APPROXIMATION OF FUNCTIONS BY OPTIMALLY FITTED POLYGONS

By

F. Kiss

Department of Precision Mechanics and Optics, Technical University, Budapest

Received June, 26, 1975

Presented by Prof. Dr. O. Petrik

In programming analog computers, function generators of various working principles are used to produce the non-linear functional relationship between the variables. The so-called diode function generator is frequently employed, approximating the required function section-wise by straight lines. The closeness of approximation depends on the number of straight sections and on the accuracy of fitting the straight lines.

The number of approximating lines is limited. In practice, the course of the lines is determined graphically after having plotted the function (Fig. 1). This method cannot ensure optimum fit.

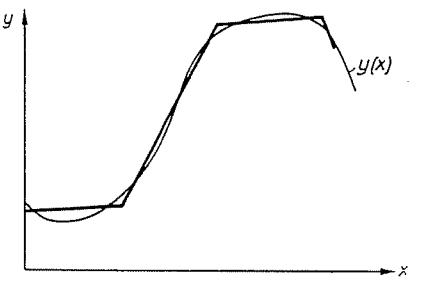

Fig. 1

A similar problem may arise when measurement results are included in an empirical formula, and a function of a given class is to be fitted to a given set of points $(x_i, y_i), (i = 1, 2, \ldots, n)$. The problem is simple to solve if the approximating function is a straight line. Therefore, it is advisable to transfer the points (x_i, y_i), related by a non-linear functional relationship, into positions fitting a straight line, by suitably imaging the plane (Fig. 2). At the end, inverse transformation of the equation of the line fitted to the set of points (u_i, v_i), leads to the equation of the approximating function in the original system of co-ordinates.

The problem becomes more complicated, if the parameters in the function describing the relationship between the variables, change step-like for

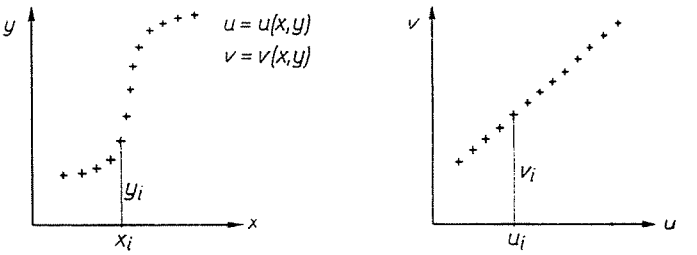

Fig. 2

certain critical values of the variables, i.e., there are discontinuities or break points in the describing function. In such cases the points obtained as the result of the suitable linearizing transformation are fitted not to a single, but to several lines, accordingly, the best approximation function is a polygon.

Formulation of the problem

Considering that the function shown in Fig. 1 is frequently given by tabulated values (e.g. characteristics plotted from measurements), or in the case of a known function, the table of values can be given, therefore, the problem can be formulated in general as follows:

In the closed interval $[x_1, x_n]$ n points $P_i(x_i, y_i)$ are given.

The set of independent variables is ordered: $x_i < x_{i+1}$.

Determine the system of equations for the polygon consisting of a given number l of straight sections, optimally fitting to the given set of points:

$$e_k(x) = m_k x + b_k \qquad k = 1, \ldots, l$$

For determining the numerical values for the parameters in the equations of the approximating lines, the principle of least squares was chosen as optimization condition, since it is suited for preparing the computer algorithm and it is one of the most accurate among the methods employed in practice. Accordingly, parameter values for which the sum of squares of deviations is the minimum, are sought for. In the course of thed eduction the following symbols are used (Fig. 3):

$P_i(x_i, y_i)$	$(i = 1, \ldots, n)$ the given ordered set of points
l	number of approximating straight lines
j, k, s, r	running subscripts for the parameters of the approximating lines
$e_k(x)$	$e_k(x) = m_k \cdot x + b_k$, equation of the k-th approximating line
m_k	slope of the k-th line
$\mathbf{m}$	column vector formed of parameters m_k $(k = 1, \ldots, l)$

b_k axial section of the k-th line on the ordinate

$\mathbf{b}$ column vector formed of parameters b_k $(k = 1, \ldots, l)$

a_k abscissa of the intersection of the k-th and $(k+1)$-th lines (break point of the polygon)

$\mathbf{a}$ column vector formed of parameters a_k $(k = 1, \ldots, l-1)$

α_k subscript of the point having the highest abscissa in the range of interpretation of the k-th line

$q_{k,i}$ $q_{k,i} = e_k(x_i) - y_i$, the deviation of ordinates at x_i

q_k sum of squares of ordinate deviations for the points in the range of interpretation of the k-th line

Q sum of squares of ordinate deviations for the whole range of interpretation.

