CHAPTER 3

NUMERICAL METHODS AND OPTIMIZATION TECHNIQUES

CHAPTER - 3

NUMERICAL METHODS AND OPTIMIZATION TECHNIQUES

In chemical engineering practise data is often given for discrete values along a continium. However, a chemical engineer may often require estimates between the discrete values and in certain cases, beyond the ranges of measured variables. If a data fit has already been obtained in terms of a generalised correlation, then such a correlation could be used conveniently for interpolation or extrapolation of the data.

Parametric data having multivariable dependencies cannot be utilised for interpolation or extrapolation unless all but one of the variables, fall in the parametric ranges of the measured data. Therefore, in order to predict or forecast the values of such multivariable dependent parameters, an appropriate generalised correlation between the concerned parameter and the independent variables is necessary.

Generalised correlations in chemical engineering are often expressed in terms of dimensionless numbers instead of fundamental variables. The important advantage of obtaining correlations using dimensionless numbers is that it interweaves a physical picture with the parametric dependency in terms of force ratio's etc. and thereby elevates the mathematical curve fitting process into realms of physical interpretation. The other advantage is that by using

dimensionless numbers often the dimensionality of the problem gets substantially reduced due to the lumping of different variables. Thus, for example, to obtain a correlation for liquid side mass transfer coefficient (k_L) one has to use five fundamental variables such as liquid density, viscosity, velocity, characteristic length and diffusivity of the solute in the liquid ; however to obtain the same correlation for k_L one has to use only two dimensionless numbers, that is, the Reynold's and Schmidt number.

Numerical methods and optimization techniques which could be used conveniently for obtaining generalised correlations have been outlined in the following pages.

3.1.0 MULTIPLE LINEAR REGRESSION :

In general, any parameter (y) could be correlated to different variables (x_1, x_2,....., x_n) by the following power law equation:-

$$y = a_o\, x_1^{a1}\, x_2^{a2} \ldots\ldots\ldots x_m^{am} \qquad (3.1)$$

The above equation (3.1) can be transformed into a linear equation as under :-

$$\log y = \log a_o + a_1 \log x_1 + a_2 \log x_2 \ldots\ldots\ldots + a_m \log x_m \qquad (3.2)$$

By expressing all the lograthimic terms in terms of X, one can rewrite the equation (3.2) as under :-

$$Y = a_o' + a_1 X_1 + a_2 X_2 \ldots\ldots\ldots + a_m X_m \qquad (3.3)$$

Using multiple linear regression, the values of coefficients (a_o, a_1, ... a_m) can be obtained by the solution of the undermentioned matrix.

$$\begin{bmatrix} \sum(n) & \sum(X_{1i}) & \sum(X_{2i}) & \cdots\cdots & \sum(X_{mi}) \\ \sum(X_{1i}) & \sum(X_{1i})^2 & \sum(X_{1i}\,X_{2i}) & \cdots\cdots & \sum(X_{1i}\,X_{mi}) \\ \vdots & \vdots & \vdots & & \vdots \\ \sum(X_{mi}) & \sum(X_{mi}\,X_{1i}) & \sum(X_{mi}\,X_{2i}) & \cdots\cdots & \sum(X_{mi})^2 \end{bmatrix} \begin{Bmatrix} a_o \\ a_1 \\ \vdots \\ a_m \end{Bmatrix} = \begin{Bmatrix} \sum(Y_i) \\ \sum(X_{1i}\,Y_i) \\ \vdots \\ \sum(X_{mi}\,Y_i) \end{Bmatrix}$$

The relevant details about the formulation of set of simultaneous (normal) equations are reported in reference (122,123,124).

3.1.1 Solution techniques for simultaneous (normal) equations :

The normal equations formulated by multiple linear regression may be solved by any of the standard elimination methods viz. Gauss elimination, L.U. decomposition etc. The precision of solution obtained by elimination methods is likely to be affected by round off errors. Hence, Chapra and Canale (124) recommend the matrix inversion aproach such as, Gauss Jordan method for solution of the normal equations, wherein the standard computer software of Constantinides (125) could be conveniently used.

