ОБЪЕДИНЕННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ

Лаборатория теоретической физики

S.N. Sokolov

E-1067

CONTINUOUS PLANNING OF REGRESSION EXPERIMENTS

Теория вероятностей и ее применение 1963, т 8, вып 1, с 95-101

Дубна 1962 год

S.N. Sokolov

E-1067

CONTINUOUS PLANNING OF REGRESSION EXPERIMENTS

Дубна 1962 год

Abstract

In the first part of this work containing a new version of the results obtained in 1960 [1,11] a local-optimal continuous planning of regression experiments is suggested which is convenient from the practical point of view.

In the second part it is shown that in case of the stable condition of experiment the continuous planning is asymptotically global-optimal.

Аннотация

В первой части этой работы, содержащей новую формулировку результатов, полученных в 1960 году [1,11], предлагается локально-оптимальное непрерывное планирование регрессионных экспериментов, удобное с практической точки зрения.

Во второй части показывается, что, в случае устойчивых условий эксперимента, непрерывное планирование является асимптотически глобально-оптимальным.

Настоящий препринт издаётся только на английском языке.

where c, f are the known functions and $\theta = \theta_1, \dots, \theta_m$ is the parametrical point under estimation, whose number of dimensions, m, is known beforehand. Let, if for measuring the curve $\eta(x)$ at the point x the time t is spent, then the result of measurement $y = y(\xi)$ be a random (independent of previous) sampling from the population $y(x, t, \xi)$ with properties

$$\langle y(x,t,\xi)\rangle = \int y(x,t,\xi)\, p(\xi)\, d\xi = \eta(x),$$
$$\langle [y(x) - \eta(x)]^2 \rangle = [\int_0^t \lambda(x,\tau)\, dt(\tau)]^{-1} \equiv w^{-1}, \tag{1}$$

where $\lambda(x,\tau)$ is the measurement efficiency, τ is the moment when the measurement is made, ξ is a random parameter of the sample. As is known, the best (having the smallest variance) estimate for θ linear in $y_i = y(x_i)$ is $\hat{\theta} = M^{-1}Y$, where M is the Fischer information matrix and

$$M_{\alpha\beta} = \sum_i f_\alpha(x_i)\, f_\beta(x_i)\, w_i; \qquad Y_\alpha = \sum_i f_\alpha(x_i)[y_i - c(x_i)]\, w_i.$$

The estimate $\hat{\theta}$ is unbiased $\langle\hat{\theta}\rangle = \theta$ and has the variance – covariance matrix $D_{\alpha\beta}(\hat{\theta}) = \langle(\hat{\theta}_\alpha - \theta_\alpha)(\hat{\theta}_\beta - \theta_\beta)\rangle = (M^{-1})_{\alpha\beta}$. In combining independent experiments their information matrices are summed up. Therefore the experiment and the corresponding Fischer matrix will be denoted by the same letter, the matrices M assumed to be positive-definite and ΔM, positive semi-definite.

Let the parameters θ be separated into the group of parameters $\phi = \theta_1, \dots, \theta_r$ we are interested in, and the group of nuisance ones $\omega = \theta_{r+1}, \dots, \theta_m$ the measurement of which is not our aim. We shall assume that of two competing experiments M_1, M_2 the experiment M_2 contains more information about ϕ if $\Delta q = q(\phi; M_2) - q(\phi; M_1) > 0$, where

$$\Delta q = -\ln |D_2(\hat{\phi})\, D_1^{-1}(\hat{\phi})|,$$
$$q(\phi; M_1) = -\ln |k \cdot D_1(\hat{\phi})|, \tag{2}$$

and where D_1, D_2 are the corresponding $r \times r$ submatrices of the matrices M_1^{-1}, M_2^{-1} and the constant $k \neq 0$ is chosen by an arbitrary way ($k = 1$, if there is no special reservation). In particular, if $M_2 = M_1 + \Delta M$ we shall speak that the experiment ΔM gives the information $\Delta q = q(M_2) - q(M_1)$. The main reasons why the experiments can be compared in terms of the information (2) are given in paper of Stone */8/. If consideration is restricted to the linear estimates only then these reasons remain also valid in the case of the formal application of (2) for the non-Gaussian distributions of probability $y(\xi)\, p(\xi)\, d\xi$.

Suppose that at the time moment τ the efficiency $\lambda(x,\tau)$ is known from the analysis of the experimental conditions or the measurements already carried out. We call the experiment $dM_{\alpha\beta} = \sum_i f_\alpha(x_i)\, f_\beta(x_i)\, \lambda(x_i, \tau)\, dt(x_i)$ a local-optimal one if it gives the maximum of $dq(\phi) = \sum_i [\partial q(\phi)/\partial t(x_i)]\, dt(x_i)$ under the condition $\sum_i dt(x_i) = d\tau$; $dt(x_i) \geq 0$. Then the suggested procedure of the continuous local-optimal planning consists in making at each time moment τ the measurement at the point x where the rate of accumulation of information $\dot{q}[\phi, x, M(\tau)] = \partial q[\phi, M(\tau)]/\partial t(x) = \lambda(x,\tau)\partial q[\phi, M(\tau)]/\partial w(x)$ is maximal. From the point of view of calculation (see 3) the continuous planning reduces to the calculation of the function $\dot{q}[\phi, x, M(\tau)]$ and the choice of the point (one of the points) x for which $\dot{q}(x) = \max_x \dot{q}(x)$.

2. Optimality and variance

We consider the experiment

$$M = \begin{pmatrix} {}^0M & {}^1M \\ {}^1M' & {}^2M \end{pmatrix}; \quad M^{-1} = D(\hat{\theta}) = \begin{pmatrix} {}^0D & {}^1D \\ {}^1D' & {}^2D \end{pmatrix}, \tag{4}$$

* Stone uses the formula $\Delta q = \frac{1}{2} \ln |M_2 M_1^{-1}|$. The complication of expressions by the factor $\frac{1}{2}$ seems quite superfluous and we do not introduce this factor.