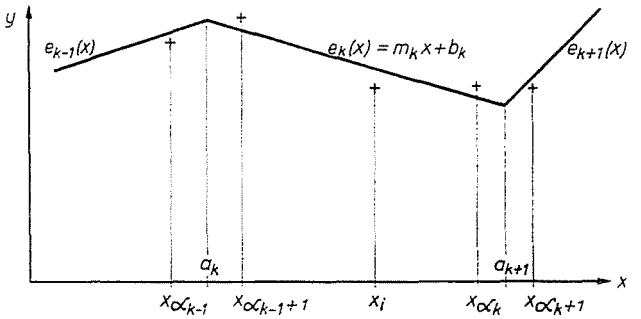

Fig. 3

Considering the ranges of interpretation of the individual line sections, the system of equations of the polygon is found to be

$$e_k(x) = m_k \cdot x + b_k, \quad \text{if} \quad a_{k-1} < x \leq a_k \quad (k = 1, \ldots, l). \tag{1}$$

The lower limit a_0 of the range of interpretation of the first line and the upper limit a_l of the range of interpretation of the l-th line are irrelevant for the solution of the problem. Be $a_0 \leq x_1$ and $a_l \geq x_n$.

The subscript of the point having the hightest abscissa in the range of interpretation of the k-th line is α_k $(x_{\alpha_k} \leq a_k, k = 1, \ldots, l)$. Subscripts α_k are functions of the break points a_k of the polygon (Fig. 3).

$$\alpha_k(a_k) = r, \quad \text{if} \quad x_r \leq a_k < x_{r+1} \quad (k = 1, \ldots, l-1). \tag{2}$$

According to the sense, $\alpha_0 = 0$ and $\alpha_l = n$.

Suppose that $\alpha_{k-1} < i \leq \alpha_k$, i.e. point $P_i(x_i, y_i)$ falls into the range of interpretation of the k-th line. The deviation of ordinates at x_i (Fig. 3) is given by

$$q_{k,i} = m_k \cdot x_i + b_k - y_i. \tag{3}$$

The sum of squares of deviations in the interval (a_{k-1}, a_k) is found to be

$$q_k = \sum_{i=\alpha_{k-1}+1}^{\alpha_k(a_k)} q_{k,i}^2 .$$

For the whole examined range

$$Q = \sum_{k=1}^{l} \sum_{i=\alpha_{k-1}+1}^{\alpha_k(a_k)} (m_k x_i + b_k - y_i)^2 , \tag{4}$$

that is

$$Q = Q(m_1, \ldots, m_l, b_1, \ldots, b_l, a_1, \ldots, a_{l-1})$$

or in vector form

$$Q = Q(\mathbf{m}, \mathbf{b}, \mathbf{a}).$$

The function Q to be minimized is, accordingly, the function of 3. $l-1$ parameters. These, however, are not independent of each other, since connection is established by means of the $l-1$ equations written for the points of intersection of neighbouring lines:

$$e_k(a_k) = m_k \cdot a_k + b_k = m_{k+1} \cdot a_k + b_{k+1} = e_{k+1}(a_k) \qquad (k=1, \ldots, l-1). \tag{5}$$

With the help of Eqs (5), $l-1$ parameters m_k, b_k $(k = 1, \ldots, l)$ can be eliminated from (4). Parameters a_k $(k = 1, \ldots, l-1)$ occur in (4) only in functions of the limits of summation subscripts $\alpha_k(a_k)$, thus the elimination of these does not serve the purpose. In (5) the parameters $b_2, \ldots, b_l$ are rather simple to express:

$$\begin{aligned} b_2 &= b_1 - a_1(m_2 - m_1) \\ &\cdot \\ &\cdot \\ &\cdot \\ b_k &= b_1 - \sum_{j=1}^{k-1} a_j (m_{j+1} - m_j) \quad (k = 2, \ldots, l) \end{aligned} \tag{6}$$

To simplify the formula, introduce parameter $a_0 \equiv 0$ and order (6):

$$b_k = b_1 + \sum_{j=1}^{k-1} m_j(a_j - a_{j-1}) - m_k \cdot a_{k-1} \quad (k=1, \ldots, l)$$

Substituting this into (3):

$$q_{k,i} = m_k (x_i - a_{k-1}) + \sum_{j=1}^{k-1} m_j(a_j - a_{j-1}) + b_1 - y_i \tag{7}$$

