ОБЪЕДИНЕННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ

Лаборатория теоретической физики

S.N.Sokolov, I.N.Silin

D-810

DETERMINATION OF THE COORDINATES OF THE MINIMA OF FUNCTIONALS BY THE LINEARIZATION METHOD

Дубна 1962

S.N.Sokolov, I.N.Silin

D-810

DETERMINATION OF THE COORDINATES OF THE MINIMA OF FUNCTIONALS BY THE LINEARIZATION METHOD

Abstract

A method of minimization is suggested applicable to the functionals dependent on the parameters sought for exclusively through the functional argument, and having convergence significantly better than that of the general methods in the common use. The method has proved highly efficient in solving a number of concrete problems.

At present any expressions dependent on the parameters are minimized almost exclusively by the gradient method (by the method of steepest descent)/1,2/.

A natural consequence of the universal character of this method is, first of all, an unsatisfactory rate of convergence, the number of steps necessary for reaching the minimum quickly increasing with the multiplicity of the free parameters which are varied. The same shortcoming is inherent to other general methods, for instance, to the relaxation one/1/.

Meanwhile, the overwhelming majority of the expressions one has to minimize in practice are rather commonplace by their structure. Therefore, it appears reasonable to single out two-three basic types and to work out the corresponding specialized methods of searching for minima which are more effective than the gradient method. Such specialized methods may be useful by themselves or in the combination with a certain encouraging cybernatic procedure, as, for instance, the gradient methods used in the method of ravines/3/.

This paper suggests as some of the specialized methods of locations of minima, the method applicable to the functionals dependent upon the unknown parameters $a = \{ a_1, \dots , a_m \}$ exclusively through the functional argument $y(a, x)$, i.e.,

$$\frac{\partial M}{\partial a_k} = \int \frac{\delta M}{\delta y(a,x)} \frac{\partial y(a,x)}{\partial a_k} dx \qquad (1)$$

(by the variational derivative $\frac{\delta M}{\delta y(x)}$ we mean the kern of the Frechet derivative $M'(y, f) \equiv$ $\equiv \int \frac{\delta M}{\delta y} f \, dx$) .

For instance, $M = \int f [y(a, x)] \, dx$. The variable x may be descrete* and continuous, one-dimensional and many-dimensional. The functional M is assumed to be twice continuously differentiable with respect to its functional argument throughout all the regions where it may be needed in the future.

Formally, the method suggested consists in replacing the exact equations of extremum by a certain system of linear equations, that is why it was called linearization method.

As for the advantages of the method, it should be mentioned that the number of steps necessary for the location of one minimum is not almost growing with the number of the parameters a .

* Here and further it is implied that if the functional is set on a descrete set of points x_ξ , $\xi = 1, \dots n$, then the variational derivatives are replaced by the partial ones, and the integration -by the summation, for instance: $\frac{\partial M}{\partial a_k} = \sum_{\xi=1}^{n} \frac{\partial M}{\partial y(a, x_\xi)} \frac{\partial y(a, x_\xi)}{\partial a_k}$. The number of points x_ξ must not be less than that of the parameters a .

The linearization method has been applied in a number of cases/4,5,6/. The minimum has been found after 5–10 iterations, the number of parameters varying from 2 up to 16.

The functional M may have many minima of different types . However, far from all of them correspond necessarily to the solution of the original problem. To reject false solutions it is necessary to investigate the stability of the positions of the minima with respect to the shifts of the outer (not entering the functional argument $y(a, x)$) parameters. Let us emphasize that such an investigation does not reduce to the study of the shape of the pits found. In the linearization method the type of the minimum can be determined without any additional calculation, since there is a close correspondence between the type of a minimum and the character of the searching process in its vicinity.

Sec. 1. Step Formula

In the linearization method the functional $M\{y\}$ is approximated by the quadratic functional

$$M\{y\} \simeq \tfrac{1}{2} \int \frac{\delta^2 M}{\delta y(z)\, \delta y(x)} [y(a,z) - y(a^\circ, z)][y(a,x) - y(a^\circ,x)]\, dz dx + \int \frac{\delta M}{\delta y(x)} [y(a,x) - y(a^\circ,x)] dx + \text{const}, \tag{1.1}$$

while the dependence $y(a)$ - by the linear dependence*

$$y(a) \simeq y(a^\circ, x) + \sum_{k=1}^{m} \frac{\partial y}{\partial a_k} \Delta a_k . \tag{1.2}$$

The functional neighbourhood of the initial approximation where approximation (1.1) holds is supposed to be great enough that the minimum M belongs to this neighbourhood **. As for (1.2), no such assumption is made.

---

* If applied to the least squares method the expediency of approximation (1.2) has been pointed out in different books, e.g.,/7,8/.

** By the functional neighbourhood of the minimum of M we mean the region in which $\int f(z) \frac{\delta^2 M}{\delta y(z)\, \delta y(x)} f(x)\, dz\, dx > 0$ for an arbitrary non-zero function f .

