zero, the system is unsolvable.

(b) If all the b_i are zero, the system is homogeneous and the appropriate least–maximum solution vanishes.

If neither of these two conditions applies, we can select the m–th equation as the one that gives the most stringent bound on ϵ; the equation with the largest value of

$$\epsilon_m = \frac{|b_m|}{a_m} \qquad (a_m, b_m \neq 0) \tag{2.1}$$

For effective vectorization of various numerical operations, the unknowns will be reordered to move the non-zero coefficients a_{mj} of the last equation to the right.

To further increase this lower bound, we will attempt to form a linear combination of equation m and an amount λ of another equation i. The bound of the combination is

$$\epsilon'_m = \frac{|b_m + \lambda b_i|}{\sum_{j=1}^{J-1} |\lambda a_{ij}| + \sum_{j=J}^{n} |a_{mj} + \lambda a_{ij}|} \tag{2.2}$$

where J is the index of the first non-zero coefficient in equation m. Since (2.2) is piecewise monotonous, the maximum occurs at a vertex where one of the coefficients $a_{mj} + \lambda a_{ij}$ vanishes.

In selecting equation i, we choose the equation that leads to the largest initial increase of ϵ_m with λ. Expanding (2.2) for small λ,

$$d\epsilon_m \sim \frac{\mathrm{sign}(b_m)\lambda b_i a_m - \lambda \sum_{j=J}^{n} a_{ij} sign(a_{mj})|b_m| - |\lambda||b_m| \sum_{j=1}^{J-1} |a_{ij}|}{a_m^2} \tag{2.3}$$

This expression can be reduced by noting that the first sum involves the deterministic part (1.4) of the least-maximum solution e_i of the m-equation alone. We will define a 'reduced' equation i^* by taking the unknowns $j \geq J$ equal to this deterministic part and moving them to the right hand side:

$$b_i^* = b_i - \sum_{j=J}^{n} a_{ij} e_j \tag{2.4}$$

$$a_i^* = \sum_{j=1}^{J-1} |a_{ij}| \tag{2.5}$$

This is so far merely a matter of definition; there is no assurance that the reduced equation is solvable even when the original system is. Yet, it will be seen that at convergence, the reduced system is solvable, and its least–maximum solution will be smaller than the least–maximum solution of the m-equation.

In terms of the reduced quantities, the initial increase in ϵ_m is

$$d\epsilon_m \sim \frac{|\lambda|}{a_m^2} \left(\text{sign}(b_m) sign(\lambda) a_m b_i^* - a_i^* |b_m|\right) \tag{2.6}$$

Since the sign of λ can be arbitrary, to get the largest initial increase in ϵ_m we choose the equation i as the one with the largest value of

$$d_i = a_m |b_i^*| - a_i^* |b_m| \tag{2.7}$$

Since d_i is positive, the linear combination will increase ϵ_m to approach the true least–maximum more closely.

But, when all the d_i are non-positive, we must turn to the reduced system, temporarily ignoring the m-th equation. Because of (2.7), each equation of the reduced system is now at least initially individually solvable. If the subsystem happens to be homogeneous, we found the true least–maximum solution, since we can satisfy the equations by taking the remaining unknowns $j = 1, \ldots, J-1$ zero.

If the reduced system is not homogeneous, we can perform a similar iterative process on it. At least initially, the least–maximum solution of any equation of the reduced system is less than that of the m–th equation:

$$\epsilon_i^* = \frac{|b_i^*|}{a_i^*} \leq \epsilon_m \tag{2.8}$$

When we repeat the equation combination for the subsystem to increase the maximum value of ϵ_i^*, we can use the new subsystem to further increase ϵ_m when (2.8) is no longer true, including the case that unsolvability arises. In case (2.8) remains true at convergence, we repeat the iterative process on the subsystem of the subsystem.

At convergence, when none of the least–maximum solutions can be further increased, the system must assume the form

$$\begin{pmatrix} \mathbf{a}_{1\,1}^T & \mathbf{a}_{1\,2}^T & \cdots & \cdots & \mathbf{a}_{1\,m-1}^T & \mathbf{a}_{1\,m}^T \\ & \mathbf{a}_{2\,2}^T & \cdots & \cdots & \mathbf{a}_{2\,m-1}^T & \mathbf{a}_{2\,m}^T \\ & & \ddots & & \vdots & \vdots \\ & & & \ddots & \vdots & \vdots \\ & & & & \mathbf{a}_{m-1\,m-1}^T & \mathbf{a}_{m-1\,m}^T \\ & & & & & \mathbf{a}_{m\,m}^T \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \vdots \\ \mathbf{x}_{m-1} \\ \mathbf{x}_m \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_{m-1} \\ b_m \end{pmatrix} \tag{2.9}$$

where the vectors $\mathbf{x}_i$ may be of zero length and the vectors $\mathbf{a}_{i\,i}^T$ have all non-zero coefficients.

Backward substitution gives

$$B_i = b_i - \sum_{j=i+1}^{m} \mathbf{a}_{i\,j}^T \mathbf{x}_j \tag{2.10a}$$

$$\mathbf{a}_{i\,i}^T \mathbf{x}_i = B_i \tag{2.10b}$$

In case the vector $\mathbf{x}_i$ is of length zero, B_i vanishes, and if not, we solve (2.10) in least–maximum sense,

$$\epsilon_i = \frac{|B_i|}{||\mathbf{a}_{ii}^T||_1} \tag{2.10c}$$

$$(\mathbf{x}_i)_k = \text{sign}\left(\left(\mathbf{a}_{i\,i}^T\right)_k\right) \text{sign}(B_i)\epsilon_i$$

On behalf of (2.8) the ϵ_i are monotonously increasing, and since ϵ_m must still be less than or equal to the maximum norm of the least–maximum solution, our solution must be a least–maximum one. Thus ϵ_m gives the maximum norm of the least–maximum solution.

3. Subroutine INFSOL

INFSOL performs an iterative procedure of the general nature described in the previous section. INFSOL first attempts to triangularize the matrix using full pivoting. This allows linear dependence between equations to be recognized, and it increases the linear independence of the equations.

In each iteration, INFSOL makes only a single attempt to increase ϵ_m. INFSOL next continues with the subsystems.