A GRADIENT METHOD UTILIZING APRIORI STOCHASTIC INFORMATION ABOUT THE LOCATION OF THE MINIMUM

Ito, Kumiko
Department of Mathematics, Kyushu University

https://doi.org/10.5109/13076

出版情報：統計数理研究. 15 (3/4), pp.101-111, 1973-03. Research Association of Statistical Sciences
バージョン：
権利関係：

A GRADIENT METHOD UTILIZING APRIORI STOCHASTIC INFORMATION ABOUT THE LOCATION OF THE MINIMUM

By

Kumiko ITOH*

(Received October 5, 1972)

§ 1. Introduction.

Among various approaches to solve the nonlinear simultaneous equations numerically, there is a cost minimization approach by a gradient method, in which we search the locations of the roots of equations by minimizing a square error in the gradient method. In the deterministic case, recently the gradient method is often replaced by a conjugate gradient method, in which the locations of the points having zero derivative in the cost function are searched. In many cases, however, apriori stochastic informations about the location of roots are given to us, and moreover we can assume the minimum value of the cost to be equal to zero. Thus, in this paper, the cost minimization problem with the zero minimum value is to be solved by application of a nonlinear filter technique, which has been conceived to estimate the state variables of nonlinear systems by Athans [4], to utilize apriori informations efficiently.

A remarkable merit of the conjugate method occurs in the quadratic convergence: the minimum of the quadratic function having n variables can be attained by at most n iterations for any starting value. For a class of functions, which are closely similar to the quadratic, the convergence may be also possible during the reasonable times of iterations, judging from the numerical example in the paper [2]. However it needs so much times for the linear search, because the large amount of calculations is required to evaluate the functions and to evaluate its gradients at the various points on the searching line. In the Newton method, on the other hand, the appropriate solution can be obtained in an approximate linearization by the tangents at searching points of a performance index. So, if the initial value is taken near the solution, then the convergence may be very fast, but we do not have any assurance of convergence for the initial value far from the solution.

Thus the main objects of this paper is to extend the Newton method to a more proper linearization by making stochastically the effect of the second order term in Taylor expansion welcome. Once we introduce the stochastic properties, some informations may be available to the prediction of the locations of solutions, and

* Department of Mathematics, Kyushu University, Fukuoka.

then our linear approximations will be expected to be more adequate. As the probability distribution of the locations of the roots, we shall consider the normal distribution, whose standard deviation is assumed to be equal to the distance between a present searching point and the predicted point, in order to apply the linear approximation associated with the stochastic bias correction due to the second order term. That is, in each stage, we first approximate the nonlinear functions by Taylor expansion up to the second order terms at the current searching point, and then approximate the Taylor expansion in a linear form defined by the condition of unbiasedness and of minimum variance under the given normal distribution.

Newton method is considered as the special case of this method, that is, the case when the weight is concentrated to the current searching point. Our method is superior to Newton method both in the times of itterations and in the characteristics of convergence, because our approximation is achieved by imposing the distributed weight over the region covering from the searching point to the predicted point.

To be brief, our method is to estimate the optimal point by using the conditional probability distribution. Actually, our iterative calculation procedure is performed in a way similar to Newton method.

The numerical examples will be given in Section 4, and these will give us full satisfaction in their convergence.

§ 2. The simultaneous equations.

Consider the simultaneous equations

$$\boldsymbol{f}(\mathbf{x})=\mathbf{o}\,,\qquad \boldsymbol{f}=(f^1,\cdots,f^m)\,,\qquad \mathbf{x}=(x^1,\cdots,x^n) \tag{1}$$

with C^2 vector function $\boldsymbol{f}$ and we shall search the solution $\mathbf{x}$. Suppose that the first prediction value $\hat{\mathbf{x}}^*_{0'-1}$ is given. If $\boldsymbol{f}(\hat{\mathbf{x}}^*_{0'-1})\neq 0$, then give a normal distribution $N(\hat{\mathbf{x}}^*_{0'-1}, C^*_{0'-1})$ as the probability distribution of $\mathbf{x}$, where $\hat{\mathbf{x}}^*_{0'-1}$ is a mean and $C^*_{0'-1}$ is a covariance which will be given later. We expand $\boldsymbol{f}(\mathbf{x})$ by Taylor series in the neighbourhood of $\hat{\mathbf{x}}^*_{0'-1}$:

