Orthomin(k) Method for Linear Least Squares Problem

ZHANG SHAO LIANG*† and YOSHIO OYANAGI**††

A class of iterative algorithms for solving large-scale linear least squares problems is proposed. The algorithms are based on the conjugate residual direction, and are well suited to linear least squares problems where the coefficient matrix is large and sparse. The linear convergence and the error bounds of the process are proved, and the convergence condition is analyzed.

1. Introduction

Recently several methods have been proposed for solving large systems of linear equations

$$Ax=b \tag{1}$$

with a non-symmetric coefficient matrix [1, 2, 3]. Their technique is based upon minimizing the Euclidean norm of the residual $\|b-Ax\|$ over the space $x_{i-k}+\mathrm{SPAN}\{p_i, p_{i-1}, \ldots, p_{i-k}\}$, where the iterate x_i is the i-th approximation and p_i is the i-th direction vector.

In the present paper we extend this method to a linear least squares problem of minimizing $\|b-Ax\|$, where A is a large sparse $m\times n$ rectangular matrix. We assume that the coefficient matrix is so large that the amount of work and storage required in direct methods such as QR and singular value decomposition is nearly prohibitive. A common technique for solving such least squares problems is to apply the conjugate gradient method [4] to the normal equation

$$A^TAx=A^Tb, \tag{2}$$

where the coefficient matrix A^TA is symmetric and positive-definite. On the $(i+1)$-th iteration, the conjugate gradient method gives an optimal approximate solution in some sense over a Krylov subspace $x_0+\mathrm{SPAN}\{A^Tr, (A^TA)A^Tr, \ldots, (A^TA)^{i-1}A^Tr\}$, where $r=b-Ax_0$. Since the condition number of A^TA is the square of that of A, this dependence on A^TA tends to make the convergence slow.

The method we present depends on a Krylov subspace based on BA rather than A^TA, where B is an appropriately chosen $n\times m$ matrix, which we call a mapping matrix. It is expected to improve the convergence of iterative methods in terms of a choice of an appropriate B as free parameters. Such methods require that the symmetric part of AB be positive semi-definite. If B is close to a generalized inverse of A, the convergence would be fast. The mapping matrix B plays a significant role, similar to that of a preconditioner in solving large sparse linear equations by the conjugate residual method [3]. A different algorithm based on a similar Krylov space has been presented for the case in which $m=n$ [5], where B is given in the form of C^{-1}. We note that the sum of squares itself has a definite statistical meaning and should not be changed by a preconditioning such as $\|B(b-Ax)\|$.

In the next section, we present an Orthomin(k) type algorithm for the least squares problem. In Section 3, we show the convergence conditions and the rate of decrease of the residuals. In Section 4, we discuss the choice of B. In Section 5, we give some numerical examples that show the usefulness of our algorithm. Section 6 is the conclusion.

2. CR-LS(k) Methods

A linear least squares problem is to obtain the minimum point $\bar{x}\in R^n$ of the objective function with n variables

$$S(x)=\|b-Ax\|^2, \tag{3}$$

where A is an $m\times n$ matrix and b is a vector of dimension m.

The conjugate residual method starts with an initial point x_0, and minimizes $S(x)$ successively along a line $x_i+\alpha p_i$, where p_i is a vector of dimensional n, called a correction vector, and α is a constant. In linear squations (A is a square matrix), p_i is chosen to be r_i plus a linear combination of former correction vectors, p_{i-1}, p_{i-2}, . . . , where r_i is the i-th residual vector. In our case, however, the dimension of the m-vector r_i is

*Doctoral Program in Engineering, University of Tsukuba, Tsukuba-shi, Ibaraki 305, Japan.
**Institute of Information Sciences, University of Tsukuba, Tsukuba-shi, Ibaraki 305, Japan.
†Now at Inst. of Computational Fluid Dynamics, Hara-machi 2-1-4, Meguro, Tokyo 152, Japan.
††Now at Department of Information Science, University of Tokyo, Hongo, Tokyo 113, Japan.

different from that of the n-vector p_i, and therefore we need a mapping matrix B, which maps an m-vector to an n-vector.

