A LEAST-SQUARE-DISTANCE CURVE-FITTING TECHNIQUE

by John Q. Howell

Langley Research Center
Hampton, Va. 23365
NASA TN D-6374

2. Government Accession No.  

3. Recipient's Catalog No.  

4. Title and Subtitle  
A LEAST-SQUARE-DISTANCE CURVE-FITTING TECHNIQUE

5. Report Date  
July 1971

6. Performing Organization Code  

7. Author(s)  
John Q. Howell

L-7675

9. Performing Organization Name and Address  
NASA Langley Research Center  
Hampton, Va. 23365

10. Work Unit No.  
125-21-21-01

11. Contract or Grant No.  

12. Sponsoring Agency Name and Address  
National Aeronautics and Space Administration  
Washington, D.C. 20546

13. Type of Report and Period Covered  
Technical Note


15. Supplementary Notes  

16. Abstract  
A method is presented for fitting a function with n parameters \( y = f(x_1, x_2, \ldots, x_n; x) \) to a set of N data points \( (X_i, Y_i) \) in a manner that minimizes the sum of the squares of the distances from the data points to the curve. A differential-correction scheme is used to solve for the parameters in an iterative manner until the best fit is obtained. Two methods for finding the distances from the data points to the curve and a listing of the curve-fitting computer program are also given.

17. Key Words (Suggested by Author(s))  
Curve fitting  
Least-square distance  
Least squares

18. Distribution Statement  
Unclassified – Unlimited

19. Security Classif. (of this report)  
Unclassified

20. Security Classif. (of this page)  
Unclassified

21. No. of Pages  
24

22. Price*  
$3.00

*For sale by the National Technical Information Service, Springfield, Virginia 22151
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUMMARY</td>
<td>1</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>SYMBOLS</td>
<td>2</td>
</tr>
<tr>
<td>DERIVATION AND DISCUSSION OF NEW TECHNIQUE</td>
<td>3</td>
</tr>
<tr>
<td>EXAMPLES OF APPLICATION OF NEW TECHNIQUE</td>
<td>6</td>
</tr>
<tr>
<td>CONCLUDING REMARKS</td>
<td>9</td>
</tr>
<tr>
<td>APPENDIX A – TWO NUMERICAL METHODS FOR FINDING THE DISTANCE FROM A POINT TO A CURVE</td>
<td>13</td>
</tr>
<tr>
<td>APPENDIX B – COMPUTER PROGRAM FOR LEAST-SQUARE-DISTANCE TECHNIQUE</td>
<td>15</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>22</td>
</tr>
</tbody>
</table>
A LEAST-SQUARE-DISTANCE CURVE-FITTING TECHNIQUE

By John Q. Howell
Langley Research Center

SUMMARY

A method is presented for fitting a function with $n$ parameters

$$y = f(a_1, a_2, \ldots, a_n; x)$$

to a set of $N$ data points $(x_i, y_i)$ in a manner that minimizes the sum of the squares of the distances from the data points to the curve. A differential-correction scheme is used to solve for the parameters in an iterative manner until the best fit is obtained. Two methods for finding the distances from the data points to the curve and a listing of the curve-fitting computer program are also given.

INTRODUCTION

Most of the generally used methods of fitting a curve to a set of data points minimize a function of the vertical distances from the points to the curve. For example, if $(x_i, y_i)$ is a set of $N$ points and $y = f(a_1, a_2, \ldots, a_n; x)$ is a curve with $n$ parameters, then the method of least squares gives values of the parameters that minimize

$$E = \sum_{i=1}^{N} [y_i - f(a_1, a_2, \ldots, a_n; x_i)]^2$$

This may be done by taking partial derivatives with respect to the parameters and setting each of the resulting equations equal to zero; that is,

$$\frac{\partial E}{\partial a_j} = 0 = -2 \sum_{i=1}^{N} (y_i - \tilde{f}_i) \frac{\partial \tilde{f}_i}{\partial a_j} \quad (j = 1, 2, \ldots, n)$$

where

$$\tilde{f}_i = f(a_1, a_2, \ldots, a_n; x_i)$$

This set of $n$ equations, sometimes called the normal equations, is then solved for the parameters. As is well known, if $f$ is linear in the parameters, for example, a polynomial in $x$ (ref. 1), a set of simultaneous linear equations merely has to be solved. However, in general, more complicated functions yield simultaneous equations that are
nonlinear. In this case \( f \) may be expanded in a truncated Taylor series about a point in parameter space, and in this manner the nonlinear normal equations can be linearized and solved by iteration. The end result is a set of parameters that yields a minimum of equation (1). This is called either the Gauss-Newton method or the method of nonlinear least squares. However, when \( f(x) \) has a region where its derivative is large or when both \( x_i \) and \( y_i \) have similar error bounds, it may be more desirable to minimize the distance from each data point to its nearest point on the curve. This minimum distance is the same as the perpendicular distance from the data point to the curve. Scarborough (ref. 2) gives a method for curve fitting that minimizes the sum of the squares of these distances but his method is limited to first-order polynomials. Reed (ref. 3) and Kendall and Stuart (ref. 4) give schemes that are applicable to polynomials of higher order. These same methods are useful for any function in which the parameters enter in a linear fashion. Guest (ref. 5) describes a related technique that minimizes the perpendicular distance from each data point to a straight line tangent to the curve. This tangent is taken at the point on the curve having the same \( x \)-coordinate as the data point.