where ${}^{0}M$, ${}^{\bullet}D$ are the $r \times r$ submatrices related to parameters $\phi = \theta_1, \dots \theta_r$ and prime (') means transposed. Calculating the variances of the estimates

$$\hat{\eta} = c + f'\hat{\theta}\ ; \quad {}^2\hat{\eta} = c + {}^0f'\phi + {}^2f'\hat{\omega} \equiv c + \sum_{\alpha=1}^{r} f_\alpha\theta_\alpha + \sum_{\alpha=r+1}^{m} f_\alpha\hat{\theta}_\alpha\ ,$$

we get

$$D[\hat{\eta}(x),\ \hat{\eta}(y)] = f'(x)\,D(\hat{\theta})\,f(y)\ ; \quad D[{}^2\hat{\eta}(x),\ {}^2\hat{\eta}(y)] = {}^2f'(x)\,{}^2M^{-1}\,{}^2f(y).$$

When $x = y$ the second argument in D will be omitted $D[\eta(x)] \equiv D[\eta(x), \eta(x)]$. Let us introduce the matrix

$$^2\Delta[D(\hat{\theta})] = \begin{pmatrix} {}^0D & {}^1D \\ {}^1D' & {}^2D - {}^2M^{-1} \end{pmatrix}. \tag{5}$$

We call the quantity

$$D(\hat{\eta}, \phi) = D(\hat{\eta}) - D({}^2\eta) = f'\ {}^2\Delta[D(\hat{\theta})]\ f \tag{6}$$

a subvariance of the estimate $\hat{\eta}$ with respect to parameters ϕ. In a particular case when ${}^1M = 0$ and the estimates $\hat{\phi}$ and $\hat{\omega}$ are independent the subvariance coincides with the partial variance $D(\hat{\eta}, \phi) = {}^0f'\,{}^0D\,{}^0f$. If the nuisance parameters are absent ($r = m$) the subvariance coincides with the total variance $D(\hat{\eta}, \phi) = D(\hat{\eta})$. In its meaning the subvariance with respect to ϕ is a fraction of the variance of the estimate $\hat{\eta}$ which is due to the ignorance of the exact values of the parameters ϕ.

LEMMA I Let M be an arbitrary matrix having an inverse one $D = M^{-1}$ and allowing the division into the submatrices similar to (4) so that ${}^2M$ and ${}^0D$ have also their inverse ones. Then, the identity

$$1 - M\begin{pmatrix} 0 & 0 \\ 0 & {}^2M^{-1} \end{pmatrix} - \begin{pmatrix} {}^0D^{-1} & 0 \\ 0 & 0 \end{pmatrix} M^{-1} = 0 \tag{7}$$

holds. To prove this,it is sufficient to calculate explicitly the left-hand sides of (7) and of the obvious equality

$$\left[\begin{pmatrix} {}^0D^{-1} & 0 \\ 0 & 1 \end{pmatrix} M^{-1}\right]\left[M\begin{pmatrix} 1 & 0 \\ 0 & {}^2M^{-1} \end{pmatrix}\right] - \begin{pmatrix} {}^0D^{-1} & 0 \\ 0 & {}^2M^{-1} \end{pmatrix} = 0$$

and compare the results.

Multiplying (7) from the left by M^{-1} we obtain the identity

$$\begin{pmatrix} {}^0D & {}^1D \\ {}^1D' & {}^2D - {}^2M^{-1} \end{pmatrix} \equiv D\begin{pmatrix} {}^0D^{-1} & 0 \\ 0 & 0 \end{pmatrix} D\ , \tag{8}$$

from which it follows that ${}^2\Delta(D)$ is positive-semi-definite and the subvariance $D(\hat{\eta}, \phi)$ is non-negative.

Now we consider the increase of q in measuring η at the point x, i.e. for

$$M_{\alpha\beta}(\tau + d\tau) = M_{\alpha\beta}(\tau) + f_\alpha(x)\,f_\beta(x)\,\lambda(x, \tau)\,d\tau.$$

THEOREM I. The rate of accumulation of information about some group of parameters equals the product of the efficiency and the corresponding subvariance

$$\dot{q}(\phi, x) = \lambda(x, \tau)\,D(\hat{\eta}(x), \phi). \tag{9}$$

Proof. We calculate explicitly the derivative

$$\dot{q} = -\partial \ln |{}^{2}D| / \partial t = \lambda f' M^{-1} \begin{pmatrix} {}^{0}D^{-1} & 0 \\ 0 & 0 \end{pmatrix} M^{-1} f$$

and use the identity (8). The theorem is proved.

Obviously, $\dot{q}(\phi, x) \geq 0$, from where it follows that theorem V of paper/8/ can be extended to the case of the presence of nuisance parameters.

By differentiating both sides of (9), it is easy to obtain expressions for higher derivatives of the information , e.g.

$$\frac{\partial^2 q(\phi)}{\partial t(x)\, \partial t(y)} = -\lambda(x)\lambda(y)\{(D[\hat{\eta}(x), \hat{\eta}(y)])^2 - (D[{}^{2}\hat{\eta}(x), {}^{2}\hat{\eta}(y)])^2\} \quad (10)$$

(in differentiating it was taken into account that $\lambda(x, r)$ does not depend explicitly on t). For $r = m$ or $x = y$ the derivative $\ddot{q}(\phi, x, y) \leq 0$. For $r < m$ and $x \neq y$ the function $\ddot{q}(\phi, x, y)$ may have any sign.

Stone has shown/8/ that the information is invariant under the linear substitution of parameters. From theorem I and the invariance of variances $D(\hat{\eta})$, $D({}^{2}\hat{\eta})$ it is obvious that the presence of nuisance parameters does not violate the validity of this result and the rate of accumulation of information is invariant under a linear substitution of parameters inside the groups ϕ and ω . Thus, the definition of optimality assumed by us implies that the specification of the estimate of any linear combination of parameters ϕ (i.e. the decrease of the variance of this estimate some given n times), is considered as an equally useful result.

In the process of the experiment planned continuously the optimal point $\bar{x}$ is moving both due to changes of the subvariance $D(\hat{\eta}(x), \phi)$ and possible changes in the efficiency $\lambda(x, r)$. We show that for $r > 1$ some displacements of the point $\bar{x}$ occur even under stable conditions of the experiment $\lambda(x, r) = \lambda(x, r^0) = \lambda(x)$.

Let λ be independent of r and $r > 1$. We choose arbitrarily certain points x_k ; $k = k_1, \ldots k_p$ and from the moment r^0 on we shall measure at these points only.

THEOREM II. If p , the number of points x_k , is less than r , after a sufficiently long measurement at each of these points none of them will be optimal.