$$\begin{aligned}\boldsymbol{f}(\mathbf{x})=\boldsymbol{f}(\hat{\mathbf{x}}^*_{0'-1})&+\boldsymbol{f}_{\mathbf{x}}(\hat{\mathbf{x}}^*_{0'-1})(\mathbf{x}-\hat{\mathbf{x}}^*_{0'-1})\\&+\frac{1}{2}\sum_{i=1}^{m}\boldsymbol{\varphi}_i(\mathbf{x}-\hat{\mathbf{x}}^*_{0'-1})'\boldsymbol{f}^i_{\mathbf{xx}}(\hat{\mathbf{x}}^*_{0'-1})(\mathbf{x}-\hat{\mathbf{x}}^*_{0'-1})+\mathbf{v}\,,\end{aligned} \tag{2}$$

where

$$(\boldsymbol{f}_{\mathbf{x}}(\hat{\mathbf{x}}^*_{0'-1}))^{ij}=\frac{\partial f^i}{\partial x^j}\bigg|\mathbf{x}=\hat{\mathbf{x}}^*_{0'-1}\,,\qquad (\boldsymbol{f}^i_{\mathbf{xx}}(\hat{\mathbf{x}}^*_{0'-1}))^{jk}=\frac{\partial^2 f^i}{\partial x^j\partial x^k}\,,$$

$$\boldsymbol{\varphi}_i'=(0\cdots 0\overset{i}{1}0\cdots 0)\,,\quad i=1,\cdots,m\,,\quad j=1,\cdots,n\,,\quad k=1,\cdots,n\,.$$

Here $\mathbf{v}$ is the aggregation of the terms whose order is more than three i.e. the error which arises from the second order approximation. In deterministic cases $\mathbf{v}$ may be known theoretically, but, when we search a solution approximating by Taylor series up to the second order, $\mathbf{v}$ may be unknown from this standpoint and will be possible to have various values according to the realization of $\mathbf{x}$. From this

point of view, let $\mathbf{v}$ be a random variable and be a normal distribution with $\mathbf{o}$ mean and small variance.

Now let $\mathbf{y}$ denote the image $\boldsymbol{f}(\mathbf{x})$ for the realization of $\mathbf{x}$, and we shall rewrite our problem in the form of a nonlinear filter problem. Since the solution to be estimated is stationary, the transition formula is given by (3). Moreover let the observation formula be (4).

(3) $$\mathbf{x}_{k+1}=\mathbf{x}_k\,,$$

(4) $$\mathbf{y}_k=\boldsymbol{f}(\hat{\mathbf{x}}^*_{k/k-1})+\boldsymbol{f}_{\mathbf{x}}(\hat{\mathbf{x}}^*_{k/k-1})(\mathbf{x}_k-\hat{\mathbf{x}}^*_{k/k-1})+\frac{1}{2}\sum_{i=1}^{m}\boldsymbol{\phi}_i(\mathbf{x}_k-\hat{\mathbf{x}}^*_{k/k-1})'\boldsymbol{f}^i_{\mathbf{xx}}(\hat{\mathbf{x}}^*_{k/k-1})(\mathbf{x}_k-\hat{\mathbf{x}}^*_{k/k-1})+\mathbf{v}\,,\qquad k=0,1,\cdots.$$

In our problem, if $\mathbf{x}$ would be the solution, the image $\boldsymbol{f}(\mathbf{x})$ must be zero. So, conversely, if the observation value is assumed to be zero always, it is considered that the value $\hat{\mathbf{x}}^*_{k/k}$ estimated from this observation value will naturally approach to the solution. Here $\hat{\mathbf{x}}^*_{k/k}$ is an optimal estimate (a conditional mean) of $\mathbf{x}_k$ based on $\mathbf{y}_0,\cdots,\mathbf{y}_k$ in the case when the observation value $\mathbf{y}_k$ is equal to zero for all k.

We assume that the second order term of observation formula (4) is distributed according to the normal distribution. Putting $\mathbf{x}=\mathbf{x}_k$, and substituting $\mathbf{y}_k=\mathbf{o}$ into the following formulas (5), (6) concerning the conditional normal distribution, we can get the optimal estimate and its error covariance as (7), (8).