Adding the mapping matrix B to Orthomin(k) [2], we derive the following algorithm:

$$r_0=b-Ax_0, \quad p_0=Br_0$$

for $i=0$ to max_i until convergence do

$$\alpha_i=(r_i, Ap_i)/(Ap_i, Ap_i)$$

$$x_{i+1}=x_i+\alpha_i p_i$$

$$r_{i+1}=r_i-\alpha_i Ap_i \tag{4}$$

for $j=0$ to $\min(k-1, i)$ do

$$\beta_{i,i-j}=-(ABr_{i+1}, Ap_{i-j})/(Ap_{i-j}, Ap_{i-j})$$

$$p_{i+1}=Br_i+\sum_{j=0}^{\min(k-1,j)} \beta_{i,i-j}p_{i-j}$$

Here α_i is determined in such a way as to minimize the new residual vector $\|r_i-\alpha Ap_i\|$ as a function of α in the direction p_i, and $\beta_{i,i-j}$ are chosen in such a way as to make p_{i+1} A^TA-orthogonal to the last k vectors-$\{p_j\}_{i-k+1}^{i}$. We will call this algorithm the CR-LS(k) method. The number k may be 0, 1, 2, . . . , depending on the characteristics of the problem. The work vectors necessary to implement CR-LS(k) are x, r, ABr, and k sets of p and Ap. To minimize the multiplication by A, Ap_i is also updated by

$$Ap_{i+1}=ABr_i+\sum_{j=0}^{\min(k-1,j)} \beta_{i,i-j}Ap_{i-j}. \tag{5}$$

The residual and correction vectors obey the following relations due to the construction of p's:

Theorem 1.

a) $(Ap_i, Ap_j)=0 \qquad 0<|i-j|\le k, \ i\ge k;$ (6a)

b) $(r_i, Ap_j)=0 \qquad 0<i-j<k, \ i\ge k+1;$ (6b)

c) $(r_i, Ap_i)=(r_i, ABr_i);$ (6c)

d) $(r_i, ABr_j)=0 \qquad 0<i-j<k;$ (6d)

e) $(r_j, Ap_i)=(r_{i-k}, Ap_i) \ 0\le i-j\le k, \ i\ge k.$ (6e)

3. Convergence Properties

In the previous section, we did not specify the mapping matrix B. The most trivial choice would be $B=A^-$, where A^- is a generalized inverse matrix of A, that is, where $AA^-A=A$ and $(AA^-)=(AA^-)$ [6]. In this case the first step of algorithm (4) would be

$$p_0=Br_0=A^-(b-Ax_0)$$

$$(r_0, Ap_0)=(b-Ax_0, AA^-b-Ax_0)=(Ap_0, Ap_0) \tag{7}$$

$$\alpha=1$$

$$x_1=x_0+A^-b-A^-Ax=A^-b-(1-A^-A)x_0, \tag{8}$$

so that x_1 gives one of the least squares solutions. This choice is unrealistic, since if we knew A^-, which is difficult to compute, we would simply compute A^-b without applying any iterative methods.

There is a certain trade-off between the number of iterations and the cost of computing Br. The more B resembles A^-, the faster the method will converge. On the other hand, the cost of computing Br at each iteration may become large if B is close to A^-. We will discuss the rate of convergence of the method and the requirement for the mapping matrix B.

We first define the projection P onto Im (A), which can be written as AA^- by using a generalized inverse [6]. The matrix P obeys the relation $P=P^T=P^2$ and $PA=A$. We can also write $P=QQ^T$, using the QR decomposition of A, where Q is an $m\times\rho$ matrix (ρ=rank A) with orthonormal column vectors.

We first prove the following lemma:

Lemma 1.

$$(Ap_i, Ap_i)\le(ABr_i, ABr_i). \tag{9}$$

Proof

If $k=0$, Eq. (9) is an identity. For $k>0$, the correction vector p_i is given by Eq. (4) in the form

$$p_i=Br_i+\sum_{j=0}^{\min(k-1,i-1)} \beta_{i,i-j}p_{i-j}.$$

Then, from properties (6a) and (6b), we have

$$\begin{aligned}(Ap_i, Ap_i)&=(ABr_i, ABr_i)+2\sum_{j=0}^{\min(k-1,i-1)} \beta_{i,i-j}(ABr_i, Ap_{i-j})\\&\quad+\sum_{j=0}^{\min(k-1,i-1)} \beta_{i-j,i-j}^2(ABr_{i-j}, Ap_{i-j})\\&=(ABr_i, ABr_i)-\sum_{j=0}^{\min(k-1,i-1)} (ABr_i, Ap_{i-j})^2/\\&\quad(Ap_{i-j}, Ap_{i-j})\\&\le(ABr_i, ABr_i).\end{aligned}$$

QED

To provide an error bound for CR-LR(k), we now present the following main theorem:

Theorem 2.