Proof. For constant λ the weight $w = (\lambda t)^{-1}$. From (1), (6) and (9) it follows that

$$(\partial q / \partial t_i)\, t_i \leq 1 . \quad (11)$$

Considering the case when all t_i 's increase proportionally, i.e. $M(r + dr) = (1 + dr/r) M(r)$, it is easy to prove the identity

$$\sum_i (\partial q / \partial t_i) t_i = r . \quad (12)$$

By singling out explicitly those t_k which are increasing and using (11), (12) we get

$$(\max_{i \neq k} \frac{\partial q}{\partial t_i}) \sum_i t_i^0 \geq \sum_{i \neq k} \frac{\partial q}{\partial t_i} t_i^0 = r - \sum_{k=1}^{p} \frac{\partial q}{\partial t_k} t_k \geq r - p \geq 1 . \quad (13)$$

Inserting in (11) $t_k > \sum_{i \neq k} t_i^0$ and combining with (13) we have

$$\partial q / \partial t_k < (\sum_{i \neq k} t_i^0)^{-1} \leq \max_{i \neq k} \partial q / \partial t_i ,$$

what proves the theorem.

The requirement of the independence of λ on r is not necessary. Repeating the calculations (11)-(13) for

$\dot{\lambda}_\tau \neq 0$ it is not difficult to show that for theorem II to be valid it is sufficient to fulfil the condition $[\tau\lambda(x_i, \tau)\lambda^{-1}(x_k, \tau)] \to \infty$ at $\tau \to \infty$ and for any x_i, x_k, i.e. a weak dependence of λ on τ is unessential.

For $p \geq m$ theorem II is certainly not true/5/. If $m > p \geq r$ then an analog of theorem II can be obtained under some restrictions on the choice of ϕ and the function $\lambda(x)$ which we shall not consider. In a practically important case $r = 1$, $m \geq r$ such an analysis can be made by means of the graphical technics developed in/7/.

Now we consider a global-optimal experiment. Let $\lambda(x)$, $M(\tau^0)$ and τ be given and the times $t(x_i)$; $\Sigma t(x_i) = \tau - \tau^0 = T$ can be distributed in any way.

THEOREM III. For the experiment $M(\tau) = M(\tau^0) + \Delta M(T)$ to be optimal, i.e. giving maximum $\Delta q = q[\phi, M(\tau)] - q[\phi, M(\tau^0)]$, it is necessary that all the points of measurements x_i are coinciding with points x_k where the function $\lambda(x)D[\hat{\eta}(x), \phi; M(\tau)]$ reaches the absolute maximum.

Proof. Assume the contrary. Then the experiment which differs from the given optimal one only by that a fraction of the measurement time dt_i is transferred from the point where $\lambda D < \max_x \lambda D$ to the point $\bar{x}_j$ in virtue of theorem I will contain $dq = [\max \lambda D - \lambda(x_i) D(\hat{\eta}(x_i), \phi)]dt_i$ more information than the optimal one, what is impossible. The theorem is proved.

From (12) it is easy to see that if $M(\tau^0) = cM(\tau)$, $c \geq 0$ and $M(\tau)$ is optimal, then $\max_x \lambda D(\hat{\eta}, \phi) = r\tau^{-1}$ and if $M(\tau^0) \neq cM(\tau)$ then $\max_x \lambda D(\hat{\eta}, \phi) > r\tau^{-1}$.

3. Prognosis of a Local-Optimal Experiment

If $\lambda(x, \tau)$ is a known function of τ then no unknown quantities enter (9) and the process of the continuous planning can be calculated any time ahead.

In planning, continuous in the exact sense of this word, the function $\dot{q}(\phi, x)$ reaches the absolute maximum, generally speaking, at several points $\bar{x}_k(\tau)$. We impose on the distribution of time $dt(\bar{x}_k) = c_k d\tau$, $\sum_{k=1}^{p} c_k = 1$, $c_k > 0$ a condition that all maxima of the function $\dot{q}[\phi, x; M(\tau)]$ where measurements are made, are decreasing with the same speed

$$\sum_{k=1}^{p} \frac{\partial \dot{q}(\phi, \bar{x}_i)}{\partial t(\bar{x}_k)} c_k = \sum_{k=1}^{p} \frac{\partial \dot{q}(\phi, \bar{x}_j)}{\partial t(\bar{x}_k)} c_k, \qquad (14)$$

what is equivalent to the requirement of the smoothness of M as a function of τ. The condition (14) jointly with the equations

$$M_{\alpha\beta}(\tau) = M_{\alpha\beta}(\tau^0) + \int_{\tau^0}^{\tau} \sum_{k=1}^{p(\tau)} f_\alpha(\bar{x}_k) f_\beta(\bar{x}_k) \lambda(\bar{x}_k, \tau) c_k(\tau)\, d\tau$$
$$\dot{q}[\phi, \bar{x}_k(\tau); M(\tau)] = \max_x \dot{q}[\phi, x; M(\tau)] \equiv \max_x \lambda(x, \tau) D(\hat{\eta}(x), \phi; \tau) \qquad (15)$$

unambiguously define the growth of $M(\tau)$ in the continuous planning. For practical applications it is more convenient to break up the time into the steps $\Delta\tau$, small compared to τ, and instead of (14), (15) use the reccurent equations

$$M_{\alpha\beta}(\tau + \Delta\tau) = M_{\alpha\beta}(\tau) + f_\alpha(\bar{x}) f_\beta(\bar{x}) \lambda(\bar{x}, \tau)\, \Delta\tau,$$
$$\lambda(\bar{x}, \tau) D(\hat{\eta}(\bar{x}), \phi; \tau) = \max_x \lambda(x, \tau) D(\hat{\eta}(x), \phi; \tau). \qquad (16)$$

In (16) it is sufficient to consider only the single optimal point $\bar{x} = \bar{x}(\tau)$, which will imply any point of equally optimal ones, if there are many, and omitting the condition (14). However, for finite $\Delta\tau$ an exact coincidence of the heights of several maxima of the function λD is an unlikely event.

If λ depends weakly on τ then, according to theorem II, after some steps of continuous motion* along one of the curves $\bar{x}_k(\tau)$ the point $\bar{x}(\tau)$, determined by the system (16) must jump to other curve $\bar{x}_{j\neq k}(\tau)$ ensuring in this way some distribution of time between all $\bar{x}_k(\tau)$ (In the limit $\Delta\tau \to 0$ coinciding with the distribution $c_k(\tau)$).
If $\lambda(x,\tau) = \lambda(x,\tau^0)$ then at $\tau \to \infty$ the allocation of points $\bar{x}_k(\tau)$ and the distribution of time between them, $c(\tau)$, 'will forget' the initial experiment $M(\tau^0)$ and become similar to those of the global- optimal experiment when the initial one is absent. Therefore the recurrent equations (16) may be applied for calculating the global- optimal allocation of measurements, for a large number of parameters this may turn out to require less computational work than the variation of the information q with respect to 2 m variables $\bar{x}_k$ and c_k.

