Mem. Fac. Sci. Shimane Univ.,
25, pp. 15–19 Dec. 25, 1991

A Modification of the Newton Method from a Viewpoint of Statistical Testing Methods

Tsukio MORITA
Mimasaka Women's College
and
Shinto EGUCHI
Department of Mathematics, Shimane University
(Received September 4, 1991)

The Newton method has quitely universal use for a nonlinear optimization problem, $\max\{f(x): x \in U\}$. The algorithm satisfies to be quadratically convergent under a mild condition. This paper concentrates on the further convergence from a statistical viewpoint. A deformation of the objective function f is proposed by association with methods of statistical testing hypothesis. It is shown that the k-step modified function enjoys convergence of $(k+2)$-th order.

1. Introduction

Let U be an open subset of $\mathbb{R}^d$ and let f be an analytic function on U. Consider a problem of maximization of f over U. We assume that there exists a unique maximizer x^* in this problem. Furthermore, assume that the gradient vector $\nabla f(x)$ vanishes only at x^* and that the Hessian matrix $H_f(x)$ is negative-definite over U. The Newton method introduces a sequence $\{x_p: p = 0, 1, \cdots\}$ defined by

$$x_{p+1} = x_p - H_f^{-1}(x_p)\nabla f(x_p),$$

where the initial point x_0 is appropriately determined. It is known that $\|x_{p+1} - x^*\| \leq c\|x_p - x^*\|^2$ with a constant $c \leq 1$.

In a statistical context we regard the objective function f as a log-likelihood function based a sample with size n. Then in the estimation problem, x^* is nothing but the maximum likelihood estimator. For a problem of testing a hypothesis $x = x_0$ against $x \neq x_0$ the following testing methods have been established (see [3]):

$$\alpha(x_0) \equiv 2\{f(x^*) - f(x_0)\},$$

$$\mathrm{b}(x_0) \equiv -{}^t\nabla f(x_0) H_f^{-1}(x_0)\nabla f(x_0)$$

and

$$\gamma(x_0) \equiv -{}^t(x^* - x_0) H_f(x^*)(x^* - x_0).$$

which are referred to as the likelihood ratio statistic the Rao statistic and the Wald statistic, respectively. It is known that under a mild condition for randomness of observation these statistics have a common random behavior around the null hypothesis as n increases. This property comes from the property that Hessian matrices of $\alpha(x)$, $\beta(x)$ and $\gamma(x)$ are equal to each other when evaluated at $x = x^*$. Of course, one cannot find $\alpha(x)$ and $\gamma(x)$ without knowledge of x^*. Under the above assumptions for f, x^* minimizes simultaneously each of the functions $\alpha(x)$, $\beta(x)$ and $\gamma(x)$.

In this paper the key idea is to pay attention to the fact that x^* is found by only one-iteration in applying $\gamma(x)$ to the Newton method, that is,

$$x_1 = x_0 - H_\gamma(x_0)^{-1} \nabla \gamma^*(x_0) = x_0 - H_f(x^*)^{-1} H_f(x^*)(x_0 - x^*) = x^*.$$

for any initial point $x_0 \in U$. From this point of view we propose a sequence of functions $\{f_k(x): k = 1, 2, \cdots\}$ defined by

$$f_1 = \varphi(f).$$

$$f_k = c_k \varphi(f_{k-1}) + (1 - c_k) f_{k-1} \qquad (k \geq 2),$$

where $c_k = 2/(k+1)(k+2)$ and $\varphi(f)(x) = (1/2)^t \nabla f(x) H_f^{-1}(x) \nabla f(x)$. Note that $f_1 = \beta$ and that the the maximizer of f_k is commonly x^* for all $k \geq 1$. We shall show that $\lim_{k \to \infty} f_k(x) = \gamma(x)$. Thus f_k may become gradually feasible to find x^*, of which property is more exactly stated as follows.

THEOREM. *The Newton algorithm for f_k defined as above,*

$$x_{p+1} = x_p - H_{f_k}^{-1}(x_p) \nabla f_k(x_p) \qquad (p \geq 0)$$

has the $(k+2)$-th order of convergence, that is,

$$\| x_{p+1} - x^* \| \leq c_k \| x_p - x^* \|^{k+2}$$

with the constant $c_k \leq 1$.

In Section 2 we prove the theorem. Section 3 gives a numerical example and discusses the implication of this modification.