For estimating the direction and the distance to the minimum M we get a following system of linear equations

$$\frac{\partial M}{\partial a_k} + \sum_{l=1}^{m} \Delta a_l \int \frac{\partial y(z)}{\partial a_l} \frac{\delta^2 M}{\delta y(z)\, \delta y(x)} \frac{\partial y(x)}{\partial a_k}\, dz\, dx = 0 . \qquad (1.3)$$

As will be seen in the following, approximations (1.1) and (1.2) lead to the rejection of not only the terms of the second and higher order with respect to Δa, but of some terms of the first order. Therefore, such a procedure should be somewhat elucidated.

Find the vector $\Delta a = \{\Delta a_1, \ldots, \Delta a_m\}$, which in an infinitesimal vicinity of the minimum of M would touch the minimum with its free end. Expanding the derivatives $\frac{\partial M}{\partial a_k}$ in powers of the vector Δa

$$\frac{\partial M(a + \Delta a)}{\partial a_k} = \frac{\partial M(a)}{\partial a_k} + \sum_{l=1}^{m} \frac{\partial^2 M(a)}{\partial a_l\, \partial a_k} \Delta a_l + \ldots \qquad (1.4)$$

and using the extremum condition

$$\frac{\partial M(a + \Delta a)}{\partial a_k} = 0, \qquad (1.5)$$

we get, after neglecting the higher powers of Δa_i, a system of linear equations

$$\frac{\partial M(a)}{\partial a_k} + \sum_{l=1}^{m} \Delta a_l \frac{\partial^2 M(a)}{\partial a_l\, \partial a_k} = 0 , \qquad k = 1, \ldots, m , \qquad (1.6)$$

or, calculating explicitly the second derivatives

$$\frac{\partial M}{\partial a_k} + \sum_{l=1}^{m} \Delta a_l \int \frac{\partial y(z)}{\partial a_l} \frac{\delta^2 M}{\delta y(z)\, \delta y(x)} \frac{\partial y(x)}{\partial a_k}\, dz\, dx + \qquad (1.7)$$

$$+ \sum_{l=1}^{m} \Delta a_l \int \frac{\delta M}{\delta y(x)} \frac{\partial^2 y(x)}{\partial a_l\, \partial a_k}\, dx = 0 .$$

The vector Δa whose components satisfy system (1.7) provides, in a small vicinity of the extremum, both for the direction to the extremum and the distance to it.

Far off the extremum the vector Δa indicates roughly the direction to the nearest extremum regardless of its type, so that if the initial approximation a^o happened to be, e.g., close to the saddle point, then the motion along Δa will leads to the saddle point. This makes it difficult to use eq. (1.7) for searching for the minima since the functions minimized may have many different competing extrema near the initial approximation.

By comparing (1.3) and (1.7) we see, that Eqs. (1.3) have been obtained from (1.7) by means of a sort of linearization - the rejection of the term

$$\sum_{i=1}^{m} \Delta a_i \int \frac{\delta M}{\delta y(x)} \frac{\partial^2 y(x)}{\partial a_i \; \partial a_k} dx \equiv \sum_{i=1}^{m} \Delta a_i \; Q_{ik} , \qquad (1.8)$$

which takes into account the non-linearity of $y(a)$, and is not, generally speaking, small compared with the term conserved in (1.3)

$$\sum_{i=1}^{m} \Delta a_i \int \frac{\partial y(z)}{\partial a_i} \frac{\delta^2 M}{\delta y(z) \; \delta y(x)} \frac{\partial y(x)}{\partial a_k} dz \, dx \equiv \sum_{i=1}^{m} \Delta a_i \; G_{ik} . \qquad (1.9)$$

The rejection of Q_{ik} leads to some advantages of system (1.3) over (1.7). In particular, the vector Δa determined by (1.3) always indicates the direction in which M is decreasing, and the competition of the extrema of different types ceases.

The system of equations which is practically used in the linearization method differs from (1.3). Here, instead of a complete step Δa , only a certain its fraction $\overline{\Delta a} = \lambda \Delta a$, $0<\lambda \leq 1$ is taken, which is determined from the condition of the optimal convergence of the minimization process (see Sec. 3). Qualitatively λ may be estimated as the maximum $\lambda \leq 1$, for which the linear approximation

$$y (a + \lambda \Delta a) - y (a) \simeq \overline{\Delta y} \equiv \sum_{k=1}^{m} \lambda \Delta a_k \frac{\partial y}{\partial a_k} \qquad (1.10)$$

is still roughly correct.