(5) $$E(\mathbf{x}/y^k)=E(\mathbf{x}/y^{k-1})+\mathrm{Cov}\,(\mathbf{x},\mathbf{y}_k/y^{k-1})\,\mathrm{Cov}^{-1}(\mathbf{y}_k,\mathbf{y}_k/y^{k-1})(\mathbf{y}_k-E(\mathbf{y}_k/y^{k-1}))\,,$$

(6) $$\mathrm{Cov}\,(\mathbf{x},\mathbf{x}/y^k)=\mathrm{Cov}\,(\mathbf{x},\mathbf{x}/y^{k-1})-\mathrm{Cov}\,(\mathbf{x},\mathbf{y}_k/y^{k-1})\,\mathrm{Cov}^{-1}(\mathbf{y}_k,\mathbf{y}_k/y^{k-1})\,\mathrm{Cov}\,(\mathbf{y}_k,\mathbf{x}/y^{k-1})\,,$$

where $y^k=(\mathbf{y}_0,\cdots,\mathbf{y}_k)$, $y^{k-1}=(\mathbf{y}_0,\cdots,\mathbf{y}_{k-1})$ and $\mathbf{x},\mathbf{y}_0,\cdots,\mathbf{y}_k$ follows the normal distribution.

(7) $$\hat{\mathbf{x}}^*_{k/k}=\hat{\mathbf{x}}^*_{k/k-1}-C^*_{k/k-1}\boldsymbol{f}'_{\mathbf{x}}(\hat{\mathbf{x}}^*_{k/k-1})(\boldsymbol{f}_{\mathbf{x}}(\hat{\mathbf{x}}^*_{k/k-1})C^*_{k/k-1}\boldsymbol{f}'_{\mathbf{x}}(\hat{\mathbf{x}}^*_{k/k-1})+V+L_k)^{-1}(\boldsymbol{f}(\hat{\mathbf{x}}^*_{k/k-1})+\pi(\hat{\mathbf{x}}^*_{k'k-1}))\,,$$

(8) $$C^*_{k/k}=C^*_{k/k-1}-C^*_{k/k-1}\boldsymbol{f}'_{\mathbf{x}}(\hat{\mathbf{x}}^*_{k/k-1})(\boldsymbol{f}_{\mathbf{x}}(\hat{\mathbf{x}}^*_{k/k-1})C^*_{k/k-1}\boldsymbol{f}'_{\mathbf{x}}(\hat{\mathbf{x}}^*_{k/k-1})+V+L_k)^{-1}\boldsymbol{f}(\hat{\mathbf{x}}^*_{k/k-1})C^*_{k/k-1}\,,$$

where

$$\pi(\hat{\mathbf{x}}^*_{k/k-1})=\frac{1}{2}\sum_{i=1}^{m}\boldsymbol{\phi}_i\,tr\,(\boldsymbol{f}^i_{\mathbf{xx}}(\hat{\mathbf{x}}^*_{k/k-1})C^*_{k/k-1})\,,$$

$$(L_k)^{ij}=\frac{1}{2}\,tr\,[\boldsymbol{f}^i_{\mathbf{xx}}(\hat{\mathbf{x}}^*_{k/k-1})C^*_{k/k-1}\boldsymbol{f}^j_{\mathbf{xx}}(\hat{\mathbf{x}}^*_{k/k-1})C^*_{k/k-1}]$$

and V is the covariance of $\mathbf{v}$.

Furthermore from (3), we get

(9) $$C^*_{k/k-1}=C^*_{k-1/k-1}\,,$$

(10) $$\hat{\mathbf{x}}^*_{k/k-1}=\hat{\mathbf{x}}^*_{k-1/k-1}\,,\qquad k=1,2,\cdots,$$

so that the optimal estimate $\hat{\mathbf{x}}^*_{k/k}$ and its error covariance $C^*_{k/k}$ are obtained one after another (see ref. [1]).