Example: $\eta = \phi + (x-3)\omega$, $\lambda = (1+x^2)^{-2}$, $\tau^0 = 6$.

The initial measurements $t_1^0 = t_2^0 = 3$ are made at points

$$x_1^0 = 1,5 \;; \quad x_2^0 = 2 \quad \text{and} \quad M(\tau^0) = \begin{pmatrix} 0.404 & -0.546 \\ -0.546 & 0.760 \end{pmatrix}.$$

Further the experiment $M_c(\tau^0,\tau)$ is planned according to (16) with a step $\Delta\tau = 0.1$.

A global- optimal (for $M(\tau^0)=0$) allocation $c_g(x_i)$ can be easily found for this example. Denoting $\lambda^{-1/2} = h$ and noticing that $\phi = \eta(3)$, by the formulae of static planning in Ch. V of /7/ or /11/ we get:

$$\frac{d}{dx}h(x) = \frac{h(x)}{x} \;; \quad x_{1,2} = \pm 1 \;; \quad \frac{t_1}{t_2} = \left| \frac{h(x_1)(3-x_2)}{h(x_2)(3-x_1)} \right| = 2\;;$$

so that

$$c_g(1) = 2/3 \;; \quad c_g(-1) = 1/3 \;; \quad M_g(1) = \frac{1}{12}\begin{pmatrix} 3 & -8 \\ -8 & 24 \end{pmatrix}.$$

Fig. 1 A gives the position of the optimal point during the first thirty steps. Starting with the ninth step when the function $D(\hat{\eta}, \phi)$ aquires the second maximum at the point $\bar{x} \approx 1$ the process goes into the asymptotics and the time allocation coincides with the global-optimal one $c_g(x)$ with the accuracy up to quantities of the order of $\Delta\tau(\tau-\tau^0)^{-1}$.

An inverse variance $[D(\hat{\phi})]^{-1} = \exp\, q(\phi)$ is plotted in Fig. 1B. The broken line I corresponds to the continuous planning $M = M_c(\tau^0,\tau)$. The curve II plotted for the comparison corresponds to the combination of the initial and the optimal experiment $M = M(\tau^0) + M_g(\tau-\tau^0)$, the straight line III, to the experiment $M = (\tau/\tau^0)M(\tau^0)$ continued similarly to the initial one.

PART II. Global Properties of the Continuous Planning

The investigation of the global properties of the local-optimal planning is interesting not only from the practical point of view

* More exactly, a motion which turns into a continuous one with the infinite breaking up of the given several steps.

but also because the further development of the theory of the planning of indirect experiments seems to choose the direction of the synthesis of the global and local approaches.

In 4 the general properties of the information as a function of the experiment M and the conditions of the global optimality are established. The global and the local plannings are compared in 5. In 6 the connection between the optimality and the condition $\min_x \max D(\hat{\eta}(x), \phi)$, discovered by Kiefer[/12/], is discussed.

4. Information from the Mixture of Experiments and the Optimality Condition

The presence of nuisance parameters $\omega = \theta_{r+1}, \ldots, \theta_m$ makes the comparison of experiments difficult. This difficulty can be partially avoided by introducing the notion of reduced matrices. Let

$$A = \begin{pmatrix} {}^0A & {}^1A \\ {}^1A' & {}^2A \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} {}^0B & {}^1B \\ {}^1B' & {}^2B \end{pmatrix}, \quad a = \begin{pmatrix} {}^0B^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \quad {}^0\Delta(A) = A \begin{pmatrix} 0 & 0 \\ 0 & {}^2A^{-1} \end{pmatrix} A \qquad (17)$$

where the division into the submatrices is performed in a way similar to (4). According to the identity (8), $A = a + {}^0\Delta(A)$. We call ${}^0R(A) = {}^0B^{-1}$ a reduced (to the parameters ϕ) matrix A, and ${}^0\Delta(A)$, a remainder of the reduction (e.g. the matrix ${}^2\Delta(D)$ (5) is a remainder of the reduction of the variance–covariance matrix to the parameters ω). The reduction procedure ${}^0R$ is chosen in such a way so that a reduced experiment ${}^0R(M)$ and a complete experiment M contain the same information $q(\phi)$. Besides, the procedure has a following property.

THEOREM IY. The sum of complete experiments $M = M_1 + M_2$ contains not less information than that of reduced experiments $R = {}^0R(M_0) + {}^0R(M_2)$, i.e. $q(\phi; M) \geq q(\phi; R)$. Proof. According to the definition, $M = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix} + {}^0\Delta(M_1) + {}^0\Delta(M_2)$. Let $c = \begin{pmatrix} {}^0c & 0 \\ {}^1c & {}^2c \end{pmatrix}$, $|{}^0c| = 1$ be the matrix transforming $\Delta M = {}^0\Delta(M_1) + {}^0\Delta(M_2)$ to a block form: $c'\Delta M c = \begin{pmatrix} {}^0G & 0 \\ 0 & {}^2G \end{pmatrix}$. Then $c'Mc = \begin{pmatrix} {}^0c' R\, {}^0c + {}^0G & 0 \\ 0 & {}^2G \end{pmatrix}$. Hence $q(M) \equiv -\ln|{}^0(M^{-1})| = \ln|{}^0c' R\, {}^0c + {}^0G|$. From the definition of the operation ${}^0\Delta$ (17) it is obvious that ΔM and, consequently, ${}^0G$ are positive-semi-definite matrices. Then, according to theorem Y of the work of Stone [8]

$$\ln|{}^0c' R\, {}^0c + {}^0G| \geq \ln|{}^0c' R\, {}^0c| = q(\phi; R).$$

The theorem is proved.