Let $\{r_i\}$ be a sequence of residuals in algorithm (4); the following inequality then holds:

$$\frac{\|r_{j+1}-\bar{r}\|^2}{\|r_i-\bar{r}\|^2}\le 1-\frac{\lambda_{\min}^2(M)}{\lambda_{\max}(M)\lambda_{\min}(M)+\rho(R)^2}, \tag{10}$$

provided that

(a) $BP=B$ and

(b) $M\equiv Q^T(AB+B^TA^T)Q/2$ is positive definite,

where $\bar{r}=b-A\bar{x}$, $R=Q^T(AB-B^TA^T)Q/2$, $\lambda_{\min}$ and $\lambda_{\max}$ are the maximum and minimum eigenvalues of M, and $\rho(R)$ is the spectral radius of R.

Proof

The proof runs parallel to that of ref. [3]. The displacement from the minimum residual $\bar{r}$ is the projection of r to Im (A) as

$$r_i-\bar{r}=A\bar{x}-Ax_i=AA^-b-Ax_i=Pb-PAx_i=Pr_i.$$

The ratio is bounded from above if assumption (b)

holds:

$$\frac{\|r_{i+1}-\bar{r}\|^2}{\|r_i-\bar{r}\|^2}=\frac{\|Pr_{i+1}\|^2}{\|Pr_i\|^2}=1-\frac{(r_i, Ap_i)^2}{(Ap_i, Ap_i)(r_i, Pr_i)}$$
$$\leq 1-\frac{(r_i, ABr_i)}{(ABr_i, ABr_i)}\cdot\frac{(r_i, ABr_i)}{(r_i, Pr_i)}. \quad (11)$$

Here Lemma 1 has been used.

Next the positivity of the second term of the RHS of the Eq. (11) will be proved. We now estimate the first factor in the second term of the RHS of Eq. (11). Assumption (a), that $BP=BQQ^T=B$, enables us to estimate the quantities in the subspace Im (A). We suppress the subscript i.

$$(r, ABr)/(ABr, ABr)=(Pr, ABPr)/(ABPr, ABPr)$$
$$=(QQ^Tr, ABQQ^Tr)/$$
$$(ABQQ^Tr, ABQQ^Tr),$$
$$=(C^{-1}y, y)/(y/y)$$
$$\geq\lambda_{\min}((C^{-1}+C^{-T})/2) \quad (12)$$

where

$C=Q^TABQ$ (from assumption (b), C is non-singular) and $y=Q^TABQQ^Tr$. Using the relation $X^{-1}+Y^{-1}=(X(X+Y)^{-1}Y)^{-1}$, we have

$$\lambda_{\min}[(C^{-1}+C^{-T})/2]=\lambda_{\min}[((2C)^T(4M)^{-1}(2C))^{-1}]$$
$$=\lambda_{\min}[((M-R^TM^{-1}(M-R))^{-1}]$$
$$=\lambda_{\min}[(M+R^TM^{-1}R)]$$
$$=\lambda_{\max}(M+R^TM^{-1}R)^{-1}$$

But,

$$\lambda_{\max}(M+R^TM^{-1}R)=\max_{x\neq 0}\left[\frac{(x, Mx)}{(x, x)}+\frac{(x, R^TM^{-1}Rx)}{(x, x)}\right]$$
$$\leq\lambda_{\max}(M)+\max_{x\neq 0,\, Rx\neq 0}\frac{(Rx, M^{-1}Rx)}{(Rx, Rx)}$$
$$\times\frac{(Rx, Rx)}{(x, x)}$$
$$\leq\lambda_{\max}(M)+\lambda_{\max}(M^{-1})\|R^TR\|_2$$
$$=\lambda_{\max}(M)+\lambda_{\min}(M)^{-1}\rho(R)^2.$$

Thus

$$\lambda_{\min}[(C^{-1}+C^{-T})/2]\geq(\lambda_{\max}(M)+\lambda_{\min}(M)^{-1}\rho(R)^2)^{-1}. \quad (13)$$

However, the second factor can be transformed

$$(r, ABr)/(r, Pr)=(Pr, ABPr)/(Pr, Pr)$$
$$=(Q^Tr, Q^TABVQ^Tr)/(Q^Tr, Q^Tr)$$
$$\geq\lambda_{\min}(M). \quad (14)$$

Combining Eqs. (12), (13), and (14), we have Eq. (10). This completes the proof. QED.

This theorem shows that the CR-LS(k) method is at least linearly convergent.