Another method of solution of simultaneous equations is to use a iterative method, such as Gauss Seidel which is very accurate and not prone to round off error as the earlier mentioned methods.

This method is most reliable for systems which are diagonally dominant. In other cases it may not always converge or sometimes converges very slowly on the true solution. Since the normal equations obtained by multiple linear regresssion may not be diagonally dominant, it is likely that Gauss Seidel may not converge to a solution in some cases.

3.2.0 UNCONSTRAINED OPTIMIZATION :

Multiple linear regression may be utilised to obtain the coefficient values of a nonlinear powerlaw type model : However there is a loss of sensititvity due to the lograthimic transformation. Moreover the standard methods of solution mentioned earlier yield exact mathematical solutions, which may not always be amenable to physical interpretation. For hypothesis testing it is more convenient to obtain a set of solutions within a prescribed range of error and then analyse the different solutions in the light of physical interpretation of the phenomena. Thus, unconstrained optimization could be adopted conveniently to yield the appropriate coefficient values.

Selection of algorithm largely depends on the objective of the task. If the objective of optimization is to obtain suitable coefficients which minimise the relative error the objective function could be defined as "Error = |(dev/exp)|". In view of the nondifferentiable nature of the function, only direct search methods should be utilised.

An indepth review of the numerical direct search methods for unconstrained optimization has been provided by Swann (126). It appears that most of the numerical direct search unconstrained optimization methods are essentially based on the philosophy of the alternating variable method wherein a set of directions is defined which can be used to explore the parametric space. Thus, for the case of multivariable optimization problem of minimization of certain objective function, it consists of minimizing with respect to each independent variable in turn. Starting with initial approximation, the variable x_1 is altered, with variables x_2, x_3..... x_n held constant, until a minimum of the objective function is located where upon x_1 is fixed and x_2 explored in the same way and so on until x_n has been explored. In practice, the alternating variable method is usually inefficient and the progress is characterised by oscillatory behaviour. Many investigators (127 to 129) have proposed modifications to the alternating variable method based on the observation that the begining and end of a directional search cycle determine a line along which more substantial progress may be made. Fletcher (130,131) has indicated that the most efficient of such modified methods, is the method proposed by Davis, Swann and Compey (129), known as "DSC Method".

3.2.1 DSC algorithm :

In the DSC algorithm, minima in a direction is located by fitting a quadratic to three points in the neighbourhood of the minima, differentiating the resulting equation and equating it to

zero gives a reasonable estimate of the minima. This algorithm moves in the descent direction using a step size always twice the previous, until a function value f_4 is found which exceeds the previous value f_3. At this stage a step is made in the reverse direction to obtain a point equidistant between the earlier two steps (having function value f_2) thereby implying that four equidistant points are obtained of which three are selected to define a quadratic. The point which is farthest from the point having the lowest of the four function values is rejected. If the function values corresponding to the equispaced points x_1, x_2 and x_3 are f_1, f_2, and f_3 where f_2 is the current lowest value corresponding to x_2 then the quadratic through points, x_1, x_2 and x_3 has a minimum at $(x_2 + \Delta\ m)$. Where Δ is the initial step size and m is given by the following equation (3.4).

$$\Delta\ m = \frac{\Delta\ (f_1 - f_3)}{2\ (f_1 - 2f_2 + f_3)} \qquad (3.4)$$

Having obtained the estimate of minimum location, the direction of the search may be altered towards other variables.

3.2.2 Powell's algorithm :

The most successful of direct search methods is the algorithm due to Powell (132). This algorithm effectively uses the history of iterations to build up directions for acceleration and at same time avoids degenerating to a sequence of coordinate searches this algorithm however requires the minimum to be bracketed. (i.e. minima range to be prespecifed).