The purpose of the present work is to derive and demonstrate the use of a curve-fitting technique that minimizes the least-square distances from each data point to the curve. The technique described herein works for a general function \( f \) and is most useful when the function being fitted contains regions where the slope is small as well as where the slope is large. It is also useful when the data points have error bounds associated with both the \( x \)- and \( y \)-coordinates. In the latter case, this technique implicitly assumes identical error for both coordinates. Generally the data points and the function can be scaled so that this condition is met. Other techniques that are less time consuming may also be used in these situations. For example, a judicious choice of weights often makes possible the use of a standard least-squares procedure. However, in these cases a particular choice of weights seldom works for more than a few sets of data. The technique described herein does not have this disadvantage since it provides a fit even when all the weights are set equal to 1.

**SYMBOLS**

- \( A_{jk} \) defined by equation (10)
- \( B_j \) defined by equation (9)
- \( D_i \) distance from \( (x_i, y_i) \) to nearest point on curve, equations (4) and (16)
- \( \hat{D}_i \) value of \( D_i \) using old parameters, equation (6)
\[ D_i(x) \] distance from \((\tilde{x}_i, \tilde{y}_i)\) to some point on curve, equation (17)

\[ E \] sum of squares of distances from data points to curve, equations (1) and (3)

\[ f(x) \] function to be fitted to data points; \( y = f(x) \)

\[ \tilde{f}_i \] value of function at \( \tilde{x}_i \), equation (2)

\[ \hat{f}_i \] value of function at \( x_i \) with old parameters, equation (8)

\[ n \] number of function parameters

\[ N \] number of data points

\[ w_i \] weight associated with data point \((\tilde{x}_i, \tilde{y}_i)\)

\[ (\tilde{x}_i, \tilde{y}_i) \] coordinates of one data point of set to which \( y = f(x) \) is being fitted

\[ (x_i, y_i) \] coordinates of point on curve nearest data point \((\tilde{x}_i, \tilde{y}_i)\)

\( \alpha_1, \alpha_2, \ldots, \alpha_n \) parameters of \( y = f(x) \)

\( \hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n \) old parameters of \( y = f(x) \) during iteration to find least-square-distance fit

\( \sigma \) root-mean-square deviation of data points from curve, equation (14)

**DERIVATION AND DISCUSSION OF NEW TECHNIQUE**

There is given a set of \( N \) data points \( \{\tilde{x}_i, \tilde{y}_i\} \) to which is to be fitted the function \( y = f(\alpha_1, \alpha_2, \ldots, \alpha_n; x) \), where \( \alpha_1, \alpha_2, \ldots, \alpha_n \) are parameters. To obtain this fit the sum of the least-square distances (sum of the squares of the shortest distances) from the data points to the curve must be minimized. By using one of the techniques given in appendix A, the coordinates \((x_i, y_i)\) are found for the point on the curve that is nearest each data point \((\tilde{x}_i, \tilde{y}_i)\). Then it is desired to minimize

\[
E = \sum_{i=1}^{N} w_i D_i^2
\]  

(3)
where the distance from the \( i \)th data point to the curve is

\[
D_1 = \left[ (x_i - \bar{x}_i)^2 + (y_i - \bar{y}_i)^2 \right]^{1/2}
\]

(4)

The weight to be associated with each point is given by \( w_i \). To minimize equation (3) it is necessary to solve the set of \( n \) normal equations

\[
\frac{\partial E}{\partial \alpha_j} = 0 = 2 \sum_{i=1}^{N} w_i D_i \frac{\partial D_i}{\partial \alpha_j}
\]

\( (j = 1, 2, \ldots, n) \) (5)

In general this is a set of nonlinear simultaneous equations. To solve the set, an iterative procedure sometimes called the method of differential correction can be used. First a Taylor series expansion of \( D_1 \) is made about some point \((\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n)\) in parameter space. Since \( D_1 \) is a known function of the parameters, the expansion can easily be written

\[
D_1(\alpha_1, \alpha_2, \ldots, \alpha_n) = \hat{D}_i + \sum_{k=1}^{n} \frac{\partial D_i}{\partial \alpha_k} \Delta \alpha_k + \frac{1}{2} \sum_{k,l=1}^{n} \frac{\partial^2 D_i}{\partial \alpha_k \partial \alpha_l} \Delta \alpha_k \Delta \alpha_l + \ldots
\]

(6)

where

\[
\hat{D}_i = D_1(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n)
\]

and

\[
\Delta \alpha_k = \alpha_k - \hat{\alpha}_k
\]