In the filter problem, the initial covariance $C^*_{0/-1}$ is given as an apriori covariance and the calculations for successive stages are proceeded recursively. But here the covariances will be determined at each stage newly by checking the order of the accuracy of estimation. Formula (8) can be reduced by assuming that the second order term of (4) follows the normal distribution. In the case when the searching point is near the solution, this hypothesis may be well enough to supplement to the approximation with the first order term, since the second order term is smaller in comparison with the first order term and the real distribution of the second order term takes the form of concentrating at the mean. In case of the searching point far from the solution, however, this hypothesis seems to be unreasonable, since the second order term is not able to be disregarded and the real distribution of the second order term may be far from the normal distribution.

Moreover, $\mathbf{v}$ is considered as the random variable with $\mathbf{o}$ mean and the small covariance, which is independent of $\mathbf{x}_k$. Actually, $\mathbf{v}$ is not independent of $\mathbf{x}_k$, since it comprises terms on and after the third order and $\mathbf{v}$ is not disregarded in case of the seaching point far from the solution. By these reasons, the recursive formula (9) with $\mathbf{v}$ seems to be unsuitable.

The numerical results later show that the method of ref. [4] is not good in the convergence.

To resolve these questions, thus, it is preferable to return back to the establishment of covariance at each stage by the renewed estimation of the probability distribution for the location of the solution around the optimal prediction $\hat{\mathbf{x}}^*_{k/k-1}$. For this purpose, let us determine $\mathbf{x}^*_{k/k-1}$ in the nearer portion to the solution by a simple experiment based on $\hat{\mathbf{x}}^*_{k/k-1}$, which is performed by obtaining the point satisfying some equations roughly. $C^*_{k/k-1}$ is determined as follows.

$$C^*_{k/k-1}=(\mathbf{x}^*_{k/k-1}-\hat{\mathbf{x}}^*_{k/k-1})(\mathbf{x}^*_{k/k-1}-\hat{\mathbf{x}}^*_{k/k-1})' . \tag{11}$$

This distribution $N(\hat{\mathbf{x}}^*_{k/k-1}, C^*_{k/k-1})$ is the concentrated normal distribution $N(\hat{\mathbf{x}}^*_{k/k-1}, (\mathbf{x}^*_{k/k-1}-\hat{\mathbf{x}}^*_{k/k-1})'(\mathbf{x}^*_{k/k-1}-\hat{\mathbf{x}}^*_{k/k-1}))$ on the line $\hat{\mathbf{x}}^*_{k/k-1}, \mathbf{x}^*_{k/k-1}$ with a point of the inflection $\mathbf{x}^*_{k/k-1}$ and has the probability 0 otherwise. Such distribution means to put the large weight on the interval $[\hat{\mathbf{x}}^*_{k/k-1}, \mathbf{x}^*_{k/k-1}]$ comparatively in searching the solution.

By applying the above method we can obtain the roots of the simultaneous equations recursively.

§ 3. The relation to Newton method.

Now let $\boldsymbol{g}_{(K,D)}(\mathbf{y})$ be given by (12), and then the optimal parameter (K, D) is gained according to the principle of least squares, (13) and unbiased estimate, (14).

$$\boldsymbol{g}_{(K,D)}(\mathbf{y})=\hat{\mathbf{x}}^*_{k/k-1}+K[\mathbf{y}-\boldsymbol{f}(\hat{\mathbf{x}}^*_{k/k-1})]+D , \tag{12}$$

$$E\{(\boldsymbol{g}_{(K,D)}(\boldsymbol{f}(\mathbf{x}))-\mathbf{x})^2\}=\text{Min.} , \tag{13}$$

$$E\{\boldsymbol{g}(\boldsymbol{f}(\mathbf{x}))-\mathbf{x}\}=\mathbf{o} , \tag{14}$$

where K is a matrix and D is a vector. Substituting the optimal parameter gained by the reference [4] into (K, D) in (12) and putting $\mathbf{y}=\mathbf{o}$, the formula (12) is reduced to the formula (7).