Thus

$$Q = Q(m_1, \ldots, m_l, b_1, a_1, \ldots, a_{l-1}).$$

Equations necessary to determine the $2l$ independent parameters are supplied by the system of necessary condition equations of the minimum of function Q:

$$\frac{\partial Q}{\partial m_1} = 0\,; \quad \ldots \quad \frac{\partial Q}{\partial m_l} = 0\,;$$

$$\frac{\partial Q}{\partial b_1} = 0\,; \tag{8}$$

$$\frac{\partial Q}{\partial a_1} = 0\,; \quad \ldots \quad \frac{\partial Q}{\partial a_{l-1}} = 0\,; \tag{9}$$

For writing function (4) and for solving the obtained system of equations, the value of the limits of summation subscripts designated so far symbolically by $\alpha_k(a_k)$, is to be known. As a first approximation, let parameters a_k and thus the optimum values of the limits of summation subscripts be supposed to be known. Thus only the system of equations (8) is to be solved.

Parameters of the best fitted polygon in the case of fixed break points

Let us introduce parameter vector

$$\mathbf{p} = \begin{bmatrix} m_1 \\ \cdot \\ \cdot \\ \cdot \\ m_l \\ b_1 \end{bmatrix}$$

with $l + 1$ dimensions. In the case of fixed break points the sum of squares of deviations depends only on vector $\mathbf{p}$:

$$Q(\mathbf{p}) = \sum_{k=1}^{l} \sum_{i=\alpha_{k-1}+1}^{\alpha_k} q_{k,i}^2(\mathbf{p})$$

The necessary condition of a minimum to exist is:

$$\frac{\partial Q}{\partial p_s} = 2 \cdot \sum_{k=1}^{l} \sum_{i=\alpha_{k-1}+1}^{\alpha_k} q_{k,i} \frac{\partial q_{k,i}}{\partial p_s} = 0 \quad (s = 1, \ldots, l+1). \tag{10}$$

Derivatives in the system of equations:
For $s \leq l$; $p_s \equiv m_s$ and

$$\frac{\partial q_{k,i}}{\partial p_s} = \frac{\partial q_{k,i}}{\partial m_s} = \begin{cases} x_i - a_{s-1} & \text{for } s = k \\ a_s - a_{s-1} & \text{for } s < k \\ 0 & \text{for } s > k. \end{cases} \tag{11}$$

For $s = l + 1$; $p_s \equiv b_1$ and

$$\frac{\partial q_{k,i}}{\partial p_{l+1}} = \frac{\partial q_{k,i}}{\partial b_1} = 1 . \tag{12}$$

In the followings, for the sake of simplicity of writing, the lower limit of subscripts will not be indicated if the upper limit of summation with respect to subscript i is α_k, and the lower limit $\alpha_{k-1} + 1$. Substitute (7) into (10):

$$\sum_{k=1}^{l} \sum_{i}^{\alpha_k} \frac{\partial q_{k,i}}{\partial p_s} \left[m_k (x_i - a_{k-1}) + \sum_{j=1}^{k-1} m_j (a_j - a_{j-1}) + b_1 - y_i \right] = 0 .$$

After multiplication and ordering, we find that

$$\sum_{k=1}^{l} \left[m_k \sum_{i}^{\alpha_k} \frac{\partial q_{k,i}}{\partial p_s} (x_i - a_{k-1}) + \sum_{i}^{\alpha_k} \frac{\partial q_{k,i}}{\partial p_s} \sum_{j=1}^{k-1} m_j (a_j - a_{j-1}) \right] +$$

$$+ b_1 \sum_{k=1}^{l} \sum_{i}^{\alpha_k} \frac{\partial q_{k,i}}{\partial p_s} = \sum_{k=1}^{l} \sum_{i}^{\alpha_k} y_i \frac{\partial q_{k,i}}{\partial p_s} .$$

Identical parameters m_j occur in the equation at various places with different coefficients. By factoring out, we obtain the system of equations

$$\sum_{k=1}^{l} m_k \left[\sum_{i}^{\alpha_k} \frac{\partial q_{k,i}}{\partial p_s} (x_i - a_{k-1}) + (a_k - a_{k-1}) \sum_{j=k+1}^{l} \sum_{i}^{\alpha_j} \frac{\partial q_{j,i}}{\partial p_s} \right] +$$

$$+ b_1 \sum_{k=1}^{l} \sum_{i}^{\alpha_k} \frac{\partial q_{k,i}}{\partial p_s} = \sum_{k=1}^{l} \sum_{i}^{\alpha_k} y_i \frac{\partial q_{k,i}}{\partial p_s} \quad (s = 1, \ldots, l + 1). \tag{13}$$

Incorporating the coefficients into matrix **G**, the free members into vector **c** leads to the matrix form of the system of equations:

$$\mathbf{G}\,\mathbf{p} = \mathbf{c} \tag{14}$$

Taking the values of the derivatives in (11) and (12) into consideration, the elements $g_{s,k}$ of the coefficient matrix **G** can be calculated as follows.