2. Proof of Theorem

We keep the notation used in the previous section and fix the objective function f and the optimal point x^* throughout this section. Define a family $\mathscr{F}$ of analytic functions $g: U \to \mathbf{R}$ such that

(A1) $\quad g(x^*) > g(x), \ \nabla g(x) \neq 0 \quad$ for all $x \neq x^*$ in U,

(A2) $\quad H_g(x^*) = H_f(x^*)$ and $H_g(x) < 0 \quad$ for all $x \in U$.

The functions α, β and γ, defined in Introduction, are in $\mathscr{F}$ and further, f_k is also in $\mathscr{F}$ for any $k \geq 1$. Note that $\mathscr{F}$ is a convex set. Next for any $k \geq 1$ we introduce a subclass $\mathscr{F}_k = \{g\colon U \to \mathbf{R}\}$ of $\mathscr{F}$ such that

(A3) $g(x)$ has the i-th order derivative vanishing at x^* for any i, $3 \leq i \leq k+2$.

Clearly $\mathscr{F} \supset \mathscr{F}_1 \supset \mathscr{F}_2 \supset \cdots$ and $\mathscr{F}_k$ is also convex. Note that $\gamma(x)$ is in $\mathscr{F}_k$ for any $k \geq 1$, that is, $\lim_{k\to\infty} \mathscr{F}_k = \{\gamma\}$ from the assumption of analyticity. We write

$$\nabla^* f(x) \equiv H_f^{-1}(x)\,\nabla f(x),$$

so that $\varphi(f)(x) = \langle \nabla f(x), \nabla^* f(x) \rangle$ with the Euclidean inner product $\langle\ ,\ \rangle$.

PROPOSITION 1. *If* $f \in \mathscr{F}_k$, *then*

$$H_f(x^*)\,\partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}\nabla^* f(x^*) = -(k-1)\,\partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}\nabla f(x^*)$$

for $1 \leq i_1,\dots,i_k \leq d$, *where* $\partial_i \equiv \partial/\partial x_i$ *with* $x = (x_1,\dots,x_d)$.

PROOF. By definition, $H_f(x)\,\nabla^* f(x) = \nabla f(x)$. The k-times differentiation of both sides yields

$$\sum_{r=0}^{k} {}_kC_r(\partial_{i_1}\cdots\partial_{i_r}H_f(x))(\partial_{i_{r+1}}\cdots\partial_{i_k}\nabla^* f(x))$$
$$= \partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}\nabla f(x) \tag{2.1}$$

by the Leipnitz law. The substitution of (2.1) into $x = x^*$ leads to

$$H_f(x^*)\,\partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}\nabla f(x^*) + k\partial_{i_1}\partial_{i_2}\cdots\partial_{i_{k-1}}H_f(x^*)\partial_{i_k}\nabla^* f(x^*)$$
$$= \partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}\nabla f(x^*)$$

because all the terms except for $r = 0$ and $r = k-1$ in the *RHS* of (2.1) vanishes at x^*. The result follows from $\nabla^t\nabla^* f(x^*) = I$ (identity matrix). □

REMARK. By a similar argument of the proof, it is seen that if $f \in \mathscr{F}_k$, then

$$\partial_{i_1}\partial_{i_2}\cdots\partial_{i_r}\nabla^* f(x^*) = 0 \qquad (2 \leq r \leq k-1).$$

Furthermore we have the following proposition.

PROPOSITION 2. *For any* $k \geq 3$, *if* $f \in \mathscr{F}_k$ *then*

(a) $\varphi(f) \in \mathscr{F}_k$ *and*

(b) $$\partial_{i_1}\partial_{i_2}\cdots\partial_{i_{k+1}}\varphi(f)(x^*) = -\frac{(k+1)(k-2)}{2}\partial_{i_1}\partial_{i_2}\cdots\partial_{i_{k+1}}f(x^*).$$

PROOF. (a) Differentiating $\varphi(f)$ in $x_{i_1}, x_{i_2},\dots,x_{i_m}$ with $1 \leq m \leq k$, we have

$$\partial_{i_1}\partial_{i_2}\cdots\partial_{i_m}\varphi(f)(x)$$
$$= \sum_{r=0}^{m} {}_mC_r\langle \partial_{i_1}\cdots\partial_{i_r}\nabla f(x),\ \partial_{i_{r+1}}\cdots\partial_{i_m}\nabla^* f(x)\rangle \tag{2.2}$$

which implies

$$\partial_{i_1}\partial_{i_2}\cdots\partial_{i_m}\varphi(f)(x^*) = 0$$

for $m \geq 3$ by noting Remark. This leads to $\varphi(f)\in\mathscr{F}_k$. (b) When $m = k+1$ in (2.2), we have

$$\partial_{i_1}\partial_{i_2}\cdots\partial_{i_{k+1}}\varphi(f)(x^*) = \frac{k+1}{2}\{\langle\partial_{i_1}\partial_{i_2}\cdots\partial_{i_k}\nabla f(x),\ \partial_{i_{k+1}}\nabla^* f(x)\rangle$$
$$+ \langle\partial_{i_1}\nabla f(x),\ \partial_{i_2}\cdots\partial_{i_{k+1}}\nabla^* f(x)\rangle\}.$$