In case $n=m$ and the inverse function $\boldsymbol{g}(\mathbf{y})$ of $\boldsymbol{f}(\mathbf{x})$ exists, putting $\boldsymbol{g}(\boldsymbol{f}(\mathbf{x}))$ into $\mathbf{x}$ and $\mathbf{y}$ into $\boldsymbol{f}(\mathbf{x})$ in (13), (14), we gain (15), (16).

$$E\{\boldsymbol{g}_{\langle K,D\rangle}(\mathbf{y})-\boldsymbol{g}(\mathbf{y})\}^2=\text{Min.}, \tag{15}$$

$$E\{\boldsymbol{g}_{\langle K,D\rangle}(\mathbf{y})-\boldsymbol{g}(\mathbf{y})\}=\mathbf{o}. \tag{16}$$

This formula shows that $\boldsymbol{g}_{\langle K,D\rangle}(\mathbf{y})$ is the linear approximation with the weight of the probability distribution of $\boldsymbol{g}(\mathbf{y})$. The mean of the $\boldsymbol{g}_{\langle K,D\rangle}(\mathbf{y})$ is the mean of the $\boldsymbol{g}(\mathbf{y})$. Since $\boldsymbol{g}_{\langle K,D\rangle}(\boldsymbol{f}(\hat{\mathbf{x}}^*_{k/k-1}))\to\boldsymbol{g}(\boldsymbol{f}(\hat{\mathbf{x}}^*_{k/k-1}))$ and $\boldsymbol{g}_{\langle K,D\rangle}(\mathbf{y})$ becomes the tangent plane of $\boldsymbol{g}(\mathbf{y})$ as $C^*_{k/k-1}\to O$, our method approaches to the Newton method as $C^*_{k/k-1}\to O$.

This follows also from (7), that is,

$$\hat{\mathbf{x}}^*_{k/k}\to\hat{\mathbf{x}}^*_{k/k-1}-\boldsymbol{f}_{\mathbf{x}}^{-1}(\hat{\mathbf{x}}^*_{k/k-1})\boldsymbol{f}(\hat{\mathbf{x}}^*_{k/k-1}) \quad \text{as} \quad C^*_{k/k-1}\to O. \tag{17}$$

Thus Newton method is the particular case of our method, and our method is an improvement of Newton method by considering up to second order term of $\boldsymbol{f}(\mathbf{x})$.

In Newton method, it is necessary for convergence that Jacobian $|\boldsymbol{f}_{\mathbf{x}}(\mathbf{x})|\neq 0$ at searching points. In our method, on the other hand, only the weaker condition that the derivative $\boldsymbol{f}_{\mathbf{x}}(\mathbf{x})\neq\mathbf{o}$ is needed, since if $\boldsymbol{f}_{\mathbf{x}}(\mathbf{x})=\mathbf{o}$ in (7), then our procedure stops. The nemerical example 1 by the method of the reference [4] becomes the case when $\boldsymbol{f}_{\mathbf{x}}(\mathbf{x})$ is very small at the searching point unexpectedly.

§ 4. The results of the calculation.

For the simplicity of explanation our examples are restricted to the case when $\mathbf{x}$ and $\boldsymbol{f}$ are two dimensional, but this restriction does not lose any generality.

The solution of simultaneous equations is represented as the intersection points of the following two curves.

$$f_1(\mathbf{x})=0, \tag{18}$$

$$f_2(\mathbf{x})=0. \tag{19}$$

Assuming that optimal estimates have been obtained up to $k-1$ stage, we shall relate the procedure of calculating k stage's optimal estimate (see ref. [3]).

1. If $f_1^2(\hat{\mathbf{x}}^*_{k-1/k-1})+f_2^2(\hat{\mathbf{x}}^*_{k-1/k-1})<\varepsilon$, stop the calculation, where ε is a given small value.
2. Otherwise, exchange f_1 for f_2.
3. Steer $\hat{\mathbf{x}}^*_{k-1/k-1}$ by the one dimensional Newton method to the f_1-curve and the f_2-curve as follows.
   a. Fix the first component of $\hat{\mathbf{x}}^{*1}_{k-1/k-1}$ and search the solution x_{2*} of equation $f_1(\hat{x}^{*1}_{k-1/k-1}, x_2)=0$ with variable x_2 by Newton method.
   b. Put $\hat{\mathbf{x}}^*_{k/k-1}=(\hat{x}^{*1}_{k-1/k-1}, x_{2*})$.
   c. Fix the second component of $\hat{\mathbf{x}}^*_{k-1/k-1}$ and solve the equation, $f_2(x_1, \hat{x}^{*2}_{k-1/k-1})=0$ with variable x_1 by Newton method. Let x_{1*} be the solution.

d. Put $\mathbf{x}_{k^*/k-1}=(x_{1^*}, \hat{x}^{*2}_{k-1/k-1})$.

