ON A SUCCESSIVE TRANSFORMATION OF PROBABILITY DISTRIBUTION AND ITS APPLICATION TO THE ANALYSIS OF THE OPTIMUM GRADIENT METHOD

By Hirotugu Akaike

(Received July 25, 1959)

§ 0. Introduction and summary.

In this paper we define a type of transformation of probability distribution and analyze the limiting behavior of the result of successive applications of the transformation to some initial probability distribution. By using the results of this analysis we can get a fairly general insight into the so-called optimum-gradient method in numerical analysis. We can prove the conjecture which was stated by Forsythe and Motzkin [7] and was used as the logical basis of an acceleration procedure for the optimum gradient method [4][5][6]. It was stated by Forsythe [4] that this conjecture seems to be hard to prove as the related transformation is rather complicated. But our present proof is rather simple. Further, we can see the relation between the condition-number of the related matrix and the convergence rate of the optimum gradient method. By using the relation which according to [5] is first proved by Kantrovich [8], we can say that when the matrix is ill-conditioned the convergence rate tends near to its worst possible value. Using the same data as those treated by Forsythes in paper [5], we give some numerical examples.

There are many important problems, not necessarily of linear type, where the gradient method is applicable [2], [3]. Even in these non-linear case we can expect that when the approximation proceeds the problem will be reduced essentially to the linear one. One of such examples was discussed in the former paper [1] of the present author. Thus, though the results of this paper are concerned with the solution of the simultaneous linear equation $Ax = b$, these results will be generally useful for the analysis of limiting behavior of the approximate solutions in the optimum gradient method.

§ 1. A successive transformation of probability distribution.

In this section we shall treat probability distributions over a set of

mutually different real numbers $\lambda_1, \lambda_2, \cdots, \lambda_n$. First we shall give the definition of our transformation T. For a probability distribution $P \equiv \{(P)_\nu;\ \nu=1, 2, \cdots, n\}$ where $(P)_\nu$'s represent the probabilities attached to λ_ν's respectively the image $TP \equiv \{(TP)_\nu;\nu=1, 2, \cdots, n\}$ of P by T is given by

$$(TP)_\nu=\frac{(P)_\nu(\lambda_\nu-\bar{\lambda}(P))^2}{\sum_{\mu=1}^{n}(P)_\mu(\lambda_\mu-\bar{\lambda}(P))^2} \qquad \nu=1, 2, \cdots, n$$

where $\bar{\lambda}(P)=\sum_{\mu=1}^{n}\lambda_\mu(P)_\mu$ and T has as its domain the set of P's with $\sum_{\mu=1}^{n}(P)_\mu(\lambda_\mu-\bar{\lambda}(P))^2 > 0$. Here we shall give some preliminary lemmas. We shall hereafter use the notation $\overline{f(\lambda)}(P)$ for $\sum_{\mu=1}^{n}f(\lambda_\mu)(P)_\mu$.

LEMMA 1. *We have*

$$\bar{\lambda}(TP)=\bar{\lambda}(P)+\frac{\overline{(\lambda-\bar{\lambda}(P))^3}(P)}{\overline{(\lambda-\bar{\lambda}(P))^2}(P)}\ .$$

PROOF.

$$\begin{aligned}\bar{\lambda}(TP)&=\sum_{\nu=1}^{n}\lambda_\nu(TP)_\nu=\frac{\sum_{\nu=1}^{n}\lambda_\nu(\lambda_\nu-\bar{\lambda}(P))^2(P)_\nu}{\overline{(\lambda-\bar{\lambda}(P))^2}(P)}\\&=\frac{\sum_{\nu=1}^{n}(\lambda_\nu-\bar{\lambda}(P))^3(P)_\nu+\bar{\lambda}(P)\sum_{\nu=1}^{n}(\lambda_\nu-\bar{\lambda}(P))^2(P)_\nu}{\overline{(\lambda-\bar{\lambda}(P))^2}(P)}\\&=\frac{\overline{(\lambda-\lambda(P))^3}(P)}{\overline{(\lambda-\bar{\lambda}(P))^2}(P)}+\bar{\lambda}(P)\ .\end{aligned}$$

LEMMA 2. *We have*

$$\overline{(\lambda-\bar{\lambda}(TP))^2}(TP)\geqq\overline{(\lambda-\bar{\lambda}(P))^2}(P)$$

where = holds only when there are only two ν's with $(P)_\nu>0$.

Note; We are considering only P's with $\overline{(\lambda-\bar{\lambda})(P))^2}(P)>0$, and the case where there is only one ν with $(P)_\nu > 0$ is out of our consideration. The lemma states that by transformation T the variance of the distribution increases except for the special type of distributions stated in the lemma. Thus for P in the domain of T, TP is again in the domain of T.

PROOF.

By using the result of lemma 1 we have

$$\lambda_\nu-\bar{\lambda}(TP)=\lambda_\nu-\bar{\lambda}(P)+\{\bar{\lambda}(P)-\bar{\lambda}(TP)\}=\lambda_\nu-\bar{\lambda}(P)-\frac{\overline{(\lambda-\bar{\lambda}(P))^3}(P)}{\overline{(\lambda-\bar{\lambda}(P))^2}(P)}\,.$$

Thus we get

$$\begin{aligned}&\overline{(\lambda-\bar{\lambda}(TP))^2}(TP)-\overline{(\lambda-\bar{\lambda}(P))^2}(P)\\&=\frac{\sum_{\nu=1}^{n}(\lambda_\nu-\bar{\lambda}(TP))^2(\lambda_\nu-\bar{\lambda}(P))^2(P)_\nu-\{\overline{(\lambda-\bar{\lambda}(P))^2}(P)\}^2}{\overline{(\lambda-\bar{\lambda}(P))^2}(P)}\\&=\frac{\overline{\{(\lambda-\bar{\lambda}(P))^4}(P)\}\,\overline{\{(\lambda-\bar{\lambda}(P))^2}(P)\}-\overline{\{(\lambda-\bar{\lambda}(P))^3}(P)\}^2-\overline{\{(\lambda-\bar{\lambda}(P))^2}(P)\}^3}{\overline{\{(\lambda-\bar{\lambda}(P))^2}(P)\}^2}\,.\end{aligned}$$

For the sake of simplicity we shall use here the abbreviated notation $M_k(P)$ in place of $\overline{(\lambda-\bar{\lambda}(P))^k}(P)$. Then we have

$$\begin{aligned}&\overline{\{(\lambda-\bar{\lambda}(P))^4}(P)\}\,\overline{\{(\lambda-\bar{\lambda}(P))^2}(P)\}-\overline{\{(\lambda-\bar{\lambda}(P))^3}(P)\}^2-\overline{\{(\lambda-\bar{\lambda}(P))^2}(P)\}^3\\&=\begin{vmatrix}1 & 0 & M_2(P)\\ 0 & M_2(P) & M_3(P)\\ M_2(P) & M_3(P) & M_4(P)\end{vmatrix}\,.\end{aligned}$$

If we represent by λ the random variable which follows the probability distribution P i. e. prob $\{\lambda=\lambda_\nu\}=(P)_\nu$ $\nu=1, 2, \cdots, n$, the above stated determinant is the determinant of the product moment matrix of the random variables $(\lambda-E(\lambda))^0\equiv 1$, $\lambda-E(\lambda)$ and $(\lambda-E(\lambda))^2$. The product moment matrix is positive semi-definite and thus the value of the above determinant is non-negative. The value of the determinant is equal to zero if and only if there is non trivial linear relation between 1, $\lambda-E(\lambda)$ and $(\lambda-E(\lambda))^2$. Suppose that there exist constants α_0, α_1, α_2 not all equal to zero satisfying the relation $\alpha_0 1+\alpha_1(\lambda-E(\lambda))+\alpha_2(\lambda-E(\lambda))^2=0$ with probability 1. Such a relation can hold only when there are not more than two values of ν for which prob $\{\lambda=\lambda_\nu\}=(P)_\nu>0$. Taking into account the fact that we have excluded the case where λ is identically equal to a fixed constant, we get the proof of the final statement of the lemma.