Let us now establish how the information changes in replacing the sum of similar experiments by the sum of different ones. In other words, we reveal one of the basic entropic properties of the definition of information (2). Consider two pairs of experiments M_1, M_2 and $\tilde{M}_1$, $\tilde{M}_2$ in which corresponding experiments contain equal information and experiments in the second pair are similar:

$$q(\phi; M_1) = q(\phi; \tilde{M}_1) = q_1; \quad q(\phi; M_2) = q(\phi; \tilde{M}_2) = q_2; \quad \tilde{M}_2 = k\tilde{M}_1. \qquad (18)$$

THEOREM Y. If experiments M_1, M_2, $\tilde{M}_1$, $\tilde{M}_2$ satisfy (18), then the mixture of different experiments $M_\rho = \rho M_1 + (1-\rho)M_2$, $1 \geq \rho \geq 0$, contains not less information than the same mixture of similar experiments $\tilde{M}_\rho = \rho\tilde{M}_1 + (1-\rho)\tilde{M}_2$, i.e.

$$q(\phi; M_\rho) \geq q(\phi; \tilde{M}_\rho) = q_1 + r\ln[\rho + (1-\rho)k]. \qquad (19)$$

Proof. Consider the combination of reduced experiments $R_\rho = \rho\, {}^0R(M_1) + (1-\rho)\, {}^0R(M_2)$. Transform $R_1 = {}^0R(M_1)$ and $R_2 = {}^0R(M_2)$ simultaneously to the diagonal form $R^* = c' R c$, $|c| = 1$, and make use of the elementary inequality $\prod_{a=1}^{r} (a_a + b_a) \geq [(\Pi a_a)^{1/r} + (\Pi b_a)^{1/r}]^r$ which is valid for any $a, b \geq 0$. We get

$$q(R_\rho) = q(R^*_\rho) = \ln \prod_{a=1}^{r} [\rho (R^*_1)_{aa} + (1-\rho)(R^*_2)_{aa}] \geq$$

$$\geq \ln [\rho |R^*_1|^{1/r} + (1-\rho)|R^*_2|^{1/r}]^r = q_1 + \ln [\rho + (1-\rho) k]^r .$$

According to theorem IV $q(\phi; M_\rho) \geq q(\phi; R_\rho)$. The theorem is proved.

It is easy to show that for $1 > \rho > 0$ and ${}^0R(M\ ) \neq k\, {}^0R(M_2)$ a strict inequality

$$q(\phi; R_\rho) > q(\phi; \tilde{M}_\rho), \quad q(\phi; M_\rho) > q(\phi; \tilde{M}_\rho) \tag{20}$$

takes place.

Now we consider the mixture of experiments with different information.

THEOREM VI. The information is a concave function of experiments:

$$q(\phi; \rho M_1 + (1-\rho) M_2) \geq \rho q(\phi; M_1) + (1-\rho) q(\phi; M_2). \tag{21}$$

Proof. Make use of the inequality (19) and the elementary inequality $\rho a + (1-\rho) b \geq a^\rho b^{1-\rho}$ which holds for any $a, b \geq 0$. We obtain

$$q(M_\rho) \geq q(\tilde{M}_\rho) = q_1 + r \ln [\rho + (1-\rho) k] \geq q_1 + r \ln k^{1-\rho} =$$

$$= \rho q_1 + (1-\rho)(q_1 + r \ln k) = \rho q_1 + (1-\rho) q_2 .$$

The theorem is proved.

Let us introduce the notion of the weighted optimality. Let the initial experiment $M(\tau^0)$ and the efficiency $\lambda(x, \tau)$ be given and the time $T = \tau^1 - \tau^0$ can be allocated in any way $dt(x_i(\tau)) = c_i(\tau)\, d\tau$, $\Sigma c_i = 1$.* We call the experiment $M_u(\tau^0, \tau)$ optimal with the weight $u(\tau) \geq 0$ and the priming (initial) experiment $M(\tau^0)$ if

$$q_u = \int_{\tau^0}^{\tau^1} q[\phi; M_u(\tau^0, \tau)]\, u(\tau)\, d\tau = \tag{22}$$

$$= \max_{x(\tau)} \int_{\tau^0}^{\tau^1} q[\phi; M(\tau^0) + \int_{\tau^0}^{\tau} \mu(x(t), t)\, dt]\, u(\tau)\, d\tau ,$$

where $\mu_{\alpha\beta} = f_\alpha(x) f_\beta(x) \lambda(x, \tau)$. In the particular case when $u(\tau) = \delta(\tau^0 - \tau)$ and $\int_{\tau^0}^{\tau^1} F(\tau) u(\tau) d\tau = F(\tau^0)$ the weighted optimality coincides with the local one; if $u(\tau) = \delta(\tau^1 - \tau)$ and $\int_{\tau^0}^{\tau^1} F(\tau) u(\tau) d\tau = F(\tau^1)$ the weighted optimality coincides with the global one. An experiment planned continuously in a local-optimal way from the moment τ^0, will be denoted by $M_c(\tau^0, \tau)$; experiment global-optimal

---

* Writing so we mean that the time T is distributed with the density $V(x, \tau) = \Sigma c_i(\tau) \delta(x - x_i(\tau))$, where δ is a Dirac function defined so that $\int F(x) \delta(x-a)\, dx = F(a)$ for any continuous $F(x)$. For the sake of brevity we shall also write $\int F(x(\tau))\, d\tau$ meaning $\int F(x) V(x, \tau)\, dx\, d\tau$.

to the moment r^1, the priming being $M(r^0)$, by $M_g(r^0, r^1; r)$; a global-optimal one for $M(r^0)=0$ and for $\lambda(x,r)=\lambda(x)$, by $M_g(r)$.

THEOREM YII. For the experiment $M(r)$ to be optimal with the weight $u(r)$ and the priming $M(r^0)$ it is necessary and sufficient that at any moment r the measurement be made at the point $\bar{x}(r)$, where the function $\lambda(x,r)D_u(x,r)=\lambda(x,r)\int^{r^1} D[\hat{\eta}(x), \phi; M(t)]\cdot dt$ reaches the absolute maximum.