4. Search the tangent of f_1 curve at $\hat{\mathbf{x}}^*_{k/k-1}$ and let $\mathbf{x}^*_{k^*/k-1}$ be a point on the tangent whose distance from $\hat{\mathbf{x}}^*_{k/k-1}$ is $\sqrt{(\hat{\mathbf{x}}^*_{k/k-1}-\mathbf{x}_{k^*/k-1})}$.
5. Put $C^*_{k/k-1}=(\mathbf{x}^*_{k^*/k-1}-\hat{\mathbf{x}}^*_{k/k-1})(\mathbf{x}^*_{k^*/k-1}-\hat{\mathbf{x}}^*_{k/k-1})'$.
6. Calculate $\hat{\mathbf{x}}^*_{k/k}$ by the formula (7) substituting $C^*_{k/k-1}$ and $\hat{\mathbf{x}}^*_{k/k-1}$.
7. Replace $\hat{\mathbf{x}}^*_{k-1/k-1}$ by $\hat{\mathbf{x}}^*_{k/k}$ and return back 1.

Here initial value $\hat{\mathbf{x}}^*_{0/-1}$ is given on the f_2-curve and $\mathbf{x}_{0^*/-1}$ is on the f_1-curve and the procedure begins (4) above.

In convenience, we will call the above method as method 1. Now we shall show the another method, method 2. This method differs from method 1 in the algorisms from 3b to 4, but is same as method 1 in other algorisms.

1. Search the tangent of f_1-curve at $\hat{\mathbf{x}}^*_{k/k-1}$.
2. Search the intersection point of the tangent and f_2-curve by the Newton method along the tangent.
3. Put intersection point by $\mathbf{x}^*_{k/k-1}$.

EXAMPLE 1. $y_1 = x_1^2 + x_2^2 - 1 = 0$ $y_2 = 2x_1x_2 + x_2^2 = 0$

	Newton method		method 1		method 2		method of ref. [4]	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
1	0.2000×10^{2}	-0.4000×10^{2}	0.2000×10^{2}	-0.4000×10^{2}	0.2000×10^{2}	-0.4000×10^{2}	0.2×10^{2}	-0.4×10^{2}
2	0.1001×10^{2}	-0.20001×10^{2}	0.6581×10	-0.1316×10^{2}	0.7484×10	-0.1497×10^{2}	0.6581×10	-0.1316×10^{2}
3	0.5012×10	-0.1002×10^{2}	-0.5000	0.1000×10	-0.5000	0.1000×10	0.1250×10	-0.500×10
4	0.2526×10	-0.5052×10	-0.4468	0.8936	-0.4476	0.8952	0.2404×10^{-1}	-0.4807×10^{-1}
5	0.13027×10	-0.26054×10	-0.4472	0.8944	-0.4472	0.8944	0.5636×10^{-4}	-0.1127×10^{-3}
6	0.7281	-0.1456					0.1318×10^{-6}	-0.2635×10^{-6}
7	0.45014	-0.90028					0.3081×10^{-9}	-0.6161×10^{-9}
8	0.4472	-0.8944					0.7202×10^{-12}	-0.1440×10^{-11}
9	0.4472	-0.8944					0.1694×10^{-14}	-0.3363×10^{-14}

$F = y_1^2 + y_2^2$

1	0.39960010×10^{7}	0.39960010×10^{7}	0.39960010×10^{7}	0.39960010×10^{7}
2	0.2495×10^{6}	0.4647×10^{5}	0.7786×10^{5}	0.4647×10^{5}
3	0.1553×10^{5}	0.6250×10^{-1}	0.6250×10^{-1}	0.4644×10^{2}
4	0.95532×10^{3}	0.3548×10^{-5}	0.2864×10^{-5}	0.9942
5	0.5624×10^{2}	0.1919×10^{-7}	0.1298×10^{-9}	0.1×10
6	0.2724×10			0.1×10
7	0.6604×10^{-1}			0.1×10
8	0.1725×10^{-3}			0.1×10
9	0.1812×10^{-8}			0.1×10