Using the results of these lemmas we want to analyze the limiting behavior of $P^{(k)}\equiv T^kP^{(0)}$ as k tends to infinity where $T^kP^{(0)}$ denotes the result of k successive applications of the transformation T to $P^{(0)}$ and the initial probability distribution $P^{(0)}$ is supposed to satisfy the condition $M_2(P^{(0)})>0$.

It is obvious that any infinite subsequence of the sequence $\{\boldsymbol{P}^{(k)}; k=1, 2, \cdots\}$ contains its own convergent subsequence. For a convergent subsequence $\{\boldsymbol{P}^{(\alpha_j)}; j=1, 2, \cdots\}$ of original $\{\boldsymbol{P}^{(k)}\}$ we shall represent by $\boldsymbol{P}_\alpha^{(\infty)}$ its limiting distribution.

Now as $\overline{(\lambda-\bar{\lambda}(\boldsymbol{P}^{(\alpha_j)}))^2}(\boldsymbol{P}^{(\alpha_j)})=\sum_{\nu=1}^{n}(\lambda_\nu-\bar{\lambda}(\boldsymbol{P}^{(\alpha_j)}))^2(\boldsymbol{P}^{(\alpha_j)})$ and $\bar{\lambda}(\boldsymbol{P}^{(\alpha_j)})=\sum_{\nu=1}^{n}\lambda_\nu(\boldsymbol{P}^{(\alpha_j)})_\nu$ it can be seen that

$$\lim_{j\to\infty}\bar{\lambda}(\boldsymbol{P}^{(\alpha_j)})=\bar{\lambda}(\boldsymbol{P}_\alpha^{(\infty)})$$

$$\lim_{j\to\infty}\overline{(\lambda-\bar{\lambda}(\boldsymbol{P}^{(\alpha_j)})^2}(\boldsymbol{P}^{(\alpha_j)})=\overline{(\lambda-\bar{\lambda}(\boldsymbol{P}_\alpha^{(\infty)}))^2}(\boldsymbol{P}_\alpha^{(\infty)}))\ .$$

Further by taking into account the result of lemma 2 we can see that $\overline{(\lambda-\bar{\lambda}(\boldsymbol{P})^{(\alpha_j)})^2}(\boldsymbol{P}^{(\alpha_j)})$ tends monotone-increasingly to $\overline{(\lambda-\bar{\lambda}(\boldsymbol{P}_\alpha^{(\infty)})^2}(\boldsymbol{P}_\alpha^{(\infty)})$ and also that

$$\overline{(\lambda-\bar{\lambda}(\boldsymbol{P}^{(\infty)}))^2}(\boldsymbol{P}_\alpha^{(\infty)})=\lim_{k\to\infty}\overline{(\lambda-\bar{\lambda}(\boldsymbol{P}^{(k)}))^2}(\boldsymbol{P}^{(k)})=\overline{(\lambda-\bar{\lambda}(T\boldsymbol{P}_\alpha^{(\infty)}))^2}(T\boldsymbol{P}_\alpha^{(\infty)})$$

holds.

Now as $M_2(\boldsymbol{P}^{(0)})>0$, we have $M_2(\boldsymbol{P}_\alpha^{(\infty)})>0$, and we have seen that the equality $M_2(T\boldsymbol{P}_\alpha^{(\infty)})=M_2(\boldsymbol{P}_\alpha^{(\infty)})$ holds. Thus from the result of lemma 2 we can see that there are just two numbers $\nu(\alpha, 1)$ and $\nu(\alpha, 2)$ for which $(\boldsymbol{P}_\alpha^{(\infty)})_{\nu(\alpha,1)}>0$, $(\boldsymbol{P}_\alpha^{(\infty)})_{\nu(\alpha,2)}>0$ and $(\boldsymbol{P}_\alpha^{(\infty)})_{\nu(\alpha,1)}+(\boldsymbol{P}_\alpha^{(\infty)})_{\nu(\alpha,2)}=1$ hold. Thus it has been proved that the limit points of the sequence $\{\boldsymbol{P}^{(k)}\}$ are contained in the set of points or distributions of the type analogous to the $\boldsymbol{P}_\alpha^{(\infty)}$ which we have obtained above.

Now we shall further prove that the sequence $\{\boldsymbol{P}^{(k)}\}$ is itself oscillatorily convergent in the sense which will be described bellow.

Taking into account of the monotone-increasing property of the sequence $M_2(\boldsymbol{P}^{(k)})$ we have seen that any $\boldsymbol{P}_\alpha^{(\infty)}$ has one and the same variance $M_2(\boldsymbol{P}_\alpha^{(\infty)})=\lim_{k\to\infty}M_2(\boldsymbol{P}^{(k)})$. Now the distribution $\boldsymbol{P}_\alpha^{(\infty)}$ is characterized by the quantities $\nu(\alpha, 1)$, $\nu(\alpha, 2)$, $(\boldsymbol{P}_\alpha^{(\infty)})_{\nu(\alpha,1)}$, $(\boldsymbol{P}_\alpha^{(\infty)})_{\nu(\alpha,2)}$ and by the relation $M_2(\boldsymbol{P}_\alpha^{(\infty)})=(\boldsymbol{P}_\alpha^{(\infty)})_{\nu(\alpha,1)}(\boldsymbol{P}_\alpha^{(\infty)})_{\nu(\alpha,2)}(\lambda_{\nu(\alpha,1)}-\lambda_{\nu(\alpha,2)})^2=\lim_{k\to\infty}M_2(\boldsymbol{P}^{(k)})$.

Thus the set $\prod_\infty$ of all $\boldsymbol{P}_\alpha^{(\infty)}$'s are composed of at most finitely many distributions. We shall denote the element of $\prod_\infty$ by $\boldsymbol{P}_i$ $i=1, 2, \cdots, N$. $\boldsymbol{P}_i$ is characterized by the quantities $(\boldsymbol{P}_i)_{\nu(i,1)}>0$ $(\boldsymbol{P}_i)_{\nu(i,2)}>0$.

Now we can see that if there is some $\boldsymbol{P}_i$ with $(\boldsymbol{P}_i)_{\nu(i,1)}\neq(\boldsymbol{P}_i)_{\nu(i,2)}$ then the distribution $\boldsymbol{P}_i^*$ with $(\boldsymbol{P}_i^*)_{\nu(i,1)}=(\boldsymbol{P}_i)_{\nu(i,2)}$ and $(\boldsymbol{P}_i^*)_{\nu(i,2)}=(\boldsymbol{P}_i)_{\nu(i,1)}$ is also contained in $\prod_\infty$. This is proved by taking into account of the fact that if $\boldsymbol{P}^{(i_j)}\to\boldsymbol{P}_i(j\to\infty)$ then $T\boldsymbol{P}^{(i_j)}\to T\boldsymbol{P}_i(j\to\infty)$ and the fact that $T\boldsymbol{P}_i=\boldsymbol{P}_i^*$.