Now in equation (6) all higher order terms are dropped and only the terms that are linear in \( \Delta \alpha \) are kept. Then substitution of equation (6) into equation (5) gives

\[
0 = \sum_{i=1}^{N} w_i \left( \hat{D}_i + \sum_{k=1}^{n} \frac{\partial \hat{D}_i}{\partial \alpha_k} \Delta \alpha_k \right) \frac{\partial \hat{D}_i}{\partial \alpha_j}
\]

\( (j = 1, 2, \ldots, n) \) (7)

Also, equation (4) yields

\[
\frac{\partial \hat{D}_i}{\partial \alpha_k} = \frac{\hat{f}_i - \bar{y}_i}{\hat{D}_i} \frac{\partial \hat{f}_i}{\partial \alpha_k}
\]

(8)

where

\[
\hat{f}_i = f(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n; x_i)
\]
Now, by definition,

\[ B_j = -\sum_{i=1}^{N} w_i \left( \hat{f}_i - \bar{y}_i \right) \frac{\partial \hat{f}_i}{\partial \alpha_j} \]  \hspace{1cm} (9)

and

\[ A_{jk} = \sum_{i=1}^{N} w_i \left( \frac{\hat{f}_i - \bar{y}_i}{D_i} \right)^2 \frac{\partial \hat{f}_i}{\partial \alpha_k} \frac{\partial \hat{f}_i}{\partial \alpha_j} \]  \hspace{1cm} (10)

and equation (7) then becomes

\[ B_j = \sum_{k=1}^{n} A_{jk} \Delta \alpha_k \]  \hspace{1cm} (j = 1, 2, \ldots, n)  \hspace{1cm} (11)

The equations needed to fit the curve \( y = f(\alpha_1, \alpha_2, \ldots, \alpha_n; x) \) to the set of \( N \) points \( \{x_i, y_i\} \) have now been derived. To use this procedure, a starting point in parameter space is first chosen and designated \( (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n) \). Next the distances from each data point to the curve are found by using perhaps one of the techniques outlined in appendix A. Then \( B_j \) and \( A_{jk} \) are found from equations (9) and (10) and the simultaneous linear equations in equations (11) are solved for the quantities \( \Delta \alpha_1, \Delta \alpha_2, \ldots, \Delta \alpha_n \). Lastly, the new parameters are obtained from

\[ \alpha_k = \hat{\alpha}_k + \Delta \alpha_k \]  \hspace{1cm} (k = 1, 2, \ldots, n)  \hspace{1cm} (12)

This set of new parameters is then used as a new starting point and the cycle repeated. This iteration is carried out until \( \alpha_k \) converges \( (\Delta \alpha_k << \alpha_k) \) or until it is obvious that a convergence will not be achieved. In the latter case a better starting point in parameter space generally leads to convergence. It should be pointed out that the point in parameter space to which equations (9) to (11) converge may not be an absolute minimum, that is, the best fit. The end point of the process may be either a relative maximum or a relative minimum. The former case is rather unlikely but at any rate is easily detected by inspection of the value of equation (3) after each iteration. In the latter case a new starting point in parameter space must be chosen to see if convergence is achieved to a point where a smaller value of equation (3) is obtained. Unfortunately, it is in general difficult to tell when the best fit has been found, but once a fit sufficient for the particular need is located, it is not necessary to search farther.
EXAMPLES OF APPLICATION OF NEW TECHNIQUE

The two examples given herein are chosen to demonstrate that for some cases the least-square-distance curve-fitting technique gives better results than the standard least-squares method. The first example arose when the author was trying to reduce some experimental plasma-physics data and led ultimately to the least-square-distance curve-fitting technique described in this paper. The second example is chosen since it is commonly known that standard least-squares procedures do not work well on this type of function. For the examples presented here the weights are set equal to one. By properly choosing the weights it may be possible to obtain a fit with the least-squares technique that is as good as that obtained with the least-square-distance method. However, for a different function and often for a different set of data points, a new set of weights would have to be chosen to achieve a good fit again. The least-square-distance technique described in this report does not have this disadvantage.

Because of the extra computations involved in finding the closest point on the curve, the least-square-distance method takes more computer time than the least-squares method. Based on the following two examples it is determined that the least-square-distance method is longer by a factor of approximately 2.5.

Example I

Example I is taken from the field of plasma physics where a common diagnostic tool is the Langmuir probe. The current versus voltage characteristic of this probe is given approximately by

\[ y = \alpha_1 x^2 + \alpha_2 x + \alpha_3 + \alpha_4 e^{\alpha_5 x} \]  

(13)

where \( x \) is the voltage, \( y \) is the current, and \( \alpha_1, \alpha_2, \ldots, \alpha_5 \) are a set of adjustable parameters. The parameters \( \alpha_4 \) and \( \alpha_5 \) are always positive so the exponential term is large for \( x \) positive and small for \( x \) negative.