Proof of necessity. Consider M, D^r and q as functionals of $x(r)$ and write the condition (22) in the form

$$\delta \int_{r^0}^{r^1} q([\bar{x}], r)\, u(r)\, dr = \int_{r^0}^{r^1} \{ q[\bar{x}+\delta\bar{x}] - q[\bar{x}]\}\, u(r)\, dr \leq 0 .$$

Taking into account that $\delta M(r) = \int_{r^0}^{r} \delta\mu(t)\, dt$ and calculating explicitly the variation, according to theorem I we have

$$\delta q([\bar{x}], r) = \int_{r^0}^{r} \delta \{ D[\hat{\eta}(\bar{x}(t)), \phi; M(r)]\, \lambda(\bar{x}(t), t)\}\, dt ,$$

where the sign of the variation δ affects $\bar{x}(t)$ only, whence, integrating by parts, we obtain

$$\delta q_u = \int_{r^0}^{r^1} u(r) \int_{r^0}^{r} \delta \{ D[\hat{\eta}(\bar{x}(t)), \phi; r]\, \lambda(\bar{x}(t), t)\}\, dt\, dr = \tag{23}$$

$$= \int_{r^0}^{r^1} \delta \{ \lambda(x(t), t) \int_{t}^{r^1} D[\hat{\eta}(\bar{x}(t), \phi, r]\, u(r)\, dr \}\, dt = \int_{r^0}^{r^1} \delta \{ \lambda(\bar{x}, t) D_u(\bar{x}, t)\}\, dt \leq 0 .$$

Because of the arbitrariness of $\delta\bar{x}(t)$ the inequality (23) can be guaranteed only if $\delta\lambda D_u \leq 0$ and $\bar{x}$ is a point where the function $\lambda(x,t)D_u(x,t)$ is absolutely maximal.

Proof of sufficiency. Suppose there exist two experiments $M_a(r)$, $M_b(r)$ in which measurements are made at points $x_a(r)$, $x_b(r)$ where the corresponding functions $\lambda[D_u]_a$, $\lambda[D_u]_b$ reach the absolute maximum and which contain the weighted information $(q_u)_b < (q_u)_a = \max_{x(r)} q_u[x]$.

Make the mixture $M_{d\rho} = d\rho\, M_a(r) + (1-d\rho)\, M_b(r) = M(r^0) + d\rho \int_{r^0}^{r} \mu(\bar{x}_a)\, dr + (1-d\rho)\int_{r_0}^{r}\mu(\bar{x}_b)\, dr$. The experiment $M_{d\rho}$ differs from M_b only by a fraction of the time of measurement transferred from the point $\bar{x}_b$ where $\lambda[D_u]_b$ is maximal to the point $\bar{x}_a$ where $\lambda[D_u]_b$, generally speaking, is not maximal. Hence

$$q_u(M_{d\rho}) \leq (q_u)_b . \tag{24}$$

On the other hand, using theorem VI and positivity of $u(r)$ we have

$$q_u(M_{d\rho}) \geq (q_u)_b + [(q_u)_a - (q_u)_b]\, d\rho > (q_u)_b . \tag{25}$$

The incompatibility of (24) and (25) proves the theorem.

5. Optimal properties of the continuous planning

From theorem YII it is obvious that if the subvariance $D(\hat{\eta}(x), \phi; r)$ as a function of x changes its form with time then the continuously planned experiment $M_c(r^0, r)$ coincides neither with the global-optimal experiment M_g nor with the weighted-optimal one M_u. We shall try to estimate how great this difference is and whether the losses of information at $r\to\infty$ are essential when M_g is replaced by M_c.

We first consider a case when the efficiency λ does not depend on time, $\lambda = \lambda(x)$.

Lemma II. Let the experiment $M = M(c+\tau)$ contain smaller information than the global-optimal (in the same conditions $\lambda(x)$) experiment $M_g(\tau)$. Then $\max_x \dot{q}(\phi, x; M) = \dot{q}(\phi, \bar{x}; M) > \max_x \dot{q}(\phi, x; M_g) = r\tau^{-1}$ (the right-hand equality follows from (12)). Proof. For some $k > 0$ it should be the equality

$$q[\phi; M(c+\tau)] = q[\phi; M_g(\tau - k)].$$

Denote $\mu_{\alpha\beta} = f_\alpha(\bar{x}) f_\beta(\bar{x}) \lambda(\bar{x})$. Since the experiments $\mu\, d\tau$ and $M_g(1)\, d\tau$ require equal time $d\tau$ then

$$q(M + \mu\, d\tau) - q(M) \geq q(M + M_g(1)\, d\tau) - q(M) \geq$$

$$\geq q[M_g(\tau - k) + M_g(1) d\tau] - q(M) = q[M_g(\tau - k + d\tau)] - q[M_g(\tau - k)] =$$

$$= r(\tau - k)^{-1} d\tau > r\tau^{-1}\, d\tau ,$$

where the first left-hand inequality follows from the local optimality of $\mu\, d\tau$ and the second left-hand one, according to theorem Y. The lemma is proved.

Theorem YIII. If the efficiency $\lambda(x)$ is independent of τ then the continuous local-optimal planning is at $\tau \to \infty$ asymptotically global-optimal $\lim_{\tau\to\infty} \{ q[\phi, M_c(\tau^0, \tau)] - q[\phi; M_g(\tau^0, \tau; \tau)]\} \to 0$, the maximal loss of time does not exceed in this case a non-optimal fraction $\tau^0 - c$ of the time τ^0 spent for the initial experiment $M(\tau^0)$:

$$q[\phi; M_g(c - \tau^0 + \tau)] \leq q[\phi; M_c(\tau^0, \tau)] \leq$$

$$\leq q[\phi; M_g(\tau^0, \tau; \tau)] \leq q[\phi; M_g(\tau)],$$

where

$$q[\phi; M_g(c)] = q[\phi; M(\tau^0)]. \tag{26}$$

Proof. Since

$$\lim_{\tau\to\infty} \{ q[\phi; M_g(\tau)] - q[\phi; M_g(c - \tau^0 + \tau)]\} = \lim_{\tau\to\infty} \left(r \ln \frac{\tau}{c - \tau^0 + \tau} \right) = 0 ,$$

then for the validity of the theorem it is sufficient to prove that $q[\phi; M_g(c - \tau^0 + \tau)] \leq q[\phi; M_c(\tau^0, \tau)]$. Suppose the contrary, i.e. for a certain τ the difference

$$q[\phi; M_g(c - \tau^0 + \tau)] - q[\phi; M_c(\tau^0, \tau)] \equiv a(\tau) > 0 .$$

According to lemma II $\dot{a}(\tau) < 0$, i.e. in decreasing τ the difference a increases. Consequently, $a(\tau^0) > a(\tau) > 0$ what contradicts (26). The theorem is proved.

