Numerical Analysis Project
Manuscript NA-87-04

March 1987

The Convergence Rate of Inexact Preconditioned Steepest Descent Algorithm for Solving Linear Systems

Hans Munthe-Kaas

Numerical Analysis Project
Computer Science Department
Stanford University
Stanford, California 94305

March 1987

The Convergence Rate of Inexact Preconditioned Steepest Descent Algorithm for Solving Linear Systems

by

Hans Munthe-Kaas*

* Norwegian Institute of Technology
N-7034 Trondheim-NTH
Norway
This work was supported in part by the NSF (DCR 84-12314).

The Convergence Rate of Inexact Preconditioned Steepest Descent Algorithm for Solving Linear Systems

by

Hans Munthe-Kaas*

Abstract:

The steepest descent algorithm is a classical iterative method for solving a linear system Ax=b, where A is a positive definite symmetric matrix. A common way to accelerate an iterative scheme is precondition the method, i.e. to solve a simpler system Mz=r in each stage of the iteration. We analyze the effect of solving the preconditioner inexactly. A lower bound for the convergence rate is derived, and we show under what conditions this lower bound is obtained. Finally we describe some numerical experiments which shows that in practical situations the lower bound may be too pessimistic. An amusing result is that in some cases small errors may lead to higher convergence rates than if the preconditioner is solved exactly!

1 Introduction

In most cases when an iterative algorithm is used to solve a linear system, it is necessary to use a preconditioning technique to get fast convergence, i.e. in each step of the algorithm we solve the system $Mz = r$, where M is an approximation for the matrix A. In practical problems it is often difficult or expensive to solve this equation exactly. It may be also be advantageous to use an “inner” iteration scheme to solve the preconditioned system. In these cases it is of great importance to know how exact the inner system must be solved to get a desired rate of convergence. Analysis of this kind is performed for the Chebychev and the Richardson methods in Golub and Overton (81) and in Golub and Overton (86). In this paper we study the effect of inexact preconditioning on the steepest descent algorithm. This method is not much used in practice-because of its slow convergence. The algorithm may however be used as a starting point for developing more sophisticated methods, such as the conjugate gradient algorithms.

For the conjugate gradient algorithm, the effect of inexact preconditioning is not well understood. A result of the same kind for this algorithm would be of great importance. The analysis in this paper may provide some insight into this much more difficult problem. The interesting point of the analysis is the way some generalized Kantorovitch inequalities relatively simply leads to non trivial bounds for the convergence rate.

* Norwegian Institute of Technology
N-7034 Trondheim-NTH
Norway

2.1 The Inexact Preconditioned Steepest Descent Algorithm

Let A be a positive definite, symmetric matrix. The system Ax=b is solved by the following iterative process:

For $k = 0,1,2...$ do:

$r_k = b - Ax_k$

$Mz_k = r_k + q_k$; where q_k is an arbitrary vector satisfying $\|q_k\|_2 \le \delta \|r_k\|_2$ (2.1)

$\alpha_k = \frac{(z_k, r_k)}{(z_k, Az_k)}$; where $(\cdot,\cdot)$ is the l-2 inner product

$x_{k+1} = x_k + \alpha_k z_k$

The matrix M is called the pre-conditioner of the system. We assume that also M is positive definite, symmetric. It should be chosen so that the system $Mz=c$ is easy to solve within the accuracy 6. The convergence rate of the algorithm depends on the condition number of $M^{-1}A$, i.e. of

$$\kappa(M^{-1}A) = \text{cond}(M^{-1}A) \equiv \frac{\lambda_{max}(M^{-1}A)}{\lambda_{min}(M^{-1}A)} \quad (2.2)$$

Where λ_{max} and λ_{min} are the largest and smallest eigenvalues of $M^{-1}A$.

2.2 Derivation of the algorithm

Let z_k be an arbitrary direction. We want to find the "best" approximation of the solution along this vector, i.e. we let

$$x_{k+1} = x_k + \alpha_k z_k \quad (2.3)$$

where α_k is the number that minimizes the ratio

$$R = \frac{(r_{k+1}, A^{-1} r_{k+1})}{(r_k, A^{-1} r_k)} \quad (2.4)$$

From (2.3) we get the following recurrence for the residuals:

$$r_{k+1} = b - Ax_{k+1} = r_k - \alpha_k A z_k \quad (2.5)$$

If we define $\Phi(\alpha)$ by

$$\Phi(\alpha) \equiv (r_{k+1}, A^{-1} r_{k+1}) = (r_k, A^{-1} r_k) - 2\alpha(z_k, r_k) + \alpha^2 (z_k, Az_k) \tag{2.6}$$

we get

$$\frac{\partial(\Phi(\alpha))}{\partial(\alpha)} = -2(z_k, r_k) + 2\alpha(z_k, Az_k) \tag{2.7}$$

So R is minimized for

$$\alpha_k = \frac{(z_k, r_k)}{(z_k, Az_k)} \tag{2.8}$$

For this value of α we get

$$R = 1 - \frac{(z_k, r_k)^2}{(z_k, Az_k)(r_k, A^{-1} r_k)} \tag{2.9}$$

If we let Mz = r, we see that for M = A we get R = 0, so the convergence rate is getting higher the better M approximates A. When the preconditioner is solved inexactly, i.e. when we solve

$$Mz_k = r_k + q_k \text{ ; where } \|q_k\|_2 \le \delta \|r_k\|_2 \tag{2.10}$$

we get the algorithm (2.1).