When the Langmuir probe is used as a diagnostic tool, the current is typically measured for a large set of voltage points. Then some curve-fitting technique is used to obtain a fit to the experimental data. The value of \( \alpha_5 \) is of interest as the electron temperature can be obtained from it. This temperature is then used to calculate a particular voltage in the region where the exponential term is small and the fitted function is used to obtain the corresponding current. The ion density can then be obtained from this current. From this description it is apparent (1) that both \( x \) and \( y \) may be in error and (2) the fit to the experimental data must be good both in regions where the exponential term is large and where it is small.
Figure 1 shows the fit that is obtained in one particular case by using the nonlinear-least-squares technique. The values of the parameters and the weighted root-mean-square deviation obtained by the curve-fitting schemes are shown in the legend of the figure. The weighted root-mean-square deviation is defined by

$$\sigma = \left( \frac{\sum_{i=1}^{N} w_i D_i^2}{\sum_{i=1}^{N} w_i} \right)^{1/2}$$

(14)

Figure 1.- Nonlinear-least-squares fit used with equation (13). $\alpha_1 = -0.2034$; $\alpha_2 = -1.166$; $\alpha_3 = -1.424$; $\alpha_4 = 3.750$; $\alpha_5 = 2.028$; $\sigma = 3.3$. 

Figure 1 - Nonlinear-least-squares fit used with equation (13). $\alpha_1 = -0.2034$; $\alpha_2 = -1.166$; $\alpha_3 = -1.424$; $\alpha_4 = 3.750$; $\alpha_5 = 2.028$; $\sigma = 3.3$. 

Figure 1 shows the fit that is obtained in one particular case by using the nonlinear-least-squares technique. The values of the parameters and the weighted root-mean-square deviation obtained by the curve-fitting schemes are shown in the legend of the figure. The weighted root-mean-square deviation is defined by

$$\sigma = \left( \frac{\sum_{i=1}^{N} w_i D_i^2}{\sum_{i=1}^{N} w_i} \right)^{1/2}$$

(14)

Figure 1 - Nonlinear-least-squares fit used with equation (13). $\alpha_1 = -0.2034$; $\alpha_2 = -1.166$; $\alpha_3 = -1.424$; $\alpha_4 = 3.750$; $\alpha_5 = 2.028$; $\sigma = 3.3$. 

Figure 1 shows the fit that is obtained in one particular case by using the nonlinear-least-squares technique. The values of the parameters and the weighted root-mean-square deviation obtained by the curve-fitting schemes are shown in the legend of the figure. The weighted root-mean-square deviation is defined by

$$\sigma = \left( \frac{\sum_{i=1}^{N} w_i D_i^2}{\sum_{i=1}^{N} w_i} \right)^{1/2}$$

(14)

Figure 1 - Nonlinear-least-squares fit used with equation (13). $\alpha_1 = -0.2034$; $\alpha_2 = -1.166$; $\alpha_3 = -1.424$; $\alpha_4 = 3.750$; $\alpha_5 = 2.028$; $\sigma = 3.3$. 

Figure 1 shows the fit that is obtained in one particular case by using the nonlinear-least-squares technique. The values of the parameters and the weighted root-mean-square deviation obtained by the curve-fitting schemes are shown in the legend of the figure. The weighted root-mean-square deviation is defined by

$$\sigma = \left( \frac{\sum_{i=1}^{N} w_i D_i^2}{\sum_{i=1}^{N} w_i} \right)^{1/2}$$

(14)

Figure 1 - Nonlinear-least-squares fit used with equation (13). $\alpha_1 = -0.2034$; $\alpha_2 = -1.166$; $\alpha_3 = -1.424$; $\alpha_4 = 3.750$; $\alpha_5 = 2.028$; $\sigma = 3.3$. 

Figure 1 shows the fit that is obtained in one particular case by using the nonlinear-least-squares technique. The values of the parameters and the weighted root-mean-square deviation obtained by the curve-fitting schemes are shown in the legend of the figure. The weighted root-mean-square deviation is defined by

$$\sigma = \left( \frac{\sum_{i=1}^{N} w_i D_i^2}{\sum_{i=1}^{N} w_i} \right)^{1/2}$$

(14)

Figure 1 - Nonlinear-least-squares fit used with equation (13). $\alpha_1 = -0.2034$; $\alpha_2 = -1.166$; $\alpha_3 = -1.424$; $\alpha_4 = 3.750$; $\alpha_5 = 2.028$; $\sigma = 3.3$. 

Figure 1 shows the fit that is obtained in one particular case by using the nonlinear-least-squares technique. The values of the parameters and the weighted root-mean-square deviation obtained by the curve-fitting schemes are shown in the legend of the figure. The weighted root-mean-square deviation is defined by

$$\sigma = \left( \frac{\sum_{i=1}^{N} w_i D_i^2}{\sum_{i=1}^{N} w_i} \right)^{1/2}$$

(14)

Figure 1 - Nonlinear-least-squares fit used with equation (13). $\alpha_1 = -0.2034$; $\alpha_2 = -1.166$; $\alpha_3 = -1.424$; $\alpha_4 = 3.750$; $\alpha_5 = 2.028$; $\sigma = 3.3$. 