The theoretical features of this algorithm are discussed by Reklaitis et al. (133). The most salient feature is that a quadratic approximation is carried out using the first three points obtained in the direction of search, and these quadratic approximations are continued until the minimum f(x) is located to the required precision. Therefore, the minimum position x_m from function values f_1, f_2, f_3 at points x_1, x_2, x_3 is given by the undermentioned equation (3.5).

$$x_m = \frac{1}{2}\left[\frac{(x_2^2 - x_3^2)f_1 + (x_3^2 - x_1^2)f_2 + (x_1^2 - x_2^2)f_3}{(x_2 - x_3)f_1 + (x_3 - x_1)f_2 + (x_1 - x_2)f_3}\right] \qquad (3.5)$$

3.2.3 Combination algorithm : D S C - Powell algorithm :

The DSC search described earlier does not require the optimum to be bracketed (i.e. range to be prespecified), however it moves very rapidly (as Δ increases) to bracket the optimum. Powell's method which is very efficient requires the minimum to be bracketed. Therefore, Box et al. (134) recommend that these algorithms should be combined. Thus, initially one requires to perform a single stage of the DSC method to obtain a bracket on x_m and then switch over to Powell's algorithm, thereby benefitting from the advantages of both the techniques.

The algorithms of the DSC method and Powell's method for univariate optimization are discussed by numerous authors (131,133 - 137). Beightler et al. (136) have demonstrated the use of both these algorithms for solution of simple quadratic functions.

Robinson (137) has presented complete logic diagrams for both these methods for univariate search.

The above mentioned information available in the literature could be utilised to design a multivariable general purpose optimization software.

3.2.4 Simplex algorithm :

A completely different approach to the problem of multivariable minimization using the direct-search method, is that which explores parameter space by means of some geometric configuration of points rather than a set of directions. This approach was utilised by Spendley et al. (138) to develop the original simplex method.

The simplex search is based on the observation that the first order experimental design requiring smallest number of points is the regular simplex. In N dimension, a regular simplex is a polyhedron composed of N + 1 equidistant points which form its vertices, for example a simplex in two dimension is a equilateral triangle. The main property that a simplex possesses, is that, a new simplex can be formed on any face of a given simplex by addition of only one point.

The method begins by setting up a regular simplex in the space of the independent variables and evaluating the function of each vertex. The vertex with highest functional value is located. This worst vertix is then reflected through the centeroid of

remaining vertices to generate a new point which is used to complete the next simplex. Thus, the simplex proceeds to search the minimum. In regions near the minima, there is a scope to reduce the size of simplex by certain scaling rules suggested by Spendley et al. (138). Simplex movement, convergence criteria and such relevant details related to the simplex algorithm are mentioned in the references (133, 139 - 143).

3.2.5 Modified simplex algorithm :

The original simplex method is a slow algorithm and has scaling problems i.e. the size of simplex cannot be expanded. To overcome such difficulties, Nelder and Mead (140) have suggested modifications, wherein the regularity of design is abandoned and the simplex automatically rescales itself according to the local geometry of the function under investigation, through steps like reflection, expansion and contraction.

Suppose that for iteration k, the vertices of the simplex are x_o^k, x_1^k, ... x_n^k and corresponding function values F_o, F_1, ... F_n ordered such that $F_n > F_{n-1} > \ldots\ldots > F_1 > F_o$ where x_o^k is the best vertix and x_n^k is the worst. Let 'c' be the centroid of the vertices x_o^k, x_1^k, ... x_{n-1}^k given by equation (3.6).

$$c_i = \frac{1}{n} \sum_{j=0}^{n-1} x_{ji}^{(k)} , \qquad (3.6)$$

where i = 1 to n

Then,as in the original simplex method,the worst vertex x_n^k is to be replaced and a simple reflection move is tried first using a reflection coefficient α (where $\alpha > 0$) ; thereby obtaining a new point (x_r^k), as shown in equation (3.7).

$$x_r^k = c^k + \alpha (c^k - x_n^k) \qquad (3.7)$$

There are then three possible cases to be considered : x_r^k is a point such that $F_o < F_r < F_{n-1}$; $F < F_o$ so that x_r^k would be a new best point ; $F_r > F_{n-1}$ so that x_r^k would be a new worst point.

In the case $F_o < F_r < F_{n-1}$,then x_r^k replaces x_n^k and the iteration in complete.