It is obvious that if $(\boldsymbol{P}_i)_{\nu(i,1)}=(\boldsymbol{P}_i)_{\nu(i,2)}=1/2$ we have $T\boldsymbol{P}_i=\boldsymbol{P}_i$. Here-

after, we shall sometimes use the notation $P(\nu(i,1),\nu(i,2))$ to represent the P_i to clarify the character of the distribution. We shall here observe the domain of T in its natural relative topology generated by the ordinary topology of n-dimensional Euclidean space. Now as is easily seen T is continuous in its domain, and for any neighborhood $U\{P(\nu(i,1),\nu(i,2))\}$ of $P(\nu(i,1),\nu(i,2))$ there exists a proper neighborhood $V\{P(\nu(i,2),\nu(i,1)\}$ of $P(\nu(i,2),\nu(i,1))$ with the property that for any P in $V\{P(\nu(i,2),\nu(i,1))\}$ its transform TP is contained in $U\{P(\nu(i,1),\nu(i,2))\}$. Here we take the neighborboods $U\{P(\nu(i,1),\nu(i,2))\}$ of P_i's so as to be mutually disjoint. We define the neighborhood $W\{P_i\}$ of P_i as the intersection of $U\{P_i\}$ and $V\{P_i\}$ where V is defined as above. Now there exists a number M such that $P^{(k)}$ with k greater than M are all contained in $\sum_{i=1}^{N} W\{P_i\}$, otherwise there must be some limit point outside Π_∞.

We shall represent by $\{P^{(i_j)}, j=1,2,\cdots\}$ the subsequence of $\{P^{(k)}; k=M+1, M+2, \cdots\}$ contained in $W\{P_i\}$. Then it is seen that $\{P^{(k)}; k=M+1, M+2, \cdots\}=\sum_{i=1}^{N}\{P^{(i_j)}; j=1,2,\cdots\}$ holds as the relation between the sets of points in the sequences. Now from the definition of M and $P^{(i_j)}$ we can see that $TP^{(i_j)} \in \sum_{i=1}^{N} W\{P_i\}$ holds for any i and j and further by the definition of W and V we can see that $\{TP^{(i_j)}; \{j=1,2,\cdots\} \subset U\{TP_i\}=U\{P(\nu(i,2),\nu(i,1))\}$. As the $U\{P_i\}$'s are taken to be mutually disjoint, so $TP^{(i_j)} \notin W\{P_i\}$ for P_i different from TP_i. Thus we have $\{TP^{(i_j)}; j=1,2,\cdots\} \subset W\{TP_i\}$. From this last relation we can also see that $\{T^2P^{(i_j)}; j=1,2,\cdots\} \subset W\{P_i\}$, consequently, that $T^{2r+1}P^{(i_j)} \in W\{TP_i\}$ and $T^{2r+2}P^{(i_j)} \in W\{P_i\}$ hold for $r=0,1,2,\cdots$ and that if $P^{(i_j)}$ is identical to some $P^{(k)}$ of the original sequence $\{P^{(k)}\}$, then $P^{(k+s)} \in W\{P_i\} \cup W\{TP_i\}$ holds for $s=0,1,2,\cdots$. From these relations it follows that Π_∞ is composed of at most two points $P^{(\infty)}$ and $P^{*(\infty)}$ satisfying the relation $P^{*(\infty)}=TP^{(\infty)}$. When $TP^{(\infty)}=P^{(\infty)}$, Π_∞ is composed of only one point $P^{(\infty)}=P(\nu(\infty,1),\nu(\infty,2))$ with $(P^{(\infty)})_{\nu(\infty,1)}=(P^{(\infty)})_{\nu(\infty,2)}=1/2$. Thus we have proved the following;

THEOREM 1: *For any $P^{(0)}$ with $\overline{(\lambda-\lambda(P^{(0)}))^2}(P^{(0)})>0$ the sequences of distributions $\{T^{2r}P^{(0)}; r=1,2,\cdots\}$ and $\{T^{2r+1}P^{(0)}; r=0,1,2,\cdots\}$ are convergent to some limiting distributions $P^{(\infty)}$ and $P^{*(\infty)}$, respectively. These distributions are characterized by $(P^{(\infty)})_{\nu(\infty,1)}>0$ and $(P^{(\infty)})_{\nu(\infty,2)}>0$ satisfying the relation $(P^{(\infty)})_{\nu(\infty,1)}+(P^{(\infty)})_{\nu(\infty,2)}=1$ and $(P^{*(\infty)})_{\nu(\infty,1)}=$*

$(\boldsymbol{P}^{(\infty)})_{\nu(\infty,2)}$, $(\boldsymbol{P}^{*(\infty)})_{\nu(\infty,2)}=(\boldsymbol{P}^{(\infty)})_{\nu(\infty,1)}$. *We have* $T\boldsymbol{P}^{(\infty)}=\boldsymbol{P}^{*(\infty)}$ *and* $\boldsymbol{P}^{*(\infty)}$ *is identical to* $\boldsymbol{P}^{(\infty)}$ *if and only if* $(\boldsymbol{P}^{(\infty)})_{\nu(\infty,1)}=(\boldsymbol{P}^{(\infty)})_{\nu(\infty,2)}=1/2$ *holds.*

Note; We can see that λ_i's with $(\boldsymbol{P}^{(0)})_i=0$ should have been entirely discarded at the very out-set of our whole discussion, and we shall hereafter assume that $(\boldsymbol{P}^{(0)})_i>0$ for $i=1, 2, \cdots, n$.

We shall here discuss some of the effects of the initial distribution $\boldsymbol{P}^{(0)}$ to the limiting distributions $\boldsymbol{P}^{(\infty)}$ and $\boldsymbol{P}^{*(\infty)}$.

THEOREM 2: *If* $\lambda_1<\lambda_i<\lambda_n (i=2, 3, \cdots, n-1)$, then we have

$$\{\nu(\infty, 1), \nu(\infty, 2)\}=\{1, n\}$$

that is, the limiting distributions have their total probability attached to both extremal points.

PROOF; It is obvious that $\lambda_1<\bar{\lambda}(\boldsymbol{P})<\lambda_n$ holds when $(\boldsymbol{P})_1>0$ and $(\boldsymbol{P})_n>0$ hold. Thus from the definition of T it is seen that $(\boldsymbol{P}^{(k)})_1>0$ and $(\boldsymbol{P}^{(k)})_n>0$ hold when $(\boldsymbol{P}^{(k-1)})_1>0$ and $(\boldsymbol{P}^{(k-1)})_n>0$ hold. From this fact, under the condition of the theorem, we have $(\boldsymbol{P}^{(k)})_1>0$ and $(\boldsymbol{P}^{(k)})_n>0$ for $k=0, 1, 2, \cdots$. Suppose $\nu(\infty, 2)<n$ holds. Then we have $\lambda_{\nu(\infty,1)}<\bar{\lambda}(\boldsymbol{P}^{(\infty)})$, $\bar{\lambda}(\boldsymbol{P}^{*(\infty)})<\lambda_{\nu(\infty,2)}<\lambda_n$ and we can find a K and $\bar{\lambda}$, $\underline{\lambda}$ such that, $\lambda_{\nu(\infty,1)}<\underline{\lambda}<\bar{\lambda}(\boldsymbol{P}^{(k)})<\bar{\lambda}<\lambda_{\nu(\infty,2)}$ holds for all $k\geqq K$. Now taking into account of the relation

$$\begin{aligned}(\boldsymbol{P}^{(k+j)})_n/(\boldsymbol{P}^{(k+j)})_{\nu(\infty,2)} &= \{(\boldsymbol{P}^{(k)})_n/(\boldsymbol{P}^{(k)})_{\nu(\infty,2)}\}\prod_{i=1}^{j}\{(\lambda_n-\bar{\lambda}(\boldsymbol{P}^{(k+i)}))^2/(\lambda_{\nu(\infty,2)}-\bar{\lambda}(\boldsymbol{P}^{(k+i)}))^2\} \\ &\geqq \{(\boldsymbol{P}^{(k)})_n/(\boldsymbol{P}^{(k)})_{\nu(\infty,2)}\}\prod_{i=1}^{j}\{(\lambda_n-\underline{\lambda})^2/(\lambda_{\nu(\infty,2)}-\underline{\lambda})^2\}\ ,\end{aligned}$$

we can see that $(\boldsymbol{P}^{(k)})_n$ cannot tend to 0, this contradicts to the definition of $\boldsymbol{P}^{(\infty)}$. The case where it is supposed that $1<\nu(\infty, 1)$ holds is treated in the same manner.