By substituting $\Delta a = \lambda \Delta a$ and (1.9) into (1.3), we get

$$\lambda \frac{\partial M}{\partial a_k} + \sum_{i=1}^{m} \overline{\Delta a_i} \; G_{ik} = 0, \qquad (1.11)$$

from here we find the step in the space of the parameters

$$\overline{\Delta a_i} = -\lambda \sum_{k=1}^{m} G^{-1}_{ik} \frac{\partial M}{\partial a_k} \qquad (1.12)$$

and the functional step

$$\overline{\Delta y(z)} = -\lambda \sum_{i,k=1}^{m} \frac{\partial y(z)}{\partial a_i} G^{-1}_{ik} \frac{\partial M}{\partial a_k} . \qquad (1.13)$$

Sec. 2. A Simple Example

Let $y(A)$ be an analytical function of the parameter $A = a_1 + ia_2$ and $M\{y\} = |y|^2$. Then the minima of M are equal to zero and correspond to the roots of the equation $y(A)=0$. By substituting $M = |y|^2$ into (1.9), (1.12) and assuming $\lambda = 1$, we obtain

$$G = 2|y'_A| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Delta A = A^{(n+1)} - A^{(n)} = -y/y'_A . \qquad (2.1)$$

Evidently, we are led to the Newton method for finding the complex roots of the equation* $y(A)=0$.

It should be noted that the linearization method is, generally speaking, not equivalent to the Newton method. In particular, if we are going to seek for the solution of the extremum equations using the Newton method we do not get system (1.3), but (1.7) whose shortcomings were already discussed .

Sec. 3. Some Properties of the Step Formula and the Choice of λ.

The functional step $\overline{\Delta y(z)}$ is invariant with respect to any (non linear inclusive) substitution of the parameters

* As far as formula (2.1) is concerned, one can turn to paper /9/. Note, that in case of nonanalytical functions $y(A)$ the matrix G is no longer diagonal and formulas (2.1) become more complicated.

$$a_k \rightarrow b_k(a_1, \dots, a_m). \tag{3.1}$$

Exp. (1.13) for $\overline{\Delta y(z)}$ may be interpreted as a certain gradient, namely, as a gradient M in the space $\frac{\partial y}{\partial a}$ with the metric tensor $\frac{\delta^2 M}{\delta y(z)\, \delta y(x)}$ (i.e. , with the scalar product $(f_1, f_2) = \int f_1(z) \frac{\delta^2 M}{\delta y(z)\ \delta y(x)} f_2(x)\, dz\ dx$). Indeed for any $dy = \sum_{p=1}^{m} \frac{\partial y}{\partial a_p} da_p$ from (1.13) with $\lambda = -1$, we have

$$(dy, \overline{\Delta y}|_{\lambda=-1}) = \sum_{p,k,l=1}^{m} d\, a_p\, G_{pk}\, G_{kl}^{-1} \frac{\partial M}{\partial a_l} = d\, M, \tag{3.2}$$

i.e., the condition which defines the gradient is fulfilled. Similarly, exp . (1.12) for $\overline{\Delta a}$ with $\lambda = -1$ is a gradient M in the space a with the metric* tensor G_{ik}.

The proof that by using (1.12), (1.13) we are approaching the minimum M is not essentially different from that employed in the conventional gradient method. Indeed, for small $\lambda > 0$ the increment M as a result of the step $\overline{\Delta y}$ will be necessarily negative, if only the metric G is positive definite. Evidently, the latter circumstance always takes place if the initial approximation $y(a^{\circ}, x)$ is situated in the functional neighbourhood of the minimum M. Now we shall discuss some practical aspects of the method. If $y(a^{\circ}, x)$ gets outside the functional neigbourhood of the minimum, the metrics G_{ik} may be provisionally replaced for the metrics

$$\bar{G}_{ik} = \int \left| \frac{\delta^2 M}{\delta y(z)\ \delta y(x)} \right| dz\ \frac{\partial y(x)}{\partial a_i}\ \frac{\partial y(x)}{\partial a_k}\ dx,$$

which is everywhere positive definite.

For small λ the increment of the functional $\Delta M = M\{ y(a + \overline{\Delta a}) \} - M\{ y(a) \}$ is getting larger with λ , but in a further increase of λ the effect of the terms rejected in approximating

* The metric G_{ik} is transformed in the substitution (8.1) so that the gradient (1.12), in contrast to the usual gradient $\nabla i = \frac{\partial}{\partial a_i}$ is transformed in the linear substitution of the parameters like the radius vector a .

(1.10) may become so great that ΔM will change its sign. The maximum step ℓ_i in each parameter a_i for which the non-linear part of the increment y cannot affect the sign of ΔM, depends upon the magnitude and the structure of the matrices G and Q. For a considerable number of the parameters $\boldsymbol{a}$ it is much more complicated to calculate the matrix Q than to make some extra steps, then it is more advantageous to find the optimal step by a trial and error method. The magnitudes of ℓ_i are changing more slowly than λ, and may be revised rarer. Therefore, it is more convenient to look for the best ℓ_i, and after that to calculate λ by the formula

$$\lambda = \frac{1}{\max\left\{1, \frac{\Delta a_i}{l_i}\right\}} . \qquad (3.4)$$

Further, the quantities ℓ_i will be assumed chosen in the following way. The original ℓ_i are taken somewhat overestimated so that the nonlinear part of the increment $y(a_i \pm \ell_i) - y(a_i)$ would be, on the average, equal to the linear part. If, as a result of the step $\overline{\Delta a}$, the value of the functional M has increased, all ℓ_i are made two times smaller and the step is repeated. If, as a result of the two preceding steps, the functional was decreasing, during the preparation for the next step, those ℓ_i for which $\Delta a_i > \ell_i$, are doubled.