EXAMPLE 2. $y_1 = x_1^2 + x_2^2 - 1 = 0$ $y_2 = 2x_1x_2 + x_2^4 = 0$

	Newton method		method 1		method 2		method of ref. [4]	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
1	-0.5×10^{3}	0.1×10^{2}	-0.5×10^{3}	0.1×10^{2}	-0.5×10^{3}	0.1×10^{2}	-0.5×10^{3}	0.1×10^{2}
2	-0.2499×10^{3}	0.8333×10	-0.2499×10^{3}	0.8333×10	-0.2499×10^{3}	0.8333×10	-0.2499×10^{3}	0.8333×10
3	-0.1249×10^{3}	0.6823×10	-0.5000	0.1000×10	-0.5000	0.1000×10	-0.1250×10^{3}	0.7467×10
4	-0.6233×10^{2}	0.5533×10	-0.3929	0.9286	-0.3929	0.9286	-0.1125×10^{3}	0.7372×10
5	-0.3102×10^{2}	0.4459×10	-0.3903	0.9207	-0.39024815	0.92071112	-0.1046×10^{3}	0.7309×10
6	-0.1533×10^{2}	0.3575×10	-0.39023651	0.92070470			-0.9897×10^{2}	0.7261×10
7	-0.7451×10	0.2851×10	-0.3902445	0.92071018			$-0\ 9462\times10^{2}$	0.7222×10
8	-0.3475×10	0.2257×10					-0.9110×10^{2}	0.7189×10
9	-0.1470×10	0.1762×10					-0.8816×10^{2}	0.7160×10
10	-0.5142	0.1349×10					-0.8565×10^{2}	0.7134×10
11	-0.2484	0.1048×10					-0.8347×10^{2}	0.7111×10
12	-0.3694	0.9430					-0.8154×10^{2}	0.7089×10
13	-0.3895	0.9215					-0.7982×10^{2}	0.7069×10
14	-0.390245574	0.9207114742					-0.7872×10^{2}	0.7050×10

$F=y_1^2+y_2^2$

	Newton method	method 1	method 2	method of ref. [4]
1	0.62549510×10^{11}	0.6254510×10^{11}	0.62549510×10^{11}	0.625495×10^{11}
2	0.3911×10^{10}	0.3911×10^{10}	0.3911×10^{10}	0.3911×10^{10}
3	0.2448×10^{9}	0.6250×10^{-1}	0.6250×10^{-1}	0.2471×10^{9}
4	0.1538×10^{8}	0.4673×10^{-3}	0.4673×10^{-3}	0.1629×10^{9}
5	0.9769×10^{6}	0.1442×10^{-7}	0.7949×10^{-11}	0.1229×10^{9}
6	0.6385×10^{5}	0.3634×10^{-9}		0.9878×10^{8}
7	0.4483×10^{4}	0.1555×10^{-10}		0.8291×10^{8}
8	0.3664×10^{3}			0.7158×10^{8}
9	0.3808×10^{2}			0.6306×10^{8}
10	0.4871×10			0.5643×10^{8}
11	0.4982			0.5110×10^{8}
12	0.9499×10^{-2}			0.4674×10^{8}
13	0.1157×10^{-4}			0.4309×10^{8}
14	0.212337×10^{-10}			0.3999×10^{8}

EXAMPLE 3. Now let us return back to the cost minimization problem, which is the object of this paper. This cost minimization problem whose minimum value is known, is reduced to solve the following simultaneous equations.

$$J(\mathbf{x})-\alpha=0\,, \qquad (20)$$

$$\partial J/\partial \mathbf{x}=0\,, \qquad (21)$$

where $J(\mathbf{x})$ is the cost, $\dfrac{\partial J}{\partial \mathbf{x}}$ the derivative of the cost and α the minimum value of the cost. Then applying the above method to these equations, we can solve our problem.