TREOREM 3: *Under the condition of theorem* 2 *if there is a* λ_i $(i\neq 1, n)$ *for which* $\bar{\lambda}(\boldsymbol{P}^{(k)})\neq\lambda_i$ *for all* $k=0, 1, 2, \cdots$ *then the following inequality holds*

$$\left(\frac{\lambda_n-\lambda_1}{2}\right)^2+\left(\lambda_i-\frac{\lambda_n+\lambda_1}{2}\right)^2\geqq 2\left(\bar{\lambda}(\boldsymbol{P}^{(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)^2.$$

PROOF; We have

$$(P^{(k+2)})_i/(P^{(k+2)})_1$$
$$= \{(P^{(k)})_i/(P^{(k)})_1\}\ \{(\lambda_i-\bar{\lambda}(P^{(k)}))^2(\lambda_i-\bar{\lambda}(P^{(k+1)}))^2\}$$
$$/\{(\lambda_1-\bar{\lambda}(P^{(k)}))^2(\lambda_1-\bar{\lambda}(P^{(k+1)}))^2\}\ .$$

For any $\delta>0$ there is a $K(\varepsilon)$ such that for $k\geq K(\varepsilon)$ we have

$$\{(\lambda_i-\bar{\lambda}(P^{(k)}))^2(\lambda_i-\bar{\lambda}(P^{(k+1)}))^2\}/\{(\lambda_1-\bar{\lambda}(P^{(k)}))^2(\lambda_1-\bar{\lambda}(P^{(k+1)}))^2\}$$
$$>\{(\lambda_i-\bar{\lambda}(P^{(\infty)}))^2(\lambda_i-\bar{\lambda}(P^{*(\infty)}))^2\}/(\lambda_1-\bar{\lambda}(P^{(\infty)}))^2(\lambda_1-\bar{\lambda}(P^{*(\infty)}))^2\}-\varepsilon\ .$$

From this relation we have

$$(\lambda_i-\bar{\lambda}(P^{(\infty)}))^2(\lambda_i-\bar{\lambda}(P^{*(\infty)}))^2/(\lambda_1-\bar{\lambda}(P^{(\infty)}))^2(\lambda_1-\bar{\lambda}(P^{*(\infty)}))^2\leqq 1$$

and

$$\left(\lambda_i-\frac{\lambda_n+\lambda_1}{2}-\left(\bar{\lambda}(P^{(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)\right)^2\left(\lambda_i-\frac{\lambda_n+\lambda_1}{2}-\left(\bar{\lambda}(P^{*(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)\right)^2$$
$$\leqq\left(\lambda_1-\frac{\lambda_n+\lambda_1}{2}-\left(\bar{\lambda}(P^{(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)\right)^2\left(\lambda_1-\frac{\lambda_n+\lambda_1}{2}-\left(\bar{\lambda}(P^{*(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)\right)^2.$$

By using the relation $\bar{\lambda}(P^{*(\infty)})=\lambda_n+\lambda_1-\bar{\lambda}(P^{(\infty)})$ we have

$$\left(\lambda_i-\frac{\lambda_n+\lambda_1}{2}-\left(\bar{\lambda}(P^{(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)\right)^2\left(\lambda_i-\frac{\lambda_n+\lambda_1}{2}+\left(\bar{\lambda}(P^{(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)\right)^2$$
$$\leqq\left(\frac{\lambda_1-\lambda_n}{2}-\left(\bar{\lambda}(P^{(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)\right)^2\left(\frac{\lambda_1-\lambda_n}{2}+\left(\bar{\lambda}(P^{(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)\right)^2$$

or

$$\left(\left(\lambda_i-\frac{\lambda_n+\lambda_1}{2}\right)^2-\left(\bar{\lambda}(P^{(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)^2\right)^2$$
$$\leqq\left(\left(\frac{\lambda_1-\lambda_n}{2}\right)^2-\left(\bar{\lambda}(P^{(\infty)})-\frac{\lambda_n+\lambda_1}{2}\right)^2\right)^2.$$

From this last inequality we can get the desired result.

§ 2. Application to the optimum gradient method

In this section we shall discuss the application of the results of the former section to the so-called optimum gradient method. Let A be an n-by-n matrix, not singular, and let x, b denote the n-rowed column vectors. The optimum gradient method for the solution of simultaneous linear equation $Ax=b$ with respect to the metric $\|\ \|_P$ is defined as fol-

lows*); Take a positive definite symmetric matrix P. Then the gradient of the error function $\| Az-b \|_P^2=(Az-b,\ P(Az-b))$ at z is given by $2(A'P(Az-b))$. Given the k-th approximate solution $x_k=x+\varepsilon_k$ where x represents the desired solution the optimum gradient method proceeds by finding the $(k+1)$-th approximate solution $x_{k+1}=x_k-\gamma_k\zeta_k$ where

$$\min_{\gamma} \| A(x_k-\gamma\zeta_k)-b \|_P^2=\|A(x_k-\gamma_k\zeta_k)-b\|_P^2$$
$$\gamma_k=(A\varepsilon_k,\ A\zeta_k)_P/\|A\zeta_k\|_P^2$$
$$\zeta_k=\frac{1}{2}\ \text{gradient at}\ x_k\ =A'PA\varepsilon_k\ .$$

We shall represent by $(0<)\lambda_1\leqq\lambda_2\leqq\cdots\leqq\lambda_n$ the eigenvalues of $A'PA$ and by $\xi_1, \xi_2, \cdots, \xi_n$ the corresponding eigenvectors which are supposed to be orthonormal. We exclude the trivial case where $\lambda_1=\lambda_2=\cdots=\lambda_n$. Then we have the following.

THEOREM 4; *In the optimum gradient method with respect to the metric* $\| \ \|_P$

i) ε_k $(=x_k-x)$ *tends to be approximated by a linear combination of two fixed eigenvectors of* $A'PA$ *with the eigenvalues equal to* $\mathrm{Max}\,(\lambda_i\,;\,(\varepsilon_0,\xi_i)\neq 0)$ *and* $\mathrm{Min}\,(\lambda_i\,;\,(\varepsilon_0,\xi_i)\neq 0)$, *respectively, and*

ii) ε_k *alternates asymptotically in two fixed directions.*

PROOF: Here we prove the theorem for the case where $\lambda_1<\lambda_2<\cdots<\lambda_n$ and $(\varepsilon_0,\xi_1)\neq 0$, $(\varepsilon_0,\xi_n)\neq 0$ hold. Modifications necessary for the proofs of other cases are obvious and are omitted here.

Suppose that the error ε_k of the k-th approximate solution is represented as $\varepsilon_k=\sum_{i=1}^{n}\alpha_i^{(k)}\lambda_i^{-1}\xi_i$. Then ε_{k+1} is given by the following;

$$\varepsilon_k=\sum_{i=1}^{n}\alpha_i^{(k+1)}\lambda_i^{-1}\xi_i=\varepsilon_k-\gamma_k\zeta_k=\varepsilon_k-\gamma_kA'PA\varepsilon_k$$
$$=\sum_{i=1}^{n}\alpha_i^{(k)}(1-\gamma_k\lambda_i)\lambda_i^{-1}\xi_i=\gamma_k\sum_{i=1}^{n}\alpha_i^{(k)}(\gamma_k^{-1}-\lambda_i)\lambda_i^{-1}\xi_i$$

where $\gamma_k^{-1}=\| A\zeta_k \|_P^2/(A\varepsilon_k,\ A\zeta_k)_P=\sum_{i=1}^{n}(\alpha_i^{(k)})^2\lambda_i/\sum_{i=1}^{n}(\alpha_i^{(k)})^2>0$.