Figure 1 shows the fit that is obtained in one particular case by using the nonlinear-least-squares technique. The values of the parameters and the weighted root-mean-square deviation obtained by the curve-fitting schemes are shown in the legend of the figure. The weighted root-mean-square deviation is defined by

$$\sigma = \left( \frac{\sum_{i=1}^{N} w_i D_i^2}{\sum_{i=1}^{N} w_i} \right)^{1/2}$$

(14)
where $D_i$ is the vertical distance in the case of the least-squares technique and the shortest distance from the point to the curve for the least-square-distance method. The least-squares fit for $x$ negative is not acceptable (fig. 1) but the fit is good for $x$ positive where the exponential term is large.

The same set of data points is then used in a program based on the least-square-distance technique derived in the present paper. The result of this fit is shown in figure 2. It is immediately apparent that the fit is much better and is in fact good enough to extract the desired information for the further data analysis as described earlier. The legends of figures 1 and 2 show that only $a_5$ agrees to within 20 percent. As

![Figure 2.- Least-square-distance fit used with equation (13) and same data points as in figure 1. $a_1 = 0.05013; a_2 = 0.3994; a_3 = 0.5654; a_4 = 2.398; a_5 = 2.2611; \sigma = 0.077.$](image)
would be expected from a comparison of figures 1 and 2, the coefficients of the polynomial portion of the function are in violent disagreement. The reason the least-squares technique does not do so well is that it finds a fit that is good in the high-slope and high-magnitude region at the expense of the fit in the small-slope and small-magnitude region.

Example II

Example II is chosen to show that the least-square-distance curve-fitting technique fits functions with a singularity. The function chosen is

\[ y = \alpha_1 x^3 + \alpha_2 x^2 + \alpha_3 x + \alpha_4 + \frac{\alpha_5}{x - 25} \]  

(15)

In figure 3 the results of the nonlinear-least-squares curve-fitting scheme are shown. The values of the parameters and the weighted root-mean-square deviation of the points from the curve are shown in the legend of the figure. The least-square-distance fit of the same function to the same points is shown in figure 4. As in example I, the fit to the small-magnitude points is better when the least-square-distance technique is used while both methods give similar fits for the large-magnitude points. In example II the initial parameter guess for the least-square-distance method is more critical than usual. With a bad initial guess both distance-finding techniques described in appendix A sometimes achieve convergence to a point on the wrong side of the singularity. This can also happen if the fitting function has a very sharp peak, in which case the distance-finding scheme may achieve convergence to a point on the wrong side of the peak. Of course, if it is desired to fit a function of this type to several sets of data, the program can be designed to alleviate this problem, but the fitting routine has to be different for each particular function.

CONCLUDING REMARKS

In the present paper the least-square-distance curve-fitting method is derived and examples of its use are presented. This technique fits a function with \( n \) parameters \( y = f(\alpha_1, \alpha_2, \ldots, \alpha_n; x) \) to a set of \( N \) data points \( \{x_1, y_1\} \) by minimizing the sum of the squares of the distances from the data points to the curve. A differential-correction scheme is used to solve for the parameters in an iterative manner until the best fit is obtained. Two examples of the use of this technique are presented, both involving functions having large slope variations. In both cases the least-squares fit is found to be lacking when compared to the least-square-distance fit. It is found that the least-squares technique fits the curve to points in the regions of large slope and large magnitude at the expense of the fit in regions of small slope and small magnitude. This does not happen for the least-square-distance method presented in this paper since the sum of the squares
Figure 3. Nonlinear-least-squares fit used with equation (15). $a_1 = 0.000819; a_2 = -0.06442; a_3 = 0.9939; a_4 = 3.886; a_5 = -12.75; \sigma = 2.6.$
Figure 4.- Least-square-distance fit used with equation (15) and same data points as in figure 3. $a_1 = 0.000709; \quad a_2 = -0.05129; \quad a_3 = 0.5789; \quad a_4 = 6.147; \quad a_5 = -14.04; \quad \sigma = 0.21.$
of the distances from the data points to the curve is minimized. Hence, for functions of this type the least-square-distance technique fits a function to a set of points more accurately than the least-squares method, unless much time is spent in customizing the least-squares weights to the particular function and particular set of data.

Langley Research Center,
National Aeronautics and Space Administration,
APPENDIX A

TWO NUMERICAL METHODS FOR FINDING THE DISTANCE FROM A POINT TO A CURVE

In appendix A two methods are presented for finding the distance from the data point \((x_i, y_i)\) to the curve \(y = f(x)\). The first method minimizes the distance from the curve to the data point, while the second method finds the perpendicular from the curve to the data point. Both these methods locate the point on the curve \(x_i, f(x_i)\) nearest the data point. The distance from the data point to the curve is then given by

\[
D_i = \left[\left(\bar{y}_i - y_i\right)^2 + (\bar{x}_i - x_i)^2\right]^{1/2}
\]  

For some very simple cases this point can be found analytically but the assumption is made here that \(f(x)\) is of such complexity that this is impossible.