If near the minimum

$$| Q_{ik} | << \sqrt{G_{ii}\, G_{kk}} \qquad (3.5)$$

for all i,k, then $\lambda \to 1$, the iteration process becomes close to the Newton one, and the convergence in the end of the process will be fast.* In this case the corrections of the coordinates of the minimum will go practically according to the one-dimensional Newton formula (2.1), where the most nonlinear of the parameters a_i will play the role of the parameter A. The case of the violation of condition (3.5) will be discussed in the next Section .

Let us mention two main reasons which favour the fulfillment of condition (3.5) in practice. First, the functions $\frac{\partial^2 y(x)}{\partial a_i \partial a_k}$ are usually well approximated by a linear combination of the derivatives $\frac{\partial y}{\partial a_i}$ and in integrating with the derivatives $\frac{\delta M}{\delta y(x)}$ yield (near the minimum) zero owing to the extremum equations $\frac{\partial M}{\partial a_i} = 0$. Second, the function $\frac{\delta M}{\delta y(x)}$ near the minimum is often appraaching ze-

* Concerning the convergence of the approximate Newton interations see /2/, Chapter XVIII, Sec.2.

ro because the absolute minimum of the functional M is close to the minimum for the given family of the functions $y(a. x)$. The latter circumstance takes place, e.g., for the least squares method when $M = \sum_{\xi} \{ y(a, x_{\xi}) - t_{\xi} \}^{2} w_{\xi}$.

It should be noted that in both aforementioned cases condition (3.5) has a tendency to be fulfilled the better, the larger the number of the parameters a and the richer the family of $y(a, x)$. This, in particular, accounts for the fact that in the linearization method the number of steps necessary for reaching the minimum does not practically increase with increasing the number of the parameters a .

Sec. 4. The Stability of the Minima

Let the functional M , besides the parameters $\boldsymbol{a}$ which are in the minimization, depends also on certain parameters $t = (t_1, \ldots, t_{\mu}, \ldots)$. We restrict ourselves to the case when the parameters t do not enter the functional argument $y(a, x)$ at all

$$\frac{\partial\, y\,(a, x)}{\partial\, t_{\mu}} \equiv 0, \tag{4.1}$$

and the functional M depends upon them only explicitly $M = M\{ t;\, y\,(a, x\,) \}$. We shall take interest in the dependence of the minimum of M position on the displacements of the parameters t .

Physical examples of such a problem are quite numerous.

The parameters t may be experimental magnitudes which are known with a limited accuracy. What will be the displacement of minimum M if the parameters are moved their standard errors aside? As another example we mention different corrections and higher terms of expansion in series which as small quantities were neglected and put to zero when the expression for the functional M was written. Are the results sensitive to these corrections?

Take the linear term in the expansion of M (in the neighbourhood of the minimum) in a Taylor series in the increments of the parameters a_k and t_{μ}

$$M \simeq M_{\min} + \sum_{\mu} \frac{\partial M}{\partial t_{\mu}} \Lambda t_{\mu} + \sum_{k=1}^{m} \frac{\partial M}{\partial a_k} \Lambda a_k \tag{4.2}$$

and substitute this expansion into the extremun equations $\frac{\partial M}{\partial a_i} = 0$. We get the system of equations

$$\sum_{\mu} \frac{\partial^2 M}{\partial t_\mu \, \partial a_i} \Delta t_\mu + \sum_{k=1}^{m} \frac{\partial^2 M}{\partial a_k \, \partial a_i} \Delta a_k = 0, \quad i=1 , \dots ,m, \qquad (4.3)$$

which establishes the relationship between the displacement of the minimum position Δa and the increment of the parameters t.

The qualitative proper ties of system (4.3) are as follows. If system (4.3) is soluable with respect to Δa_k and the determinant of the matrix

$$\tilde{G}_{ki} = \frac{\partial^2 M}{\partial a_k \, \partial a_i} \qquad (4.4)$$

is not small, then insignificant displacements of the parameters t will give rise to small movements of the minimum Δa, i.e., the minimum will be stable.

When in the neighbourhood of the minimum det $\tilde{G}$ is small, then the elements of the inverse matrix $\tilde{G}^{-1}$ turn out to be great, and even insignificant displacements of the parameters t may cause considerable shifts in the minimum position. We shall call such minima relatively unstable; in more detail this case is considered in Sec. 5.

A special case takes place if certain derivatives $\frac{\partial y(a,x)}{\partial a_k}$ (or their linear combinations) vanish in the minimum. Such minima we shall call degenerated.