If we consider the set of values $((\alpha_1^{(k)})^2/\sum_{i=1}^{n}(\alpha_i^{(k)})^2, (\alpha_2^{(k)})^2/\sum_{i=1}^{n}(\alpha_i^{(k)})^2, \cdots, (\alpha_n^{(k)})^2/\sum_{i=1}^{n}(\alpha_i^{(k)})^2)$ as a probability distribution $\boldsymbol{P}^{(k)}$ over $(\lambda_1, \lambda_2, \cdots, \lambda_n)$, then

*) Here we use the notations $\| x \|_P^2$ and $(x, y)_P$ to represent the quantities (x, Px) and (x, Py), respectively, where (x, y) denotes the ordinary inner product of vectors x and y. A' denotes the transposition of A.

$\boldsymbol{P}^{(k+1)} = ((\alpha_1^{(k+1)})^2/\sum_{i=1}^{n}(\alpha_i^{(k+1)})^2, (\alpha_2^{(k+1)})^2/\sum_{i=1}^{n}(\alpha_i^{(k+1)})^2, \cdots, (\alpha_n^{(k+1)})^2/\sum_{i=1}^{n}(\alpha_i^{(k+1)})^2)$ is represented as $\boldsymbol{P}^{(k+1)}=T\boldsymbol{P}^{(k)}$, transformation T being defined in § 1. Thus the results of the former section are applicable to the present sequence of $\boldsymbol{P}^{(k)}$'s and we have for some non-zero c

$$\lim_{h\to\infty}\boldsymbol{P}^{(2h)}=\left(\frac{1}{1+c^2}, 0, \cdots, 0, \frac{c^2}{1+c^2}\right)\equiv\boldsymbol{P}^{(\infty)}$$

$$\lim_{h\to\infty}\boldsymbol{P}^{(2h+1)}\left(\frac{c^2}{1+c^2}, 0, \cdots, 0, \frac{1}{1+c^2}\right)\equiv\boldsymbol{P}^{*(\infty)}$$

and the limit points of the sequence of direction-cosines

$$\left(\frac{\alpha_1^{(2h)}}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(2h)})^2}}, \frac{\alpha_2^{(2h)}}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(2h)})^2}}, \cdots, \frac{\alpha_n^{(2h)}}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(2h)})^2}}\right)$$

are limited to the set of points of the type

$$\left(\pm\frac{1}{\sqrt{1+c^2}}, 0, \cdots, 0, \pm\frac{c}{\sqrt{1+c^2}}\right).$$

Now consider the transformation $\hat{T}$ defined over the set of points $\{(\delta_1, \delta_2, \cdots, \delta_n); \sum_{i=1}^{n}\delta_i^2=1, \delta_i^2<1 (i=1, 2, \cdots, n)\}$ by the following;

$$\hat{T}(\delta_1, \delta_2, \cdots, \delta_n)$$
$$=\left(\frac{\delta_1(\bar{\lambda}-\lambda_1)}{\sqrt{\sum_{i=1}^{n}\delta_i^2(\bar{\lambda}-\lambda_i)^2}}, \frac{\delta_2(\bar{\lambda}-\lambda_2)}{\sqrt{\sum_{i=1}^{n}\delta_i^2(\bar{\lambda}-\lambda_i)^2}}, \cdots, \frac{\delta_n(\bar{\lambda}-\lambda_n)}{\sqrt{\sum_{i=1}^{n}\delta_i^2(\bar{\lambda}-\lambda_i)^2}}\right)$$

where $\bar{\lambda}=\sum_{i=1}^{n}\delta_i^2\lambda_i$. Then we have

$$\left(\frac{\alpha_1^{(k+1)}}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(k+1)})^2}}, \frac{\alpha_2^{(k+1)}}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(k+1)})^2}}, \cdots, \frac{\alpha_n^{(k+1)}}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(k+1)})^2}}\right)$$
$$=\hat{T}\left(\frac{\alpha_1^{(k)}}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(k)})^2}}, \frac{\alpha_2^{(k)}}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(k)})^2}}, \cdots, \frac{\alpha_n^{(k)}}{\sqrt{\sum_{i=1}^{n}(\alpha_i^{(k)})^2}}\right)$$

and relations such as

$$\hat{T}\left(\frac{1}{\sqrt{1+c^2}}, 0, \cdots, 0, \frac{c}{\sqrt{1+c^2}}\right)=\left(\frac{|c|}{\sqrt{1+c^2}}, 0, \cdots, 0, \frac{-c}{|c|\sqrt{1+c^2}}\right)$$

$$\hat{T}\left(\frac{|c|}{\sqrt{1+c^2}}, 0, \cdots, 0, \frac{-c}{|c|\sqrt{1+c^2}}\right)=\left(\frac{1}{\sqrt{1+c^2}}, 0, \cdots, 0, \frac{c}{\sqrt{1+c^2}}\right).$$

Obviously $\hat{T}$ is continuous in its domain and by applying the eintirely similar arguments in the proof of theorem 1 of § 1, which were used to show the oscillatory convergence of $\boldsymbol{P}^{(k)}$, we can get the conclusions i) and ii) of the present theorem.

Note: It may be of some use to remark here the relations $(\zeta_k, \zeta_{k+1})=0$,

$$\frac{||\zeta_{k+1}||_I^2}{||\zeta_k||_I^2}=\frac{\overline{(\lambda-\bar{\lambda}(\boldsymbol{P}^{(k)}))^2}(\boldsymbol{P}^{(k)})}{\{\bar{\lambda}(\boldsymbol{P}^{(k)})\}^2}=(\text{coefficient of variation of } \boldsymbol{P}^{(k)})^2 .$$

Thus taking into account the fact that $\gamma_k^{-1}=\bar{\lambda}(\boldsymbol{P}^{(k)})$ and lemma 2 of § 1, we can see that the ratio $||\zeta_{k+1}||_I^2/||\gamma_k\zeta_k||_I^2=\overline{(\lambda-\bar{\lambda}(\boldsymbol{P}^{(k)}))^2}(\boldsymbol{P}^{(k)})$ tends monotone-increasingly to the limit $(\lambda_n-\lambda_1)^2c^2(1+c^2)^{-2}$.*)

This theorem provides the theoretical foundations of the following acceleration procedure which is a slightly modified version of the acceleration procedure proposed by Forsythe and Motzkin [5].