**Method I**

The distance from some point on the curve \(y = f(x)\) to the data point \((\bar{x}_i, \bar{y}_i)\) is

\[
D_i(x) = \left[\left(f(x) - \bar{y}_i\right)^2 + (x - \bar{x}_i)^2\right]^{1/2}
\]  

(17)

Now an \(x\) such that \(D_i\) is minimum may be found by solving

\[
\frac{dD_i}{dx} = 0
\]  

(18)

For the case of \(D_i \neq 0\) (for \(D_i = 0\), the trivial solution is \(x_i = \bar{x}_i\) and \(y_i = \bar{y}_i\)), it is seen from equation (18) that \(x_i\) must satisfy the equation

\[
\left[f(x_i) - \bar{y}_i\right] \frac{df(x_i)}{dx} + (x_i - \bar{x}_i) = 0
\]  

(19)

Once \(x_i\) is found, \(y_i\) is obtained from \(y_i = f(x_i)\), and equation (16) is used to find \(D_i\). Equation (19) can be solved by any convenient method. For cases where the second derivative of \(f(x)\) is obtainable, the author has used the Newton-Raphson method with good success. It should be kept in mind that in some cases the solution of equation (19) may yield a \(D_i\) that is a relative maximum or a relative minimum instead of the absolute minimum that is desired. Fortunately these cases are rare.
The second method may be called the method of successive tangents and does not require higher derivatives of \( f(x) \). Consequently it is much more useful for complicated functions. To use this method a point on the curve is initially chosen near where the closest point is thought to be. This initial guess may be designated \( [x_i, f(x_i)] \) and a straight line fitted through this point tangent to the curve. This can be done by using either \( f(x_i) \) and \( f'(x_i) \) or \( f(x_i) \) and \( f(x_i + \Delta x) \). In the latter case \( \Delta x \) is some small arbitrarily chosen increment. Once the straight line is found, a perpendicular is dropped to it from the data point \( (x_i, y_i) \). A better estimate of the closest point on the curve is now obtained by letting the new \( x_i \) be the \( x \)-coordinate of the foot of the perpendicular on the straight line. Then a second straight line tangent to the curve may be fitted through the new point \( [x_i, f(x_i)] \). This process is repeated until two successive \( x_i \)'s agree to within some previously chosen increment. For cases where \( f'(x) \) is easily obtained the author has used this scheme with good success. If the Newton-Raphson method is used with the first method and if \( f''(x) \) is zero these two methods are equivalent.
APPENDIX B

COMPUTER PROGRAM FOR LEAST-SQUARE-DISTANCE TECHNIQUE

Appendix B contains a description and listing of a least-square-distance curve-fitting program written in FORTRAN. The procedure for finding the distance from the data point to the curve is built into the curve-fitting subroutine. The method used is the successive-tangent method described in appendix A. The curve-fitting subroutine also has a damping procedure (ref. 6) included for increased stability. This program has operated satisfactorily for the author with several different functions but has not been tested extensively.

Main Program

It is felt that a description of the main program is not needed since any potential user has to write the main program around his own particular application.

Least-Square-Distance Curve-Fitting Subroutine

This subroutine assumes the existence of a linear-simultaneous-equation solver called SIMSOL. It is called by the statement

\[ \text{Call SIMSOL}(A,B,M) \]

and solves the equation in \( M \) unknowns given by \( AX = B \). The solution vector for \( X \) is returned in \( B \). The curve-fitting subroutine also calls the subroutine FUNC described subsequently. A description of the calling procedure for the curve-fitting subroutine follows.

Use:

\[ \text{Call LSD}(X,Y,W,N,AL,M,ERR,RMS) \]

\[ \{X, Y\} \quad \text{Vectors containing x- and y-coordinates of data points to which function is being fitted.} \]

\[ W \quad \text{Vector containing weight associated with each point.} \]

\[ N \quad \text{The number of points being supplied to subroutine by main program.} \]

\[ AL \quad \text{Vector containing values of function parameters. Initially a trial set must be supplied. The curve-fitting subroutine iterates and returns a better set.} \]

\[ M \quad \text{The number of parameters in function being fitted.} \]
APPENDIX B – Continued

ERR  An error criterion that must be supplied to subroutine by main program.
The subroutine iterates until $\text{RMS}_{\text{old}} - \text{RMS}_{\text{new}} < \text{ERR} \times \text{RMS}_{\text{new}}$.

RMS  Weighted root-mean-square deviation of data points from curve defined by

$$
\text{RMS} = \left( \frac{\sum_{i=1}^{N} w_i D_i^2}{\sum_{i=1}^{N} w_i} \right)^{1/2}
$$

where $D_i$ is distance of $i$th point from nearest point on curve.

Restrictions: (1) $X,Y$ and $W$ are all dimensioned 50 and hence $N \leq 50$.
(2) $AL$ is dimensioned 10 and hence $M \leq 10$.
(3) A linear-simultaneous-equation solver must be provided as described previously.
(4) A subroutine called FUNC containing information about the function must be supplied. An example is described next.

Description of Subroutine FUNC

The subroutine FUNC contains information about the function being fitted. This subroutine is called by the curve-fitting subroutine described previously. The subroutine listing included herein is used to fit equation (13) to a set of data points as shown in figure 2 and is intended to be an example of how this subroutine may be written.