The abundance of the degenerated minima is typical for functionals (1) which do not depend explicitly on a. Indeed, let $\frac{\partial y(a,x)}{\partial a_j} = 0$, $\frac{\partial^2 M}{\partial a_j^2} > 0$ for some $a_j = \bar{a}_j$, (4.5)

where the derivative equals to zero identically with respect to x, and $\bar{a}_j$ may be a function of the other parameters $a_{k \neq j}$. Let us put $a_j = \bar{a}_j$. According to (1), this entails the fulfillment of the extremum equation $\frac{\partial M}{\partial a_j} = 0$. By minimizing M over the parameters $a_{k \neq j}$, we can secure that the remaining extremum equations $\frac{\partial y}{\partial a_k}$ would be fulfilled. Thus, each vanishing of one of the derivatives $\frac{\partial y}{\partial a_k}$ may lead to the appearance of the degenerated minimum of the functional M

The degenerated minima are 'superstable' in certain directions in the parametric space a *, what

* The stability with respect to other directions can be investigated in a usual manner if we cross out the equations which do not contain the increments Δt_μ from system (4.8).

means that no change of t can shift the minimum along the mentioned directions. For instance, in case (4.5), the minimum may move only on the surface set by the equation $a_j = \bar{a}_j$. Note, that the statements just made essentially rely upon limitation (4.1).

A set of the minima which the functional has depends on how its functional argument is parametrized. Let the functional M have a certain spectrum of the minima $M^{(i)}_{min} = M\{ y^{(i)}(x) \}$.
If the family $y(a, x)$ is parametrized so that the functional argument may take the same value $y^{(i)}(x)$ for some different values of the parameters a , then M as the function of a-s will have a number of identical copies of this minimum with the precisely coinciding, depths, and the corresponding spectral value $M^{(i)}_{min}$ will be multiple.

Introduce a new parametrization $b_k = b_k(a_1, \dots, a_m)$ of the same family. Evidently, any minimum $M\{ y^{(i)}(x) \}$ of the functional $M\{ y(a, x) \}$ in which the Jacobian $\frac{\partial(b_1,\dots,b_m)}{\partial(a_1,\dots,a_m)}$ is different from zero must be, as well, present (a copy at least) in the spectrum of $M\{ y(b, x)\}$, since the changes in the spectrum of the minima may be related only with the zeros of the Jacobians
$\frac{\partial(b)}{\partial(a)}$ and $\frac{\partial(a)}{\partial(b)}$.

In changing the parameters t the undegenerated minimum is moving in the m - dimensional region, and Jacobians $\frac{\partial(b)}{\partial(a)}$, $\frac{\partial(a)}{\partial(b)}$ may vanish only in the sub-region of fewer dimensions. Therefore, the undegenerated minimum cannot be created or annihilated by the substitution of the parameters, although the number of its copies may change. It follows herefrom that if there exists such a parametrization of the family $y(a, x)$ for which M has the only non-degenerated minimum $M^{(0)}_{min}$, then the spectrum of the minima $M\{ y(a, x) \}$ begins with $M^{(0)}_{min}$, and all $M_{min} > M^{(0)}_{min}$ are degenerated.

When the functional M is only an auxilliary magnitude and only those values of the parameters a for which it is minimal are of immediate interest, the degenerated minima correspond to the false solutions of the problem. Indeed, the degenerated minimum can be easily created artificially for any value $\bar{a}_k$ of any parameter a_k without changing essentially the functional M itself, but having changed only formally the way of the parametrization of its functional argument $y(a, x)$. Let us substitute, e.g., the parameters b for a by the formula $b_k^2 + b_k^3 = (a_k - \bar{a}_k)\ \mathrm{sign}\ \frac{\partial M}{\partial a_k}$.
Now $\frac{\partial y}{\partial b_k} = (2b + 3b^2)\frac{\partial y}{\partial a_k} = 0$, $\frac{\partial^2 M}{\partial b_k^2} > 0$, at $b_k = 0$, and the functional M has the m-multiple degenerated minimum at the point $a_k = \bar{a}_k$ we have chosen . By making similar procedure in an inverse order, by a formal substitution of the parameters a_k it is possible for the functional M to get rid of any degenerated minimum found (at the same time new degenerated minima may appear in other places).

In the degenerated minima the iteration process is convergent owing to the decrease of the numbers ℓ_i et each step. In case of such a 'forced' convergence at the end of the process $\lambda \to 0$ what allows to distinguish easily the degenerated minima from the usual ones. Note, that there may exist minima in which λ goes on varying up to the very end of the process. For instance, these are the minima which are close to the degenerated and pass into the latter ones in changing slightly the parameters t (in such minima condition (3.5) is violated).

Sec. 5. Unstability Problem

A considerable unstability of the minima is accounted, as a rule, for an incorrect formulation of the minimization problem and cannot be overcome formally. Regardless of the minimization method applied, the unstability leads inevitebly to the slowing down of the searching for the minimum and to the loss of accuracy in the calculations because of the accumulation of the rounding off errors. The technical reasons for which these difficulties appear may be different. In particular, in the linearization method λ is becoming small and the relative error in the matrix elements of G^{-1} is increasing as well as the error in the quantities which are calculated in terms of this matrix.