However,if reflection produces the best point then,in the direction of reflection,the simplex expands by defining the point (x_e^k), using a expansion coefficient β (where $\beta > 1$) as shown in equation (3.8).

$$x_e^k = c^k + \beta (x_r^k - c^k) \qquad (3.8)$$

If $F_e < F_o$ the expansion is considered to be successful and x_e^k replaces x_n^k ; otherwise the expansion is deemed to have failed and x_n^k is replaced by x_r^k. In either event the iteration is then complete.

If original reflection resulted in a new worst point,then it is assumed that the size of design is too large to allow any progress to be made; therefore simplex contracts by defining points

x_c^k using a contraction coefficient γ $(0 < \gamma < 1)$.

$$x_c^k = c^k + \gamma (x_n^k - c^k) \qquad \text{if } F_n < F_r \qquad ..(3.9)$$

$$x_c^k = c^k + \gamma (x_r^k - c^k) \qquad \text{if } F_n > F_r$$

If $F_c < \min (F_n, F_r)$ then x_c^k replaces x_n^k ; otherwise a more comprehensive contraction is carried out by having the distances from the best point x_o^k of all the other vertices of the simplex. In either case the iteration is then complete.

Using steps of reflection, expansion and contraction the iteration continues till the function values do not vary significantly. Nelder and Mead (140) recommend the values of $\alpha = 1$, $\beta = 2$, $\gamma = 0.5$ to be employed. Numerical comparison by Box and Draper (134) indicates this algorithm to be the most efficient of all sequential techniques, very reliable and extremely robust. Parkinson and Hutchinson (141) investigated the manner of construction of initial simplex and observed that the shape of initial simplex was not important.

In presence of constraints on the objective function, the generation of initial simplex is problematic because, for any given simplex size parameter α, it is likely that many points of the regular simplex could be unfeasible. In order to overcome this difficulty, Box (144) proposed a set of trial points be generated randomly and sequentially. The total number of points (P) to be used should be no less than N+1 but can be larger. The recommended value of P based on numerical experiments by Box happens to be P = 2N.

3.3.0 SALIENT FEATURES OF THE SOFTWARE DEVELOPED IN THIS INVESTIGATION :

The software developed in this work based on the DSC-Powell combination algorithm has the flexibility to optimise any nonlinear function which may be included in the subroutine called ERROR, that is, ERROR is user specific and does not interfere with the main body of the program. Subroutine ERROR should calculate and return 'error values' corresponding to the current values of variables K_1, K_2 etc.

The main body of program is oriented towards generation of the equidistant points required in the DSC method for quadratic fitting. Once such points are generated the informaton is transferred to a subroutine DSC wherein the quadratic fitting as in equation (3.4) is done. This subroutine generates the DSC optima values in one direction thereby bracketing the optima. At this stage the information is transferred to the main program for Powell search wherein the value of minima in one direction (variable under consideration) is obtained. There after retaining the value of minima of the variable for which the search has concluded, the direction of optimization is changed and second variable is optimized in likewise manner. When the values of minima for all the variables under consideration have been obtained one cycle of iterations is over and the prescribed accuracy of convergence is checked, If the accuracy conditions are satisfied, the minima values so obtained are designated as optimum values and the program comes to its logical end. Otherwise, if convergence conditions are not

satisfied, the entire process of DSC-Powell search is repeated until optimum values are obtained which satisfy the convergence criterion.

A programme incorporating all the above mentioned aspects which could be conveniently used to optimize upto six variables is listed in Appendix (A.2.1)

In this investigation, the modified simplex algorithm of Nelder and Mead (140) has also been used. The salient features of this software adopted in this work are the following :- Standard codes for this algorithm is available in literature (142,143). The program in BASIC developed by Valko and Vajda (143) could be adopted conveniently for this purpose. For objective functions with implicit constraints, the recommendation due to Box (144) could be easily implemented by setting the value of P to 2N instead of N+1. Further, the points could be generated randomly and sequentially with minor modifications in the program. The detailed program incorporating these changes in Nelder and Meads simplex algorithm proposed by Valko and Vajda is listed in Appendix (A.2.2).