We go over to the case $\lambda = \lambda(x, \tau)$. As it is not difficult to prove, if λ does not change its form, $\lambda = \lambda_0(x) \cdot k(\tau)$, then theorem VIII remains valid. If λ changes its form with time then neither M_c nor, all the more, the experiment M_g have any definite asymptotic behaviour at $\tau \to \infty$. However, some limits for the information $q_c(\tau) = q[\phi; M_c(\tau^0, \tau)]$ can be established. Consider $v = \exp\{ r^{-1} q[\phi; M(\tau)]\}$. Denote by $\dot{v}\{\lambda(x, \tau)\}$ the derivative $\dot{v}$ in the case when $M = M_g$ where M_g is the experiment optimal (with the priming $M(\tau^0) = 0$) for the given efficiency* $\lambda(x, \tau)$. According to lemma II, $\frac{d}{d\tau} \exp\{ r^{-1} q_c(\tau)\} \geq$ $\geq \dot{v}\{\lambda(x, \tau)\}$. Hence

* That is optimal for the time-independent efficiency $\lambda(x)$ equal to the time-dependent efficiency $\lambda(x, \tau)$ taken at a moment τ.

$$\bullet\ q_c(\tau) \geq q[\phi; M(\tau^0)] + \tau \ell n \int_{\tau^0}^{\tau} \dot{v}\{\lambda(x,\tau)\}\, d\tau .$$

Majorizing for $\tau^0 \leq \tau' \leq \tau$ the efficiency λ by the product $\Lambda(x) k(\tau) \geq \lambda(x,\tau)$ we can find the upper limit

$$q_c(\tau) \leq q[\phi; M_g(\{\lambda\}; \tau^0, \tau; \tau)] \leq$$

$$\leq q[\phi; M_g(\{\Lambda\}, c)] + \tau \ell n \int_{\tau^0}^{\tau} \dot{v}\{\Lambda\}\, k(\tau')\, d\tau',$$

where the constant c is determined from the condition: the difference $M_g(\{\Lambda\}, c) - M(\tau^0) = \Delta M$ is a positive semi-definite matrix. In addition, the following qualitative considerations may be stated.

If the efficiency λ depends on time in a manner known beforehand then it is a function increasing with time since during the experiment all the changes of equipment are usually made in order to improve its efficiency, rather than to make it worse. In the long-time experiment it is usually desirable that an experiment be close to the global-optimal one $M_g(\tau^0, \tau; \tau)$ at whatever moment τ it is interrupted what is equivalent to the requirement of some uniform optimality, i.e. an optimality with a weight $u(\tau)$ close to a constant. In this case according to theorem YII the maximum of the function

$$\lambda(x,\tau) D_u(x,\tau) = \lambda(x,\tau) \int_{\tau}^{1} D(\hat{\eta}(x), \phi; t)\, u(t)\, dt . \qquad (27)$$

is to be taken as a measurement point. In case of stable efficiency the subvariance D decreases as τ^{-1}, in case of increasing λ, faster yet, and the main contribution to the integral (27) is given by the region of t close to τ. Therefore the functions $\lambda D(\hat{\eta}, \phi; \tau)$ and $\lambda D_u(x, \tau)$ must be close in their form and have a similar location of the maximum. Consequently, the continuous local-optimal planning for the efficiency $\lambda(x,\tau)$ nondecreasing with time must be close to the uniform-optimal one.

6. Kiefer' s condition of minimax

In paper of Kiefer /12/ published in 1961 it was shown that for $M(\tau^0) = 0$ the optimality condition

$$\max q[\phi; M(\tau)] \qquad (\alpha)$$

and the condition x)

$$\min \max_{x} \lambda(x) D[\hat{\eta}(x), \phi; M(\tau)] \qquad (\beta)$$

coincide. According to theorem I this means that the maximum of the information obtained results in the worst conditions for its further accumulation (lemma II). In the framework of the "global" approach suggested in /12/ the degree of generality of the last statement remains not quite clear. We show that the whole region of equivalence of the conditions (α) and (β) is described by the following Table

	$r = m$	$r < m$
$M(\tau^0) = 0$	$(\alpha) = (\beta)$	$(\alpha) = (\beta)$
$M(\tau^0) \neq 0$	$(\alpha) = (\beta)$	$(\alpha) \neq (\beta)$

x) In /12/ it was everywhere assumed that $\lambda(x) = 1$ what is unessential, since a time-independent efficiency λ can be excluded from the consideration by the following redefinition of the functions η and f: $\eta^* = \eta\sqrt{\lambda}$; $f^* = f\sqrt{\lambda}$; $\lambda^* = \lambda/\lambda = 1$.

The non-equivalence of (α) and (β) for $M(r) \neq 0$, $r < m$ is seen from the following example: $\lambda = 1$; $\phi = \theta_1$, $\omega = \theta_2$; $x = x_1, x_2$; $r^1 = r^0 + 0{,}01$;

$$\begin{pmatrix} f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_2(x_2) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ ½ & 6 \end{pmatrix} \quad ; \quad M(r^0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} .$$

The version I given below is optimal and the version II ensures the minimum of $\max_x \lambda \; D(\hat{\eta}, \phi)$

I) $t(x_1) = 0{,}01$; $t(x_2) = 0$; $q = 0{,}00991$; $\max_x D(\hat{\eta}, \phi) = 0{,}971$

II) $t(x_1) = 0$; $t(x_2) = 0{,}01$; $q = 0{,}00184$; $\max_x D(\hat{\eta}, \phi) = 0{,}956$.

Note that in this example $D(\hat{\eta}(x_1), \phi)$ decreases faster if the measurement is made not at x_1 but at the other point x_2 where $D(\hat{\eta}(x), \phi)$ is lower.

In a particular case when $r = m$ such a situation according to (10) is impossible, since in this case $D(^2\hat{\eta}) = 0$ and $| D[\hat{\eta}(x_1), \hat{\eta}(x_2)] |$ never exceeds the largest of the variances $D[\hat{\eta}(x_1)]$, $D[\hat{\eta}(x_2)]$. Therefore, for $r = m$ the conditions of the local-optimality and the fastest decrease of the maximum of $\lambda(x) D[\hat{\eta}(x)]$ coincide and theorem III, if we there replace "q" by "minus $\max_x \lambda \; D(\hat{\eta})$ " remains valid, the proof requiring only small trivial alterations. Considering then the inequalities (24), (25) it is not difficult to see that for $r = m$ the conditions (α) and (β) are equivalent globally too.