The unstability of the minimum is usually felt long before it was found. A convenient indicator may be dimensionless quantity

$$\rho = \frac{G_{11}\, G_{22} \cdots \; G_{mm}}{\det G} \,, \qquad \rho \geq 1, \tag{5.1}$$

which is close to unity in the region of the relatively stable minimum and much larger than unity in the region of the relatively unstable one. More detailed information may be provided by the correlation factors*

$$R_k = G_{kk}\; G^{-1}_{kk} \,, \qquad R_k \geq 1, \tag{5.2}$$

which behave like ρ and indicate the unstability with respect to the parameter a_k . The quantities ρ and R_k are closely connected: if the parameter a_k is fixed, the quantity ρ decreases R_k times.

The correlation factors ought to be known for checking the accuracy of the calculations. It may be

* The correlation factors are closely connected with the so-called multiple coefficients of correlation /10/ $\rho_k\,(1, \ldots, k-1, k+1, \ldots m)$, namely: $R_k^{-1} = 1 - \rho^2_{k(\ldots)}$.

shown that the relative accuracy of the k-th diagonal element of the matrix G^{-1} is not less than R_k times worse than the relative accuracy of the same element of the matrix G . Let, e.g., the matrix G is known with 9 significant figures. Then, if at least one of R_k exceeds 10^{+8}, then in the matrix G^{-1} not a single figure can be guaranteed correct no matter how G^{-1} was calculated.

ρ and R_k have a simple geometric sense. If the functions $\frac{\partial y(x)}{\partial a_k}$ are considered to be vectors in the space with the metric tensor $\frac{\delta^2 M}{\delta y(z)\ \delta y(x)}$ (cf. Sec. 3), then ρ^{-1} is equal to the square of the volume of the parallelepiped with the unit ribs constructed on these vectors, and $R_k^{-1} = \sin^2 \psi$, where ψ is the angle which the vector $\frac{\partial y(x)}{\partial a_k}$ makes with the plane in which the remaining vectors are lying.

A typical reason for the unstability is an incorrect choice of the family of the functions $y\,(a,x)$ which is a functional argument. This case can be easily identified as long as the addition of each new parameter a_k decreases the minimum M slightly but increases ρ sharply. The stability is re-established if for $y(a,x)$ one takes the family which is more to the point when the nature of the problem is concerned.

Let., e.g., $y(a,x)$ is chosen in the form $y = \sum\limits_k a_k x^{k-1}$ and the low values of M should be expected if . $y(x)$ has the shape of a curve drawn in Fig. 1.

Clearly, to describe such a curve by a polynomial is difficult, and the coefficients a_k will be determined from the condition $M = min$ rather bad. In this case it is much better to put, for instance, $y = (b_2 + b_4\, x + \dots)(1 + b_1 x + b_3 x^2 + \dots)^{-1}$

It may happen that the family $y\,(a,x)$ is chosen in accordance with the nature of the problem, but ρ is, nevertheless, great. This means that the parametrization of the family $y\,(a,x)$ is no success, and some nonlinear substitution of the parameters* with large R is required. In such a substitution one should tend for each of the functions $\frac{\partial y(x)}{\partial a_k}$ to have the maximum at the point where the remaining functions are small.

In the worst case, if the parameters a_k introduced are just the quantities, for the sake of which the minimization is made, one has to give up determining the parameters with large R by fixing some of these parameters.

* Firstly, the linear substitution does not avoid the loss in accuracy, second, to know its coefficients exactly enough it is necessary, at first, to find the minimum.