The elements in the main diagonal ($k = s$):

$$g_{s,s} = \begin{cases} \sum_{i}^{\alpha_s} (x_i - a_{s-1})^2 + (a_s - a_{s-1})^2 (n - \alpha_s), & \text{if } s \leq l \\ n, & \text{if } s = l + 1 \end{cases}$$

The elements in the upper triangle ($k > s$):

$$g_{s,k} = \begin{cases} (a_s - a_{s-1}) \left[\sum_{i}^{\alpha_k} (x_i - a_{k-1}) + (a_k - a_{k-1}) (n - \alpha_k) \right], & \text{if } k \leq l \\ \sum_{i}^{\alpha_s} (x_i - a_{s-1}) + (a_s - a_{s-1}) (n - \alpha_s), & \text{if } k = l + 1 \end{cases}$$

The elements in the lower triangle ($k < s$):

$$g_{s,k} = \begin{cases} (a_k - a_{k-1}) \left[\sum_{i}^{\alpha_s} (x_i - a_{s-1}) + (a_s - a_{s-1})(n - \alpha_s) \right], & \text{if } s \leq l \\ \sum_{i}^{\alpha_k} (x_i - a_{k-1}) + (a_k - a_{k-1})(n - \alpha_k), & \text{if } s = l + 1 \end{cases}$$

Since $g_{s,k} = g_{k,s}$, matrix **G** is symmetrical.

The free members are

$$c_s = \begin{cases} \sum_{i}^{\alpha_s} y_i (x_i - a_{s-1}) + (a_s - a_{s-1}) \sum_{i=\alpha_s+1}^{n} y_i & \text{if } s \leq l \\ \sum_{i=1}^{n} y_i & \text{if } s = l + 1 \end{cases}$$

Solving the system of Eqs (14), yields parameters **m**(**a**), **b**(**a**) of the polygon optimally fitted in the case of given break points **a**.

It may occur that the value of some parameters of the approximation lines are previously determined by the problem. For instance, a condition may be that the first straight line passes through the origin ($b_1 = 0$). In such a case, system of Eqs (14) is reduced by taking the fixed parameter values into consideration.

Determination of the optimum values of break points

The estimated values $\mathbf{a}^0$ of the break points and the parameters $\mathbf{m}(\mathbf{a}^0)$, $\mathbf{b}(\mathbf{a}^0)$ obtained from system of Eqs (14) do not satisfy system of Eqs (9) in the general case, hence $\mathbf{a}^0 \neq \mathbf{a}_{opt}$. The optimum vector of the break point $\mathbf{a}_{opt}$ satisfying also system of Eqs (9) can only be calculated by approximation methods. To solve this problem, the so-called gradient method was chosen. With a view on system of Eqs (14), function Q to be minimized depends on the break points alone:

$$Q = Q(\mathbf{a}).$$

To start the calculation, let estimation $\mathbf{a}^0$ of the optimum value of the break points be given. The value of derivatives at $\mathbf{a}^0$ is given by:

$$\left.\frac{\partial Q(\mathbf{a})}{\partial a_s}\right|_{\mathbf{a}=\mathbf{a}^0} = 2 \cdot \sum_{k=1}^{l} \sum_{i}^{\alpha_k} q_{k,i} \left.\frac{\partial q_{k,i}}{\partial a_s}\right|_{\mathbf{a}=\mathbf{a}^0} \quad (s = 1, \ldots, l-1)$$

where, taking (6) into consideration,

$$\frac{\partial q_{k,i}}{\partial a_s} = \frac{\partial e_k(x_i)}{\partial a_s} = \frac{\partial b_k}{\partial a_s} = \begin{cases} -(m_{s+1} - m_s) & \text{for } s < k \\ 0 & \text{otherwise} \end{cases}$$

Therefore

$$\frac{\partial Q}{\partial a_s} = -2 \cdot (m_{s+1} - m_s) \sum_{k=s+1}^{l} \sum_{i}^{\alpha_k} q_{k,i} \,. \tag{15}$$