Use:  \begin{align*}
\text{Call } & \text{FUNC}(X,Y,N,AL,XDER,DER) \\
X & \text{Vector containing values of independent variable.} \\
Y & \text{Vector used to return values of dependent variable to curve-fitting subroutine. For example, } Y(I) \text{ must contain the value of the function evaluated at } X(I) \text{ for } I = 1 \text{ to } I = N. \\
N & \text{The number of } X \text{ values being supplied to subroutine. If } N = 1 \text{ only } X(1) \text{ is supplied and the value of the function and its } x \text{-derivative must be returned in } Y(1) \text{ and } XDER, \text{ respectively. For other values of } N, \text{ both } Y \text{ and } DER \text{ must be filled and } XDER \text{ need not be calculated.} \\
AL & \text{Current values of function parameters being supplied to } \text{FUNC by LSD.}
\end{align*}
APPENDIX B – Continued

XDER Variable containing value of x-derivative of function evaluated at X(1).
   It need be calculated only when N = 1.

DER Matrix containing derivatives of function with respect to all function parameters, each evaluated at X(I), for I = 1 to I = N. This matrix must be filled by FUNC whenever N > 1. The defining equation is

\[
\text{DER}(K,I) = \left[ \frac{\partial f}{\partial \alpha_K} \right]_{X=X_I}
\]

where \( \alpha_K \) is Kth function parameter.

Restrictions: (1) X and Y are dimensioned 50 so N \( \leq 50 \).
(2) AL is dimensioned 10.
(3) DER is dimensioned (10,50).
APPENDIX B – Continued

MAIN PROGRAM

THIS PROGRAM READS THE INITIAL GUESS AT THE PARAMETERS AND THE DATA POINTS AND THEN CALLS THE LEAST SQUARE DISTANCE CURVE FITTER.

PROGRAM CFIT(INPUT,OUTPUT)
DIMENSION X(50),Y(50),W(50),AL(10)
ERR=1.E-5
1 READ 1000,N,(AL(I),I=1,N)
1000 FORMAT(I2/(E10))
READ 1001,N,(X(I),Y(I),W(I),I=1,N)
1001 FORMAT(I2/(3E10))
CALL LSD(X,Y,W,N,AL,M,ERR,RMS)
GO TO 1
END

LEAST SQUARE DISTANCE CURVE FITTING SUBROUTINE

SUBROUTINE LSD(X,Y,W,N,AL,M,ERR,RMS)

THIS IS A LEAST SQUARE DISTANCE CURVE FITTING SUBROUTINE.
IT HAS BUILT IN THE SUCCESSIVE TANGENT LINE SCHEME TO FIND THE DISTANCE FROM A DATA POINT TO THE CURVE.
IT CALLS SIMSOL(AX=B) TO SOLVE THE LINEAR SIMULTANEOUS EQUATION AX=B IN M UNKNOWNS.
IT CALLS FNC TO OBTAIN INFORMATION ABOUT THE FUNCTION BEING FITTED.
DIMENSION X(50),Y(50),W(50),AL(10),D(10,50),B(10),A(10,10)
DIMENSION DIS(50),X1(50),W2(50),ID(2),Y1(50),DX(50)

IF IPRINT=0, THERE IS NO PRINTED OUTPUT FROM THIS SUBROUTINE.
IPRINT=0
IPRINT=2
TRMS=1.E10
ID(1)=10HCONVERGED
ID(2)=10H
DO 2 I=1,N
2 DX(I)=0.