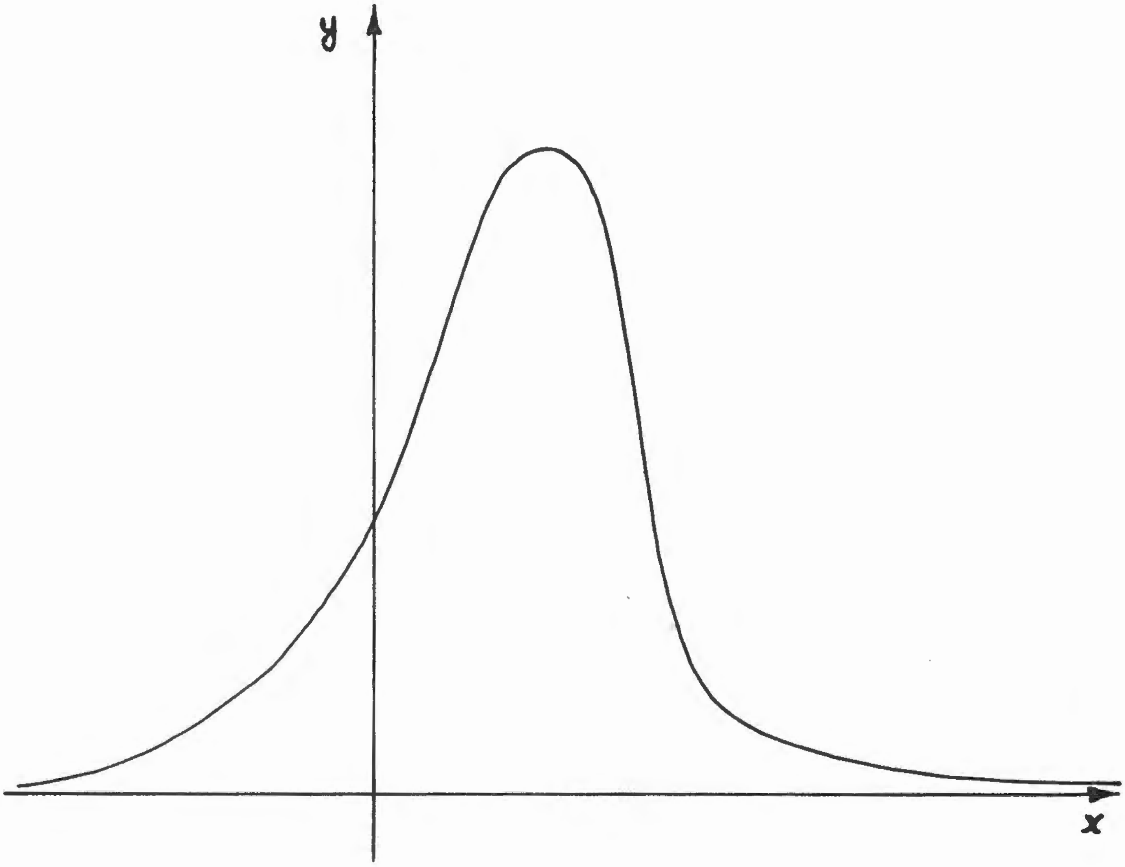

Fig. 1.

6. Some Applications. The quadratic Functional

The quadratic functional

$$M = \sum_{\xi} [\,y\,(a, x_{\xi}) - t_{\xi}\,]^{2} \quad w_{\xi}\,, \qquad (6.1)$$

vhose minimization is required in the least squares method /7,10,11/ is an ideal object for applying the linearization method. Indeed, for functional (6.1) a small functional neighbourhood of the minimum coincides with the whole functional space, and the matrix Q in the minimum is small. The latter circumstance is connected with the fact that in the minimum the difference $y\,(x_{\xi}) - t_{\xi}$ entering

$$Q_{ik} = 2 \sum_{\xi} \frac{\partial^{2} y\,(x_{\xi})}{\partial\, a_{i}\ \partial\, a_{k}} \quad [\,y\,(x_{\xi}) - t_{\xi}\,] \quad w_{\xi} \qquad (6.2)$$

is small.

Besides, if the functional argument $y\,(a, x_{\xi})$ allows the linear parametrization, then M has the only undegenerated minimum*.

Now we shall calculate the error matrix of the parameters

$$\sigma^{2}_{ik} \equiv \overline{\Delta\, a_{i}\ \Delta\, a_{k}} \qquad (6.3)$$

(the upper line means the averaging), assuming, as is usually done in the least squares method, that the experimental magnitudes ot t_{ξ} are independent

$$\overline{\Delta\, t_{\xi} \Delta\, t_{\eta}} = 0 \qquad \text{at} \qquad \xi \neq \eta \qquad (6.4)$$

and have the variances equal to the inverse weights

$$\overline{\Delta\, t^{2}_{\xi}} = w^{-1}_{\xi}\,. \qquad (6.5)$$

Substituting (4.3), (4.4) into (6.3) and taking into account (6.4), (6.5), we have

* Practically all the functionals $M = -2 \ln L$ where L is the function of likelihood have the same properties.

$$\sigma_{ik}^2 = \overline{\left(\sum_{j=1}^{m} \sum_{\xi} \tilde{G}_{ij}^{-1} \frac{\partial^2 M}{\partial a_j \partial t_\xi} \Lambda t_\xi \right) \left(\sum_{l=1}^{m} \sum_{\eta} \tilde{G}_{kl}^{-1} \frac{\partial^2 M}{\partial a_l \partial t_\eta} \Lambda t_\eta \right)} = \tag{6.6}$$

$$= \sum_{j,l=1}^{m} \tilde{G}_{ij}^{1} \tilde{G}_{kl}^{-1} 2 G_{jl} .$$

If Q is small, it is possible to put $\tilde{G}^{-1} \simeq G^{-1}$, then *

$$\sigma_{ik}^2 \simeq 2 \sum_{j,l=1}^{m} G_{ij}^{-1} G_{kl}^{-1} G_{jl} = 2 G_{ik}^{-1} . \tag{6.7}$$

In the degenerated minimum where $G^{-1} \to \infty$, the approximate expression (6.7) is not applicable. It can be seen from the exact expression (6.6), that, as it should be expected, the parameters with respect to which the minimum is degenerated have the variances σ_{jj}^2 equal to zero. With the aid of the error matrix σ_{ik}^2 it is easy to calculbte, e.g., the variance of the functional argument (the square of the corridor of errors)

$$\overline{[\Delta y (a, x)]^2} \equiv \sigma^2(x) = \sum_{i,k=1}^{m} \frac{\partial y (x)}{\partial a_i} \sigma_{ik}^2 \frac{\partial y (x)}{\partial a_k} . \tag{6.8}$$

The variances of the parameters a_k are simply connected with the correlation factors R_k . According to (5.2) and (6.7)

$$\overline{\Delta a_i^2} = 2 (G_{ii})^{-1} R_i , \tag{6.9}$$

i.e., the variance of any parameter can be decreased as much as R_i times if the other parameters are fixed.