Apply the derivatives to form unit vector e, with co-ordinates:

$$e_s(\mathbf{a}) = \frac{-\dfrac{\partial Q(\mathbf{a})}{\partial a_s}}{\sqrt{\sum_{j=1}^{l-1} \left[\dfrac{\partial Q(\mathbf{a})}{\partial a_j}\right]^2}} \quad (s = 1, \ldots l-1). \tag{16}$$

In the $l-1$ dimension space formed by vectors a, in the reduced surroundings of point a, vector e is directed towards decreasing values of function $Q(\mathbf{a})$, its direction being contrary to that of vector *grad* $Q(\mathbf{a})$. Accordingly, an approximation a of vector $\mathbf{a}_{opt}$, which is better than $\mathbf{a}^0$, is to be looked for on the line

$$\mathbf{a} = \mathbf{a}^0 + t \cdot \mathbf{e}^0, \tag{17}$$

at positive values of parameter t. To determine the optimum value of parameter t, we have again only approximation methods at our disposal. The most simple, and in our case the most advisable method is to solve the problem by arranging the intervals into boxes. In doing this, however, remind that the value of parameter t cannot be arbitrarily high, since the range of interpretation of function $Q(\mathbf{a})$ is limited. Namely, the values of break points, i.e. the co-ordinates of vector a, must satisfy the system of inequalities

$$x_1 < a_1 < \ldots < a_{l-1} < x_n. \tag{18}$$

The range of interpretation is limited by l planes in the $l-1$ dimension space formed by place vectors a, the plane equations being:

$$\begin{aligned} a_1 &= x_1 \\ &\vdots \\ a_j &= a_{j-1} \qquad (j = 2, \ldots, l-1) \\ &\vdots \\ a_{l-1} &= x_n. \end{aligned} \tag{19}$$

Straight line (17) passes through the limit planes (19). Be d_j $(j = 1, \ldots, l)$ the value of line parameter t at the point of crossing the j-th plane. Among these parameters of the passage point, be d_{min} the lowest positive one. Its value is equal to the distance of the examined point a from the nearest limit plane (in the positive direction of line (17)). Thus, the optimum value of parameter t is likely to be in the range

$$0 \leq t < d_{min}. \tag{20}$$

If for a point

$$\mathbf{a}^1 = \mathbf{a}^0 + t^0\mathbf{e}^0$$

determined by some value t^0 (in the interval (20)), the relationship

$$Q(\mathbf{a}^1) < Q(\mathbf{a}^0)$$

is valid, value t^0 can be regarded as optimum and the approximation can be continued starting from point $\mathbf{a}^1$.

$$\mathbf{a}^2 = \mathbf{a}^1 + t^1 \cdot \mathbf{e}^1 \qquad [Q(\mathbf{a}^2) < Q(\mathbf{a}^1)]$$

$$\vdots$$

$$\mathbf{a}^{r+1} = \mathbf{a}^r + t^r \cdot \mathbf{e}^r \qquad [Q(\mathbf{a}^{r+1}) < Q(\mathbf{a}^r)]$$

Approximation is continued until one or both of the following criteria are satisfied:

1. $$|\mathbf{a}^{r+1} - \mathbf{a}^r| = t^r < \varepsilon,$$

i.e. the distance of subsequent points is smaller than the previously given limit $\varepsilon > 0$.

2. $$\max\left(\left|\frac{\partial Q}{\partial a_1^r}\right|, \ldots, \left|\frac{\partial Q}{\partial a_{l-1}^r}\right|\right) < \eta,$$

i.e. system of Eqs (9) is satisfied by vector $\mathbf{a}^r$, at the required accuracy.

The solution obtained in this way supplies a local minimum of the function of square deviations Q.

Conclusion

This wearisome method is advisably programed for a digital computer. Running time depends on the number and distribution of points in the set determining the function to be approximated, further on the number of approximation lines, but it is essentially determined by the number of iteration steps. It is therefore important to optimally choose the starting value $\mathbf{a}^0$ of the break point vector and the parameter t, further to give reasonable values for limits ε and η. In the case of a well-organized program and a computer of medium capacity, with data $n = 10 \ldots 100$, $l = 5 \ldots 10$, $\varepsilon = 10^{-3} \ldots 10^{-4}$, the running time is expected at $10 \ldots 300$ sec.

Summary

A special case of linearizing functions is the approximation by several straight-line sections. This problem is to be solved e.g. in programming analog computers with a diode function generator. The paper describes a numerical method of determining the equations of the approximating straight lines, based on the principle of least squares, suited for a computer algorithm.

Ferenc Kiss, H-1521 Budapest