Acceleration procedure for the optimum gradient method:

In the optimum gradient method when the direction of vector ζ_k *is nearly the same as that of* ζ_{k-2}, *it is recommendable to insert the step defined by the following*;

$$x_{k+1}=x_k-\hat{\gamma}_k(x_{k-2}-x_k)$$

where

$$\hat{\gamma}_k=(A\varepsilon_k, A(x_{k-2}-x_k))_P/||A(x_{k-2}-x_k)||_P^2 .$$

The rationale for this procedure is as follows; The fact that the direction of ζ_k is nearly the same of that of ζ_{k-2} means that $\zeta_{k-2}\fallingdotseq||\zeta_{k-2}||_I||\zeta_k||_I^{-1}\zeta_k$ where the sign $\fallingdotseq$ is used in place of the description "is approximately equal to". Then we can see that

$$\begin{aligned} x_{k-2}-x_k=\varepsilon_{k-2}-\varepsilon_k&=(A'PA)^{-1}(\zeta_{k-2}-\zeta_k)\\ &\fallingdotseq(||\zeta_{k-2}||_I||\zeta_k||_I^{-1}-1)(A'PA)^{-1}\zeta_k\\ &=(||\zeta_{k-2}||_I||\zeta_k||_I^{-1}-1)\varepsilon_k \end{aligned}$$

holds and it is obvious that in this case $x_{k-2}-x_k$ will be a good candidate for the correcting term of the $x_k=x+\varepsilon_k$. Now if we can suppose that $\varepsilon_{k-2}\fallingdotseq\beta(\lambda_1^{-1}\xi_1+c\lambda_n^{-1}\xi_n)$ then we have

*) It is supposed here that $(\boldsymbol{P}^{(0)})_1>0$ and $(\boldsymbol{P}^{(0)})_n>0$ hold.

$$\zeta_{k-2} \doteqdot \beta(\xi_1 + c\xi_n)$$

$$\gamma_{k-2} = (A\varepsilon_{k-2}, A\zeta_{k-2})_P / \| A\zeta_{k-2} \|_P^2 \doteqdot (1+c^2)/(\lambda_1 + c^2\lambda_n)$$

$$\varepsilon_{k-1} = \varepsilon_{k-2} - \gamma_{k-2}\zeta_{k-2} \doteqdot \beta(\lambda_1 + c^2\lambda_n)^{-1}(\lambda_n - \lambda_1)c^2\{\lambda_1^{-1}\xi_1 - c^{-1}\lambda_n^{-1}\xi_n\}$$

$$\| A\varepsilon_{k-2} \|_P^2 = (\varepsilon_{k-2}, \zeta_{k-2}) \doteqdot \beta^2(\lambda_1^{-1} + c^2\lambda_n^{-1})$$

$$\| A\varepsilon_{k-1} \|_P^2 = (\varepsilon_{k-1}, \zeta_{k-1}) \doteqdot \beta^2(\lambda_1 + c^2\lambda_n)^{-2}(\lambda_n - \lambda_1)^2 c^4(\lambda_1^{-1} + c^{-2}\lambda_n^{-1})$$

$$\varepsilon_k \doteqdot \beta(\lambda_n - \lambda_1)^2(\lambda_1 + c^2\lambda_n)^{-1}(\lambda_1 + c^{-2}\lambda_n)^{-1}\{\lambda_1^{-1}\xi_1 + c\lambda_n^{-1}\xi_n\}$$

$$\doteqdot \| A\varepsilon_{k-1} \|_P^2 \| A\varepsilon_{k-2} \|_P^{-2}\varepsilon_{k-2} .$$

It is obvious that in this case the direction of ζ_k is nearly the same as that of ζ_{k-2}. This fact and the result of theorem 1 which assures that we can expect that ε_{k-2} takes the form just stated when k becomes large, give the theoretical basis of the recommendation of the present acceleration procedure.

The rate of convergence of the optimum gradient method:

By using the results of calculations in the former paragraph we can see that when the distribution $\boldsymbol{P}^{(k)}$ corresponding to ε_k tends alternatingly to the limiting distributions $\boldsymbol{P}^{(\infty)} = (1/(1+c^2), 0, \cdots, 0, c^2/(1+c^2))$ and $\boldsymbol{P}^{*(\infty)} = (c^2/(1+c^2), 0, \cdots 0, 1/(1+c^2))$ i.e. ε_k tends to be approximated by $\beta_k\{\lambda_1^{-1}\xi_1 + c\lambda_n^{-1}\xi_n\}$ with some scalar β_k and c, the rate of convergence $\| A\varepsilon_{k+1} \|_P^2 / \| A\varepsilon_k \|_P^2$ tends to the value $(\lambda_n - \lambda_1)^2\{(\lambda_1 + \lambda_n)^2 + (c - c^{-1})^2\lambda_1\lambda_n\}^{-1}$. Thus the rate of convergence of the optimum gradient method is eventually determined by the value of c^2 which is inherited from the initial vector x_0 or $\boldsymbol{P}^{(0)}$. The values $(\lambda_n - \lambda_1)^2\{(\lambda_1 + \lambda_n)^2 + (c - c^{-1})^2\lambda_1\lambda_n\}^{-1}$ attains its maximum value $(\lambda_n - \lambda_1)^2(\lambda_1 + \lambda_n)^{-2}$ when $c^2 = 1$ holds. Now $(\lambda_n - \lambda_1)^2/(\lambda_n + \lambda_1)^2 = (t-1)^2/(t+1)^2$ where $t \equiv \lambda_n/\lambda_1$ is a so-called condition-number of the matrix of $A'PA$.

We shall here make a slight digression for investigation of the meaning of the condition-number. Now define the function $f(\hat{x}, c) \equiv \| A(\hat{x} - c\hat{\zeta}) - b \|_P^2 / \| A\hat{x} - b \|_P^2$ of a vector $\hat{x}(\neq x)$ and a scaler c where $\hat{\zeta} = A'PA\hat{\varepsilon}$, $\hat{\varepsilon} = \hat{x} - x$ (x is the solution of $Ax = b$). If $\hat{\varepsilon} = \sum_{i=1}^{n} \beta_i \xi_i$, we have $f(\hat{x}, c) = \sum_{i=1}^{n} \beta_i^2\lambda_i(1 - c\lambda_i)^2 / \sum_{i=1}^{n} \beta_i^2\lambda_i$ and thus $\operatorname{Min}_c \operatorname{Max}_{\hat{x}} f(\hat{x}, c) = \operatorname{Min}_c (\operatorname{Max}((1 - c\lambda_1)^2, (1 - c\lambda_n)^2)) = (1 - 2\lambda_1/(\lambda_1 + \lambda_n))^2 = (\lambda_n - \lambda_1)^2/(\lambda_n + \lambda_1)^2$ for $c = 2/(\lambda_1 + \lambda_n)$. Obviously $\operatorname{Min}_c \operatorname{Max}_{\hat{x}} f(\hat{x}, c) \geqq \operatorname{Max}_{\hat{x}} \operatorname{Min}_c f(\hat{x}, c)$ holds and $\operatorname{Min} f(\hat{x}, c) = f(\hat{x}, \hat{\gamma})$ where $\hat{\gamma} = (A\hat{\varepsilon}, A\hat{\zeta})_P / \| A\hat{\zeta} \|_P^2$ is given by the optimum gradient method. Thus we get $f(\hat{x}, \hat{\gamma}) \leqq (\lambda_n - \lambda_1)^2/(\lambda_n + \lambda_1)^2$. We have already seen that $f(\hat{x}, \hat{\gamma}) = (\lambda_n - \lambda_1)^2/(\lambda_n + \lambda_1)^2$ holds for $\hat{x}$ with corresponding probability

distribution $\hat{P} = (1/2, 0, \cdots, 0, 1/2)$ and we can see that the value $(\lambda_n - \lambda_1)^2/(\lambda_n + \lambda_1)^2 = (t-1)^2/(t+1)^2$ gives the least upper bound of the convergence rate $f(\hat{x}, \hat{\gamma})$ of the optimum gradient method. It is stated in [5] that a proof of this fact was given by Kantrovitch [8] but as the Kantrovitch's paper was not available, we gave a proof to it for the sake of completeness of the following discussion.