The function given by Rosenbrock is as follows.

$$J(x_1, x_2)=100(x_2-x_1^2)^2+(1-x_1)^2\,. \qquad (21)$$

Since this function has a deep helical valley along $x_1^2=x_2$, it is difficult to seek a optimal point by the steepest descent method. The method to solve by the condition that the derivative of J at optimal point is 0, i.e. $\dfrac{\partial J}{\partial x_1}=0$ and $\dfrac{\partial J}{\partial x_2}=0$, is called as method a. The method to solve under the condition $J=0$ in addition to the condition of method a is called as method b, where the covariance is calculated from the method stated in the beginning of Section 4, putting $\dfrac{\partial J}{\partial x_1}$ into f_1 and $\dfrac{\partial J}{\partial x_2}$ into f_2.

The following table shows that method b is faster in convergence than the conjugate gradient method.

$J(\mathbf{x})=100(x_2-x_1^2)^2+(1-x_1)^2$

	method b		method b		method b		method a		
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2	
1	-0.12×10	0.144×10	0.3×10	0.9×01	0.1×10^2	0.1×10^3	0.1×10^2	0.1×10^3	
2	0.1×10	-0.38×10	0.9998	−0.3000	0.9995	-0.8×10^2	0.4517×10	-0.9649×10	the linear search
	0.1×10	0.1×10	0.9998	0.9996	0.9996	0.9991	0.4517×10	0.2041×10^2	
							the oscillation		
	J		J		J		J		
	method b		method b		method b		method a		
1	0.484×10		0.4×10		0.81×10^2		0.81×10^2		
2	0.234×10^4		0.16×10^4		0.65×10^6		0.2987×10^{10}		the linear search
	0.986×10^{-11}		0.3428×10^{-9}		0.19×10^{-8}		0.1515		
	9.05 sec.								

The conjugate gradient method

	x_1	x_2	x_1	x_2	x_1	x_2
1	-0.1200×10	0.1000×10	0.3×10	0.9×10	0.100×10^2	0.1000×10^3
2	-0.1032×10	0.1067×10	0.2569×10	0.6563×10	0.9604×10	0.9216×10^2
3	-0.3947	0.1187	0.2200×10	0.4815×10		
4	-0.1827	-0.3964×10^{-2}	0.1857×10	0.3423×10		
5	-0.9506×10^{-1}	-0.1740×10^{-1}	0.1580×10	0.2488×10		
6	0.3737×10^{-1}	-0.2965×10^{-1}	0.1383×10	0.1902×10		
7	0.1933	0.7594×10^{-2}	0.1170×10	0.1367×10		
8	0.5492	0.2424	0.1038×10	0.1067×10		
9	0.6723	0.4379	0.1002×10	0.1004×10		
10	0.8976	0.7794	0.1000×10	0.1002×10		
⋮						
14	0.1000×10	0.1000×10				
⋮						
20			0.1000×10	0.1000×10		
	J		J		J	
1	0.242×10^2		0.4×10		0.8100×10^2	
2	0.4129×10		0.2609×10		0.7457×10^2	
3	0.2083×10		0.1510×10			
4	0.1538×10		0.8032			
5	0.1269×10		0.3457			
6	0.1023×10		0.1622			
7	0.7394		0.2956×10^{-1}			
8	0.5533		0.1211×10^{-1}			
9	0.1270		0.2059×10^{-4}			
10	0.7928×10^{-1}		0.1143×10^{-5}			
⋮						
14	0.4372×10^{-11}					
⋮						
20			0.8392×10^{-8}			
time (sec)	7.191		69.966		120.009	

Acknowledgment. The author wishes to appreciate with gratitude the very valuable suggestion and guidance she received from Professor Nagata Furukawa.

References

[1] S. Tsuji and K. Kumamaru; *System Indentification and Function Learnings by a Nonlinear Filter,* Memoirs of Fac. of Eng. Kyushu University, Vol. XXX, No. 4, 1971.

[2] R. Fletcher and C. M. Reeves; *Functional Minimization by Conjugate Gradients,* Computer, J. 7. (1964), pp. 149–154.

[3] Y. Shinohara; *A geometric method of numerical solution of nonlinear equations and error estimation by Urabe's proposition,* Publications of the RIMS, Kyoto Univ., Ser. A, Vol. 5, No. 1 (1969), pp. 1–9.

[4] M. Athans, R. P. Wisher, A. Bertolini; *Suboptimal State Estimation for Continuous Time Nonlinear Systems from Discrete Noisy Measurements,* Preprint of J. A. C. C. 1968, pp. 364–382.