In the particular case $M(r^0) = 0$ investigated in /12/ a main cause of coincidence of the conditions (α) and (β) is the equality

$$\Sigma \; \dot{q}_k \, t(\bar{x}_k) = \Sigma \; \lambda(\bar{x}_k) \, D(\hat{\eta}(\bar{x}_k), \phi; r) \, t(\bar{x}_k) = r ,$$

which is valid for $r = \Sigma \; t(\bar{x}_k)$ and due to which $\min \max_x \dot{q}_k$ equal to $r r^{-1}$ is reached if all $\dot{q}_k$ coincide what according to theorem III is just a necessary condition of the optimum.

Thus, the equivalence of the "information" and "minimax" approaches discovered by Kiefer has not quite a universal character and in the most general case $M(r^0) \neq 0$, $r \neq m$, the maximum of information obtained is not leading to the worst condition for its further accumulation.

In conclusion the author expreses his gratitude to A.N. Shiriaev for the help and to Ja.A. Smorodinsky, L.A.Khalfin, L.N. Bolshev and Yu.V. Prokhorov for the discussions.

References

1. S.N. Sokolov. Measurement yielding maximum information and continuous planning of experiments. Preprint JINR D - 573 (1960)

2. G.Elfving. Optimum allocation in linear regression theory. Ann. Math. Stat. 23, 255- 262 (1952).

3. G.Elfving. J. Am. Stat. Ass. 51, 644.

4. H.Chernoff .Locally optimum designs for estimating parameters. Ann. Math. Stat. 24, 586- 602 (1953).

5. A de la Garza. Spacing of information in polynomial regression. Ann. Math. Stat. 25, 123 - 130 (1954).

6. P.G. Guest. The spacing of observations in polynomial regression. Ann. Math. Stat. 29, 294 - 299 (1958).

7. Н.П. Клепиков, С.Н. Соколов. Анализ экспериментальных данных методом максимума правдоподобия. гл. У, препринт О ИЯИ Р-235 (1958).

8. M.Stone. Application of a measure of information to the design and comparison of experiments. Ann.Math .Stat. 30, 55- 70 (1959).

9. J.Keifer and Wolfwitz. Optimum designs in regression problems. Ann. Math. Stat. 30, 271 - 294 (1959).

10. H.A. David, B.E Arens. Optimal spacing in regression analysis. Ann. Math . Stat. 30, 1072 - 1081 (1959).

11. N.P. Klepikov, S.N. Sokolov. Analysis and planning of experiments by the method of maximum likelihood, ch.Y, Akademie-Verlag, Berlin, DDR (1961).

12. J. Kiefer. Optimum designs in regression problems. II. Ann. Math. Stat. 32, 298- 325 (1961).

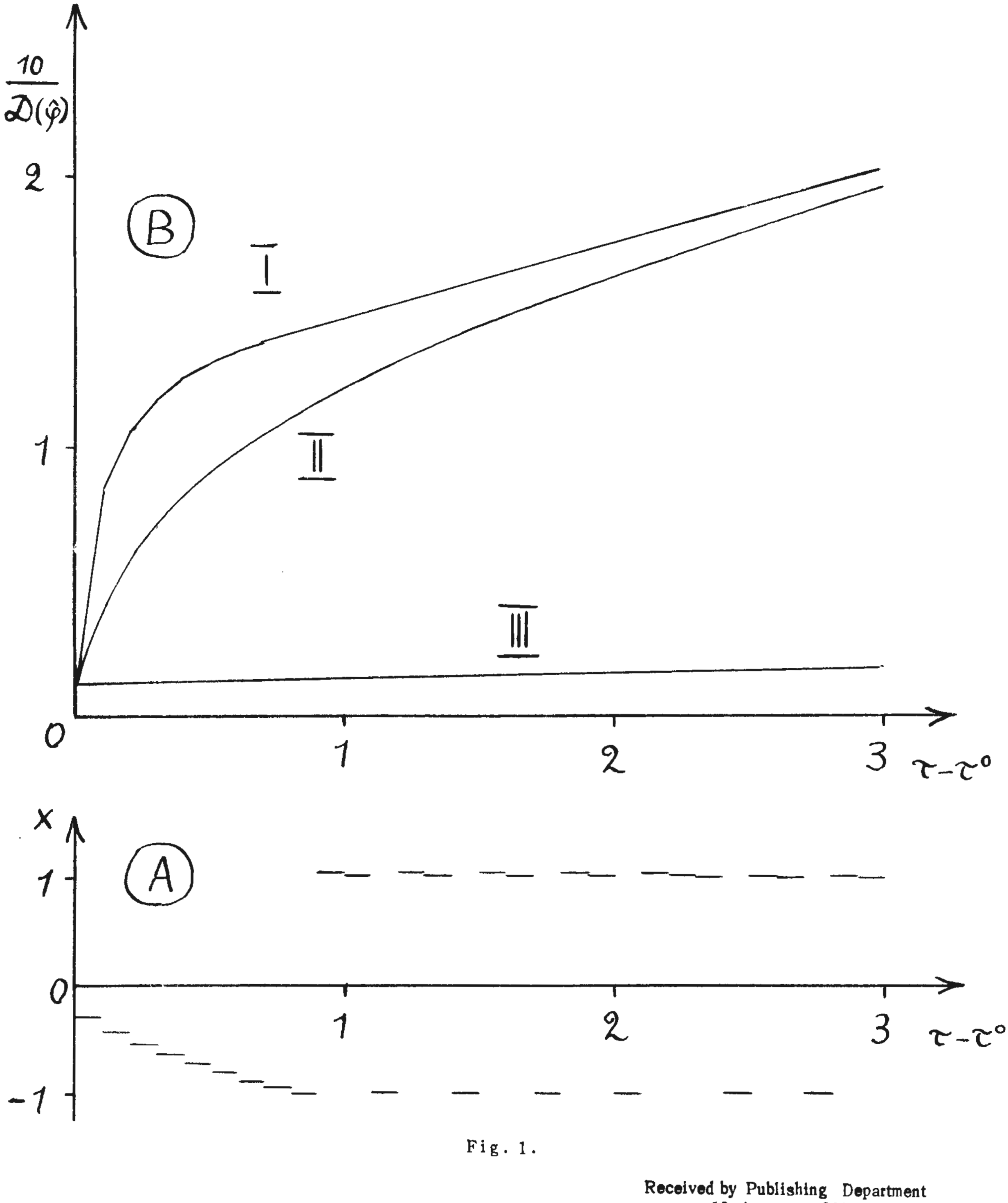

Fig. 1.

Received by Publishing Department
on 13 August, 1962.