4. Choice of B

The CR-LS(k) algorithm covers a wide class of methods that differ in the choice of the mapping matrix B and the parameter k. The particular choice of B critically depends on the application and cannot be discussed in general. We will restrict ourselves to a few general comments.

The simplest choice of B that automatically satisfies the two conditions (a) and (b) in Theorem 3 is $B=A^T$. In this case, C is symmetric and the convergence rate is controlled by $1-\lambda_{\min}(C)/\lambda_{\max}(C)$. As may be expected, this case is equivalent to the conjugate gradient method for the normal equations (2), and CR-LS(k) ($k\geq 1$) is equivalent to CR-LS(1).

Next, we consider what is meant by the convergence condition (a) $BP=B$.

Lemma 2

There is an $n\times n$ matrix D such that $B=DA^T$, provided $BP=B$.

Proof

Write $A=(a_1, \ldots, a_n)$ and $B^T=(b_1, \ldots, b_n)$. Then, since $PA=A$, and $\text{rank}(1-P)=m-\text{rank}(A)$, we have $\text{Kerl}(1-P)=\text{Im}(A)=\text{SPAN}\{a_1, \ldots, a_n\}$.

We can write $x=\sum_{i=1}^n d_ia_i$ provided that we know every root of the linear equation $(1-P)x=0$.

Since $(1-P)B^T=0$, there is some scalar number d_{ij} such that

$$b_i=\sum_{j=1}^n d_{ij}a_j, \quad i=1, 2, \ldots, m.$$

Therefore, $B^T=AD^T$, $B=DA^T$. QED.

We consider a family of mapping matrices in the form

$$B=DA^T, \quad (15)$$

and the condition imposed on D. In practice, A is a large sparse matrix, so that multiplying a vector by A^T from the left will not be too time-consuming. The matrix D should not have too complex a structure. If Eq. (15) holds, condition (a) is automatically satisfied for any D. If the symmetric part of D is positive definite, condition (b) is also satisfied, since

$$2M=Q^T(ADA^T+AD^TA^T)Q=Q^TA(D+D^T)A^TQ. \quad (16)$$

We have to make the condition number of M as small as possible. The extreme choice would be to set D equal to $(A^TA)^{-1}$. In this case, B is again a generalized inverse of A. If the columns of A are approximately orthogonal, we may take D as the inverse of the diagonal part of (A^TA). Incomplete Cholesky decomposition [7] of (A^TA) will also be applicable. The practical choice of D will be discussed elsewhere.

5. Numerical Results

We present here some numerical results obtained by applying the above CR-LS(k) method to the next linear

Table 1 A comparative result in scalar mode ($N=201$, $n=101$).

Methods		α 0.005	0.005	0.05	0.1	0.2
Gram-Schmidt (ms)		3275	3275	3275	3275	3275
CG	iterations	23	23	23	22	20
	CPU time (ms)	15	15	15	14	13
CR-L S(1)	iterations	12	12	12	12	15
	CPU time (ms)	11	11	11	11	11

Table 2 A comparative result in vectorization mode ($N=201$, $n=101$).

Methods		α 0.005	0.005	0.05	0.1	0.2
Gram-Schmidt (ms)		116	116	116	116	116
CG	iterations	23	23	23	22	20
	CPU time (ms)	3	3	3	3	3
CR-L S(1)	iterations	12	12	12	12	15
	CPU time (ms)	2	2	2	2	2

Table 3 Comparative results in vectorization mode ($N=2001$, $n=1001$).

Methods		α 0.005	0.005	0.05	0.1	0.2
CG	iterations	23	23	23	22	19
	CPU time (ms)	9	9	9	9	8
CR-L S(1)	iterations	12	12	12	13	15
	CPU time (ms)	5	5	5	5	6

least squares problem, which appears in curve fitting based on the Bayes procedure estimation [8].

In this curve fitting problem, let n be number of observations, and N the number of partitions. Then, the coefficient matrix of the least squares problem is given by

$$A_\alpha=\begin{bmatrix} E \\ \alpha D \end{bmatrix},$$

where E is a $n\times N$ fit-point matrix defined by

$$E=\begin{bmatrix} & & & & & & & & & & & & j \\ 1 & & & & & & & & & & & & \cdot \\ & 1 & & & & & & & & & & & \cdot \\ & & 0 & 1 & & & & & & & & & \cdot \\ & & & & 0 & 0 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & & & & & & & & \cdot \\ & & & & & & & & & & & & \cdot \\ & & & & & & & & & & & & \cdot \\ & & & & & & & & & & & 0 & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ i \\ \\ \\ \\ \\ \end{matrix}$$

D is a $(N-2)\times N$ second order differential matrix defined by

$$D=\begin{bmatrix} 1 & -2 & 1 & & & & & & & \\ & 1 & -2 & & 1 & & & & & \\ & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \\ & & & & 1 & & -2 & 1 & & \\ & & & & & & 1 & -2 & 1 & \end{bmatrix}$$

and α is a scalar, called a super-parameter in the Bayes procedure.