Here we use the result of theorem 3 of § 1 to see why the optimum gradient method often converges with the convergence rate nearly equal to its worst possible value $(\lambda_n-\lambda_1)^2/(\lambda_n+\lambda_1)^2$. This fact was also noticed by Forsythes [5]. For $P^{(\infty)}=(1/(1+c^2), 0, \cdots, 0, c^2/(1+c^2))$ and $P^{*(\infty)} = (c^2/(1+c^2), 0, \cdots, 0, 1/(1+c^2))$ we have $\bar{\lambda}(P^{(\infty)}) = (\lambda_1 + c^2\lambda_n)/(1+c^2)$ and $\bar{\lambda}(P^{*(\infty)}) = (c^2\lambda_1 + \lambda_n)/(1+c^2)$ and thus we have $(\bar{\lambda}(P^{(\infty)}) - (\lambda_n+\lambda_1)/2)^2 = (\bar{\lambda}(P^{*(\infty)}) - (\lambda_n+\lambda_1)/2)^2 = (\lambda_n-\lambda_1)^2(1-c^2)^2/4(1+c^2)^2$. Thus under the condition which assumes that the point λ_i is not discarded during the course of approximation procedure, we have from the result of theorem 3 of § 1

$$\left(\frac{\lambda_n-\lambda_1}{2}\right)^2+\left(\lambda_i-\frac{\lambda_n+\lambda_1}{2}\right)^2 \geqq \frac{(1-c^2)^2}{2(1+c^2)^2}(\lambda_n-\lambda_1)^2 .$$

By putting $\delta_i \equiv (\lambda_i-(\lambda_n+\lambda_1)/2)/((\lambda_n-\lambda_1)/2)$ we get from the above inequality the following

$$4\left\{\frac{1+\delta_i^2}{1-\delta_i^2}\right\} \geqq (c-c^{-1})^2 .$$

Thus, for example, when there exists some λ_i which satisfies the condition of theorem 3 and with $|\delta_i| \ll 1$ we can expect that $(c-c^{-1})^2$ is near or less than 4. Now the convergence rate at the point corresponding to the $P^{(\infty)}=(1/(1+c^2), 0, \cdots, 0, c^2/(1+c^2))$ is given by $(\lambda_n-\lambda_1)^2(\lambda_n+\lambda_1)^{-2}$ $\{1+(c-c^{-1})^2(\lambda_n\lambda_1^{-1}+\lambda_1\lambda_n^{-1}+2)^{-1}\}^{-1}$ and by putting $\varepsilon=1-(\lambda_n-\lambda_1)/(\lambda_n+\lambda_1)$ this is represented as $(\lambda_n-\lambda_1)^2(\lambda_n+\lambda_1)^{-2}\{1+(c-c^{-1})^2(2\varepsilon^{-1}+1+\varepsilon(2-\varepsilon)^{-1})^{-1}\}^{-1}$. Thus we can see that the convergence rate tends to be greater than $(\lambda_n-\lambda_1)^2(\lambda_n+\lambda_1)^{-2}\{1+(\varepsilon/2)(c-c^{-1})^2\}^{-1}$ or a fortiori than $(\lambda_n-\lambda_1)^2(\lambda_n+\lambda_1)^{-2}$ $\{1+2\varepsilon(1+\delta_i^2)(1-\delta_i^2)^{-1}\}^{-1}$. From this last relation we can see that when we use the optimum gradient method for ill-conditioned ($\varepsilon \ll 1$) matrix $A'PA$, then the rate of convergence tends near to its worst possible value, especially when there is some λ_i near the value $(\lambda_n+\lambda_1)/2$ i.e. $|\delta_i| \ll 1$.

As to one of the numerical examples described by Forsythes [5], and which will be treated fully in the following paragraph of this section, we have $\varepsilon=0.01073$ and

$$\begin{aligned}&(\lambda_n-\lambda_1)^2(\lambda_n+\lambda_1)^{-2}\{1+2\varepsilon(1+\delta_i^2)(1-\delta_i^2)^{-1}\}^{-1}\\&=(\lambda_n-\lambda_1)^2(\lambda_n+\lambda_1)^{-2}\times 0.9789=0.9580 \quad \text{for} \quad \delta_5^2=0.001321\\&=(\lambda_n-\lambda_1)^2(\lambda_n+\lambda_1)^{-2}\times 0.7171=0.7018 \quad \text{for} \quad \delta_2^2=0.896854\,.\end{aligned}$$

We can clearly see in this example the effect of the small value of ε making the convergence rate tend near to its worst possible value. Thus, it seems that our present analysis gives fairly general theoretical explanation to the fact that the optimum gradient method converges slowly when the matrix is ill-conditioned.

Numerical example

Here we shall present some numerical results obtained by using the same A and b as those treated by Forsythes [5] and by putting $P=I$ (identity matrix). These results are obtained by using a FACOM-128 relay computer of our institute. We have used the computation scheme of optimum gradient method where the acceleration step is inserted automatically when the condition $(\zeta_{k-2}, \zeta_k)_I/(\| \zeta_{k-2}\|_I\|\zeta_k\|_I)>\delta$ is satisfied for some preassigned value of $\delta\,(1>\delta>0)$.

We have

$$A=\begin{bmatrix}\sqrt{0.00268704} & & & & & \\ & \sqrt{0.01581310} & & & 0 & \\ & & \sqrt{0.08234830} & & & \\ & & & \sqrt{0.17590130} & & \\ & 0 & & & \sqrt{0.25946632} & \\ & & & & & \sqrt{0.49823436}\end{bmatrix}$$

$$b=[0,\quad 0,\quad 0,\quad 0,\quad 0,\quad 0,]\quad \text{and}$$

$$A'PA=\begin{bmatrix}0.00268704\ (=\lambda_1) & & & & & \\ & 0.01581310\ (=\lambda_2) & & & 0 & \\ & & 0.08234830\ (=\lambda_3) & & & \\ & & & 0.17590130\ (=\lambda_4) & & \\ & 0 & & & 0.25946632\ (=\lambda_5) & \\ & & & & & 0.49823436\ (=\lambda_6)\end{bmatrix}.$$

We have used three values 0.99, 0.999 and 0.9999 as δ. In fig's. 1, 2 and

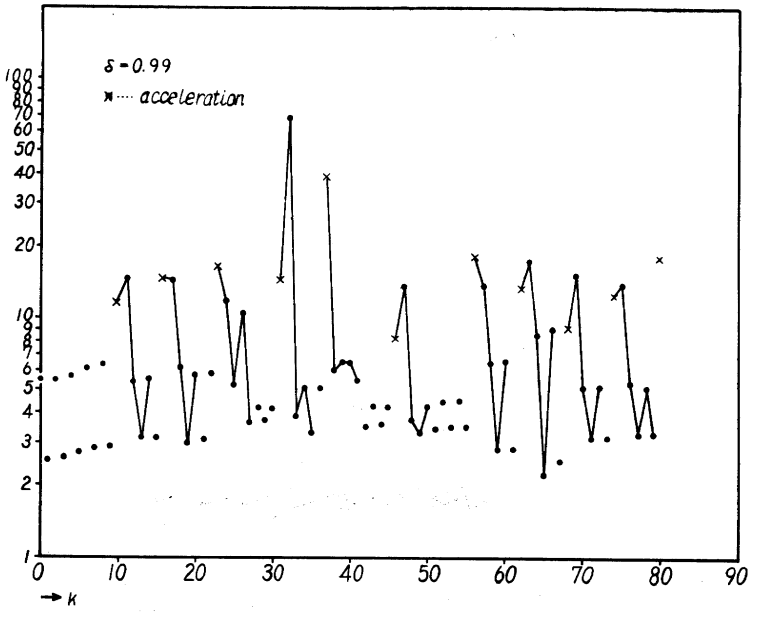
δ=0.99
acceleration
100
90
80
70
60
50
40
30
20
10
9
8
7
6
5
4
3
2
1
0
10
20
30
40
50
60
70
80
90
k