From Proposition 1,

$$\langle\partial_{i_2}\cdots\partial_{i_{k+1}}\nabla^* f(x^*),\ \partial_{i_1}\nabla f(x^*)\rangle = -(k-1)\partial_{i_1}\partial_{i_2}\cdots\partial_{i_{k+1}}f(x^*),$$

which completes the proof.

We now prove Theorem stated in Introduction by using Proposition 2.

PROOF of Theorem. By induction we have $f_k\in\mathscr{F}_{k+2}$, applying Proposition 2. Consequently

$$H_{f_k}(x_p) = H_f(x^*) + 0(\|x_p - x^*\|^{k-1}),$$
$$\nabla f_k(k_p) = H_f(x^*)(x_p - x^*) + 0(\|x_p - x^*\|^k)$$

Hence the application of f_k to the Newton method yields

$$x_{p+1} - x^* = x_p - x^* - H_{f_k}^{-1}(x_p)\nabla f_k(x_p) = x_p - x^*$$
$$-(H(x^*) + 0(\|x_p - x^*\|^{k-1}))^{-1}(H(x^*)(x_p - x^*) + 0(\|x_p - x^*\|^k))$$
$$= 0(\|x_p - x^*\|^k),$$

which completes the proof.

3. A numerical example

Consider the model in which a random variable X has a normal distribution with mean $\beta\in\mathbf{R}$ and variance β^2. In Table, we shall show the iterative process to an M.L.E. of β in the case of using f_1 and f. We can see from Table that the convergence based on f_1 is more rapid than that of f.

Table. Iterative process for f_3 and f
(convergence criterion 10^{-9})

Iteration number	Initial value 0.2 f_1	Initial value 0.2 f	Initial value 0.8 f_1	Initial value 0.8 f
1	.263074761	.259375000	.714815016	.336842105
3	.442097128	.414895124	.618214642	.505940567
4	.546513815	.500646777	.618033989	.575482249
5	.610703543	.571876554		.610870543
6	.618024798	.609677692		.617812095
7	.618033989	.617732985		.618033772
8		.618033590		.618033989
9		.618033989		

4. Discussion

We discuss an effect of change of variables in the optimization problem: $\max\{f(x): x \in U\}$. Let τ be a one-to-one transformation on U. We write $f^{(\tau)}(y) \equiv f(\tau^{-1}(y))$ and hence the optimal point x^* is exactly mapped into $y^* = \tau(x^*)$ because of $f^{(\tau)}(y^*) = f(x^*)$. However in the sequence $\{y_p^{(\tau)}: p \geq 0\}$ generated by applying the Newton method to $f^{(\tau)}$, $\tau(y_p^{(\tau)})$ is not generally equal to x_p for $p \geq 1$ even if if stated from $\tau(x_0^{(\tau)}) = x_0$. Thus the Newton method leads to different courses to the optimal point by change of variables. For example, the ideal property for $\gamma(x)$ discussed in Introduction is lost after any nonlinear change of variables. Conversely there is an approach to detecting variables for rapid convergence, see [3]. In this paper our approach is to deform the original function f in place of the above way. The k-step modification involves the k-times composition of φ, which accompanies with the computation of the $(k+2)$-order derivatives of f. Our method will be more feasible with aid of program package for processing the mathematical formulas.

References

[1] D. R. Cox and D. V. Hinkley. Theoretical Statistics. Chapman and Hall (1974), London.

[2] S. Eguchi. A differential geometric approach to statistical inference on the basis of contrast functionals. Hiroshima Math. J. (1985) **15**, 341–391.

[3] T. Morita. A Note on parameter estimation using sequential reparametrization algorithm. Bulletin of Mimasaka Women's College (1985).

[4] J. M. Ortega and W. G. Rheinbaldt. Iterative solution of nonlinear equations in several variables. Academic Press (1970), New York.