When $N\times n$ and $\alpha<0.5$, we set D in Eq. (15) equal to $\mathrm{diag}(A^TA)^{-1}$ to approximate the extreme choice $(A^TA)^{-1}$. We now compare the convergence behaviour of the Orthomin(1) method, which does not use the normal equation, and the CG method when those methods are applied to the normal equation for solving the above least squares problem. We also compare the results of using the CR-LS(1) method as an iterative method and the modified Gram-Schmidt orthogonalization as a direct method in terms of the CPU time. The computation was performed on a Fujitsu VP-200 with double precision.

The initial approximation x_0 is taken from the uniform random number, and the stopping criterion is $\|x_i-\bar{x}\|\leqq 10^{-7}$, where $\bar{x}$ is the exact solution obtained by modified Gram-Schmidt orthogonalization.

When the size of the coefficient matrix is 300×201, in the scalar mode, the results shown in Table 1 confirm that the CPU time for executing the CR-LS(1) method algorithm is far schorter than that for the modified Gram-Schmidt orthogonalization as a direct method, and also show that the CR-LS(1) method is superior to the CG method applied to the normal equation in terms of the number of iterations and the CPU time. In the vectorization mode, Table 2 similarly shows that the CR-LS(1) method is more efficient than the other two.

We will compare only the CR-LS(1) method and the CG method for large-scale problems.

When the size of the coefficient matrix is 3000×2001, the CR-LS(1) reduces the CPU time and the number of iterations for various α in vectorization mode (Table 3). The CPU time for the modified Gram-Schmidt orthogonalization is 97 seconds in this case.

We demonstrate the usefulness of the CR-LS(1) method for solving large-scale least squares problems in our numerical examples by our choice of matrix B.

6. Conclusion

We have shown that the Orthomin(k) method as a conjugate residual type can be extended to the linear least squares problem, and have indicated the computational advantages of using a mapping matrix B. We have also presented a proof for the rate of convergence of this method, and analyzed the construction of B that satisfies conditions (a) and (b). One disadvantage may be that it is not possible to obtain the variance matrix $(A^TA)^{-1}$ with this algorithm. Since the method is applied to large-scale problems where n is very large, it is not realistic to calculate the variance matrix. However,

this disadvantage should not be taken too seriously. It is simple to extend the method to the weighted least squares problem with non-diagonal weight.

In the numerical examples we have shown that methods of the type presented in this paper are efficiently performed on supercomputers with vector facilities.

Acknowledgement

We would like to thank K. Tanabe, M. Mori, M. Natori, and M. Sugihara for stimulating discussions and comments. Our work was supported in part by the Grants-in-Aid for Scientific Research of the Ministry of Education, Science and Culture (No. 61540142).

References

1. CONCUS, P. and GOLUB, G. H. A generalized conjugate gradient method for nonsymmetric systems of linear equations. Lecture notes in economics and mathematical systems, **134**, eds. R. Glowinski and J. L. Lions, Springer-Verlag, Berlin (1976), 56–65.
2. VINSOME, P. K. W. Orthomin, an iterative method for solving sparse sets of simultaneous linear equations. Proc. Fourth symposium on reservoir simulation, Society of Petroleum Engineers of AIME (1976), 149–159.
3. EISENSTAT, S. L., ELMAN, H. C. and SCHULTS, M. H. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Num. Anal.* **20** (1983), 345–357.
4. HESTENES, H. R. and STIEFEL, E. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, **49** (1952), 409–436.
5. AXELSSON, O. A generalized conjugate gradient, Least square method. *Numer. Math* **51** (1987), 209–227.
6. RAO, C. R. and MITRA, S. K. Generalized inverse of matrices and its application. Wiley, New York (1971).
7. MEIJERINK, J. A. and VAN DER VORST, H. A. An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. *Math. Comput.* **31** (1977), 148–162.
8. KASHIWAGI, N. A Bayes estimation procedure for fertilities in field experiments. Research Memorandum, No. 220 (1982), Inst. Statist. Math.

(Received June 27, 1989; revised February 19, 1990)