START ITERATION
DO 100 ITER=1,100

FIND CLOSEST POINTS ON CURVE
5 DO 30 I=1,N
W2(I)=W(I)
30 CONTINUE

18
C FIND CLOSEST POINT GIVEN AN INDIVIDUAL DATA POINT
DO 10 K=1,50
IF(ABS(W(I)) .LT. 1.E-9) GO TO 20
XT=X(I)+DX(I)
CALL FUNC(XT,Y(I)+AL,XDER,D)
DY=Y(I)-Y1(I)
DXT=(-DX(I)+DY*XDER)/(1.+XDER**2)
C
C THE NEXT 3 CARDS PREVENT OVERSHOOT BY DECREASING THE INCREMENT.
C THIS GIVES INCREASED STABILITY AT THE EXPENSE OF INCREASED
C CONVERGENCE TIME.
C
C NEXT 3 CARDS PREVENT OVERSHOOT BY DECREASING THE INCREMENT.
C THIS GIVES INCREASED STABILITY AT THE EXPENSE OF INCREASED
C CONVERGENCE TIME.
C
TT=2.
IF(ABS(DX(I)).GT.1.E-4) TT=ABS(DXT/DX(I))
DX(I)=DX(I)+DXT/(1.+TT**5)
T=ABS(DY)+ABS(DX(I))
IF(ABS(DXT).LT.1.E-4*DX(I)) OR T.LT.1.E-8) GO TO 20
10 CONTINUE
PRINT 1005
1005 FORMAT(* SHORTEST DISTANCE NOT FOUND*)
W2(I)=0.
20 CONTINUE
C
C CLOSEST POINT FOUND. NOW FIND DISTANCE
XI(I)=X(I)+DX(I)
DIS(I)=SQRT(DX(I)**2+DY**2)
IF(DIS(I).LT.1.E-8) DIS(I)=1.E-8
30 CONTINUE
C
C COMPLETE SET OF CLOSEST POINTS AND DISTANCES FOUND
C NOW FIND A NEW SET OF PARAMETERS
DDM=0.
WT=0.
DO 40 I=1,N
DDM=DDM+DIS(I)**2+DI(S(I)**W2(I)
40 WT=WT+W2(I)
RMS=SQRT(DDM/WT)
IF(IPRINT.GE.1)PRINT 1001,RMS,AL(I),I=1,M
1001 FORMAT(* RMS=*G15.7/* PARAMETERS=*SG15.7,(*12x,SG15.7))
44 IF(ABS(TRMS-RMS).LT.ERR*RMS) GO TO 110
TRMS=RMS
CALL FUNC(X1,Y1,N+AL,XDER,D)
DO 45 K=1,M
DO 45 I=1,N
45 D(K,I)=(Y(I)-Y1(I))*D(K,I)/DIS(I)
DO 60 K=1,M
DO 50 J=K,M
A(K,J)=0.
DO 47 I=1,N
47 A(K,J)=A(K,J)+D(K,I)*D(J,I)*W2(I)
50 A(J,K)=A(K,J)
B(K)=0.
DO 60 I=1,M
60 B(K)=B(K)+D(K,I)*DIS(I)*W2(I)
THE NEXT 6 CARDS ARE DERIVED FROM A DAMPING TECHNIQUE THAT INCREASES STABILITY AS MENTIONED IN THE DESCRIPTION. IT WILL INCREASE CONVERGENCE TIME TO SOME EXTENT.

C

T=0
DO 65 I=1,M
65 T=T+B(I)**2
WW=.5*DDM/T
DO 67 I=1,M
67 A(I+1)=A(I+1)+.5/WW

C
C THE FOLLOWING SUBROUTINE CALL SOLVES THE LINEAR SIMULTANEOUS EQUATION GIVEN BY AX=B WITH M unknowns.
C CALL SIMSOL(A,B,M)
DO 70 I=1,M
70 AL(I)=AL(I)+B(I)
100 CONTINUE
ID(1)=10HNOT CONVER
ID(2)=10HGED
110 CONTINUE
IF(IPRINT.GE.1)PRINT 1004,ID,RMS,ERR,ITER
1004 FORMAT(* SUBROUTINE LSD *2A10 /* RMS=*E15.7/* CONVERGENCE C RITERION=*E15.7/* ITERATION COUNT=*I3/) IF(IPRINT.GE.2)PRINT 1000,*(X(I),Y(I),X(I),DIS(I),W(I),I=1,N)
1000 FORMAT(* X,Y,X1,D1*W=*/(SG15.7)) RETURN END

C
C
C
C SUBROUTINE FUNC CALLED BY CURVE FITTING SUBROUTINE
C
C
C
C
C SUBROUTINE FUNC(X,Y,N,AL,XDER,DER)
C
C THIS SUBROUTINE IS CALLED BY LSD AND IS FOR THE LANGMUIR PROBE CURRENT VS VOLTAGE FUNCTION.
C DIMENSION X(50),Y(50),DER(10,50),AL(10)
IF(N*NE*1)GO TO 50
C
C IF N=1, CALCULATE THE FUNCTION AND ITS X-DERIVATIVE AT X(1).
X1=X(1)
Y(1)=X1*(AL(2)+AL(1)*X1)+AL(3)+AL(4)*EXP(AL(5)*X1)
XDER =2*XAL(1)*X1+AL(2)+AL(4)*AL(5)*EXP(AL(5)*X1)
RETURN
50 CONTINUE
APPENDIX B – Concluded

C IF N NOT = 1, CALCULATE THE FUNCTION AND DERIVATIVES WITH RESPECT
C TO ALL PARAMETERS AT POINTS X(I), I=1,N.
    DO 100 I=1,N
      X1=X(I)
      T1=EXP(AL(5)*X1)
      T =AL(4)*T1
      X2=X1*X1
      Y(I)=AL(1)*X2+AL(2)*X1+AL(3)+T
    DER (1*I)=X2
    DER (2*I)=X1
    DER (3*I)=1.
    DER (4*I)=T1
    DER (5*I)=X1*T
100 CONTINUE
    RETURN
END

END
REFERENCES


"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Scientific and technical information generated under a NASA contract or grant and considered an important contribution to existing knowledge.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

TECHNOLOGY UTILIZATION PUBLICATIONS: Information on technology used by NASA that may be of particular interest in commercial and other non-aerospace applications. Publications include Tech Briefs, Technology Utilization Reports and Technology Surveys.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION OFFICE

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Washington, D.C. 20546