Fig. 1

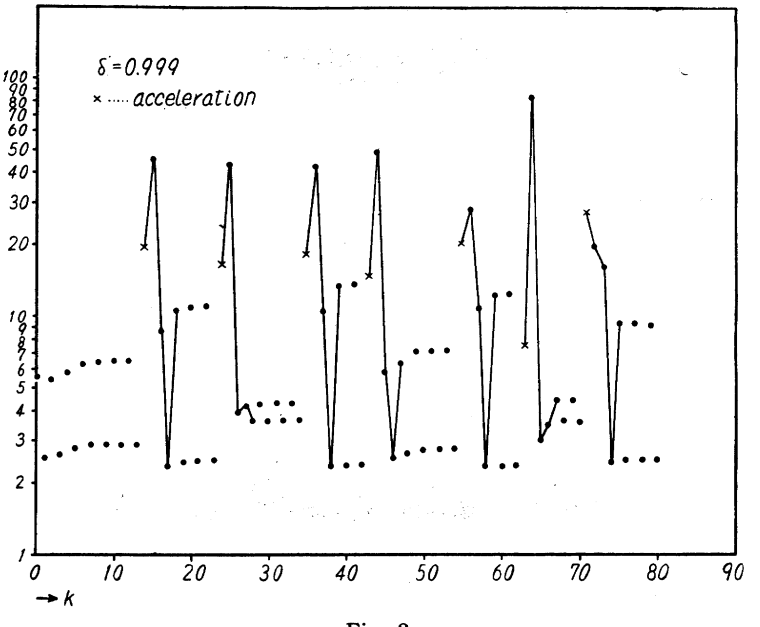
δ=0.999
acceleration
100
90
80
70
60
50
40
30
20
10
9
8
7
6
5
4
3
2
1
0
10
20
30
40
50
60
70
80
90
k

Fig. 2

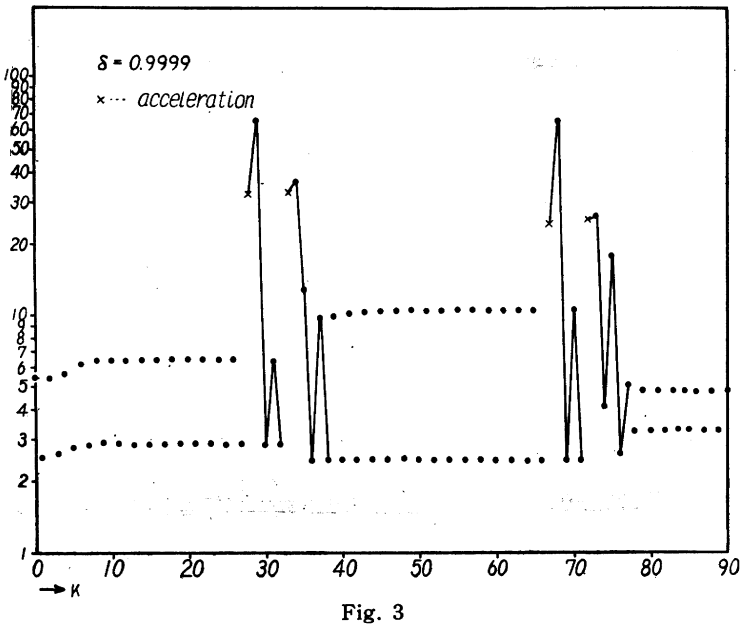

Fig. 3

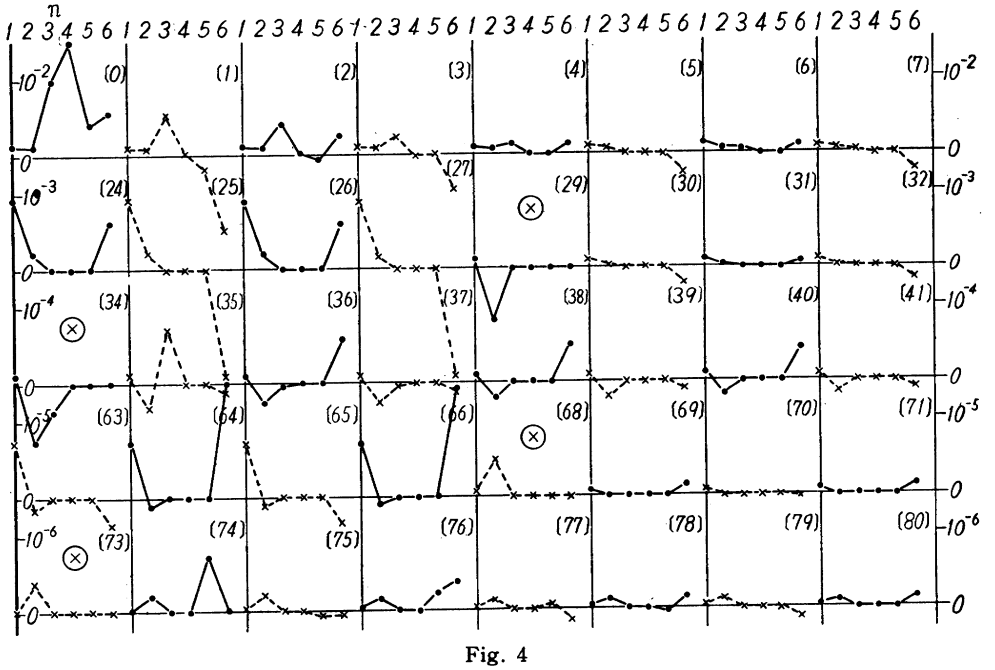

Fig. 4

3 the sequence of γ_k's are represented by $\bullet$ and $\hat{\gamma}_k$'s (acceleration step) are represented by $\times$. These three figures correspond to the above mentioned three values of δ. In fig. 4 part of the values of ζ_k's are illustrated for the case $\delta=0.9999$. The numbers in the blacket in fig. 4 correspond to the k's of fig. 3 and the symbols $\otimes$ signify the result of acceleration procedure.

We can see in these examples a good agreement between the results of our theoretical analysis made in this paper and the results of practical computations.

Acknowledgement

I wish to express my thanks to Mr. K. Isii for reading the manuscript of § 1 of this paper. Thanks are also due to Miss Y. Saigusa who performed the programmings and operations of FACOM-128 automatic relay computer to prepare the numerical results illustrated in § 2.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Akaike, H. "On a computation method for eigenvalue problems and its application to statistical analysis," *Annals of the Institute of Statistical Mathematics* 10, 1-20 (1958).

[2] Chernoff, H. and Divinsky, N. "The computation of maximum-likelihood estimates of linear structural equations," Chapter 10 in *Studies in Econometric Method, Cowles Commission Monograph* 14, Wm. C. Hood and T. C. Koopmans editors, John Wiley and Sons, Inc. New York, (1953).

[3] Crockett, J. B. and Chernoff, H. "Gradient methods of maximization," *Pacific J. Math.*, 5, 33-50 (1955).

[4] Forsythe, G. E. "Solving linear algebraic equations can be interesting," *Bull. Amer. Math. Soc.*, 59, 299-329 (1953).

[5] Forsythe, A. I. and Forsythe, G. E. "Punched-card experiments with accelerated gradient methods for linear eqations," *Contributions to the Solution of systems of Linear Equations and the Determination of Eigenvalues, N.B.S. Applied Mathematics Series* 39, Olga Taussky editor, 55-69 (1954).

[6] Forsythe, G. E. and Motzkin, T. S. "Asymptotic properties of the optimum gradient method," *Bul. Am. Math. Soc.*, 57, 183 (1951) (Abstract).

[7] Forsythe, G. E. and Motzkin, T. S. "Acceleration of the optimum gradient method," Preliminary report, *Bul. Am. Math. Soc.*, 57, 304-305 (1951) (Abstract).

[8] Kantrovich, L. V. "Functional analysis and applied mathematics," *Uspekhi Mathematicheskikh Nauk* (NS), 3, No. 6, 89-185 (1951).