Let the real solutions of the system of the non-linear equations

$$f_k (a_1 , \dots , a_m) = 0 , \quad k = 1 , \dots , m \tag{6.10}$$

have to be found.

Put $y_\xi = f_\xi$, $t_\xi = 0$ and introduce the weights in an arbitrary fashion, then the solution of (6.10) reduces to the minimization of functional (6.1). In a similar manner one can look for complex solutions of system (6.10) putting $y_\xi = Re\, f_\xi$, $y_{\xi+m} = Im\, f_\xi$, and taking $a_\xi = Re\, a_\xi$, $a_{\xi+m} = Im\, a_\xi$ as independent parameters (cf. with (2.1)). By an appropriate choice

* In order to avoid coefficient 2 in (6.7), a matrix two times less than G is often introduced instead of G.

of the weights w_ξ it is possible to change to a certain extent the magnitude of the correlation factors.

In this particular case, when the number of parameters coincides with that of points on which the functional is defined, the application of the linearization method to the minimization of functional (6.1) with $\lambda = 1$ is the same as the application of the Newton method just to system (6.10). However, due to the choice of λ , the linearization method provides for the convergence in a wider class of the cases and has a more convenient control system. In case of cumbersome systems (6.10), the minimum M may turn out to be relatively unstable with respect to the variation of the quantities t_ξ from zero, what leads to technical difficulties. Evidently, the unstability of such a kind is by no means connected with the nature of the problem and is rather formal. Therefore, the unstability can be always avoided by a certain non-linear transformation of system* (6.10).

In the degenerated minima of the functional $M = \sum_\xi y^2 w_\xi$ do not become equal to zero, so that such minima are not the solutions of system (6.10). If real solutions are looked for,then some degenerated minima may point out that near them there may be present a pair of complex solutions with a small imaginary part. If the complex solutions of system (6.10) are sought, then the degenerated minima may be found only at the points when the functions f_k are not analytical with respect to one or several parameters a_k . Indeed, if at a certain point $\frac{\partial f_k}{\partial a_p} = 0$ and the functions f_k are analytical with respect to a_p , then $| f_k |$ is necessarily decreasing in one of the directions, and the searching for the minimum will not stop at this point .

The authors express their gratitude to N.P. Klepikov, J.A. Smorodinski, E.P. Zhydkov, N.N.Govorun, Yu.M. Kazarinov, R.M. Dzhabarh Zade and G.P. Ososkov for valuable remarks.

*See the footnote on page 14.

References

1. Н.С.Березин, Н.П.Жидков. Методы вычислений гл. У1, У11, Физматгиз, Москва, 1960.

2. Л.В.Канторович, Г.П.Акилов. Функциональный анализ в нормированных пространствах. Физматгиз, Москва, 1959.

3. И.М.Гельфанд, М.Л.Цетлин. Принципы нелокального поиска в системах автоматической оптимизации. ДАН, 1961, т.137, № 2.

4. Н.П.Клепиков, В.А.Мещеряков, С.Н.Соколов. Анализ экспериментальных данных по полным сечениям взаимодействия π -мезонов с протонами. Препринт ОИЯИ Д-584, Дубна, 1960.

5. Н.С.Амаглобели, Ю.М.Казаринов, С.Н.Соколов, И.Н.Силин. Определение константы π -мезон-нуклонного взаимодействия по дифференциальным сечениям упругого np -рассеяния. ЖЭТФ, т.39, вып. 4(10), 1960.

6. Лю Юйань, Н.И.Пятов, В.Г.Соловьев, И.Н.Силин, В.И.Фурман. О свойствах ряда сильно-деформированных ядер. ЖЭТФ, 40, 1501 (1961).

7. A.Hald. Statistical Theory with Engineering Applications. New York, 1952.

8. И.Н.Бронштейн и К.А.Семендяев. Справочник по математике для инженеров и учащихся втузов. Физматгиз, Москва, 1959, стр. 570.

9. В.В.Воеводин. Применение метода спуска для определения всех корней алгебраического многочлена. ЖВММФ т.1, № 2. 1961 .

10. H.Cramér. Mathematical Methods of Statistics. Princeton, 1946.

11. Н.П.Клепиков, С.Н. Соколов. Анализ экспериментальных данных методом максимума правдоподобия. Препринт ОИЯИ Р-235, Дубна, 1958.

12. Л.А.Люстерник, В.И.Соболев. Элементы функционального анализа. § 41 ГИТТЛ, 1951.

Received by Publishing Department
on October 12, 1961.