2.3 Convergence rate of the algorithm (2.1)

To get a lower bound for the convergence rate we need a lower bound for the ratio $\frac{(z_k, r_k)^2}{(z_k, Az_k)(r_k, A^{-1} r_k)}$ when z_k satisfies (2.10). The ratio is related to the famous inequality of Kantorovitch (see Householder(72)). This-inequality is however not general enough, but Bauer & Householder(61) supplies the necessary generalizations. They give the following corollary:

Corollary 1 **(Bauer & Householder):**

For any non-null vectors **u,v** and **pos. def. symm.** A

$|v^H u| \geq \|v\| \cdot \|u\| \cdot \sin\phi \;;\; 0 \leq \phi \leq \frac{\pi}{2}$

implies

$$\frac{u^H A u \cdot v^H A^{-1} v}{u^H u \cdot v^H v} \leq \frac{[(\kappa+1) + (\kappa-1)\cdot\cos\phi]^2}{4\kappa} \;;\; \text{where } \kappa = \text{cond}(A)$$

Suppose p and q are two vectors satisfying $\frac{|p^T q|}{\|p\| \cdot \|q\|} \geq \cos\theta$. **If we let** $\phi = \frac{\pi}{2} - \theta$ **in the above** corollary, we get:

$$\frac{(p,q)^2}{(p,p)\cdot(q,q)} \cdot \frac{(p,p)\cdot(q,q)}{(p,Ap)\cdot(q,A^{-1}q)} \geq \frac{4\kappa\cos^2\theta}{[(\kappa+1) + (\kappa-1)\sin\theta]^2} =$$

$$\left(\frac{2}{\frac{\kappa^{\frac{1}{2}}\cdot(1+\sin\theta)}{\cos\theta} , \frac{\kappa^{-\frac{1}{2}}\cdot(1-\sin\theta)}{\cos\theta}} \right)^2$$

But $\frac{1+\sin\theta}{\cos\theta} = \frac{\cos\theta}{\text{i-she}} = \sqrt{\frac{1+\sin\theta}{\text{i-sine}}}$, so we get the following lemma:

Lemma2

Suppose $\frac{|(p,q)|}{\|p\| \cdot \|q\|} \geq \cos\theta \;;\; 0 \leq \theta \leq \frac{\pi}{2}$

Then $$\frac{(p,q)^2}{(p,Ap)\cdot(q,A^{-1}q)} \geq \left(\frac{2}{\left(\kappa\frac{1+\sin\theta}{1-\sin\theta}\right)^{\frac{1}{2}} + \left(\kappa\frac{1+\sin\theta}{1-\sin\theta}\right)^{-\frac{1}{2}}} \right)^2$$

Where κ = **cond(A).**

We also need another result from Bauer **&** Householder:

Theorem 3 (Bauer & Householder):

$$|y^H x| \le \|x\|\cdot\|y\|\cdot\cos\phi \; ; \; 0 \le \phi \le \frac{\pi}{2}$$

implies $|(Ay)^H Ax| \le \|Ax\|\cdot\|Ay\|\cdot\cos\psi$ where $\cot\frac{\psi}{2} = \kappa\cdot\cot\frac{\phi}{2}$

If we apply this theorem with A-1, do the change of variables: y = Au , x = Av, and finally take the contrapositive of this theorem, we get the following result:

Lemma 4

$$\frac{|(u,v)|}{\|u\|\cdot\|v\|} \ge \cos\psi \text{ implies}$$

$$\frac{(Au)^T Av}{\|Au\|\cdot\|Av\|} \ge \cos\phi \qquad \text{where } \tan\frac{\phi}{2} = \kappa\cdot\tan\frac{\psi}{2} \; ; \quad \kappa = \mathrm{cond}(A)$$

Provided $0 \le \phi \le \frac{\pi}{2}$, i.e $0 \le \kappa\cdot\tan\frac{\psi}{2} < 1$

Back to our original problem. From (2.10) we have Mz = r + q = v, so (2.9) gives us:

$$R = 1 - \frac{(v^T M^{-1} r)^2}{(v^T M^{-1} A M^{-1} v)(r^T A^{-1} r)} \tag{2.11}$$

We want to bound the second term away from zero. Instead of supposing a bound on the norm of q, as in (2.10) it is easier for the analysis to assume a bound on the angle between r and v. Assume that

$$\frac{|(v,r)|}{\|v\|\cdot\|r\|} \ge \cos\psi \tag{2.12}$$

If we do the change of variables:

$$\left.\begin{aligned} p &= \sqrt{M^{-1}}\cdot v \\ q &= \sqrt{M^{-1}}\cdot r \\ W &= \sqrt{M^{-1}}\cdot A\cdot\sqrt{M^{-1}} \end{aligned}\right\} \tag{2.13}$$

We get

$$\frac{(v^T M^{-1} r)^2}{(v^T M^{-1} A M^{-1} v)(r^T A^{-1} r)} = \frac{(p,q)^2}{(p,Wp)\cdot(q,W^{-1}q)} \tag{2.14}$$

To apply lemma 2 we need the angle between p and q. From lemma 4 we get :

$$\frac{|(p,q)|}{\|p\|\cdot\|q\|} = \frac{|(\sqrt{M^{-1}}v)^T(\sqrt{M^{-1}}r)|}{\|\sqrt{M^{-1}}v\|\cdot\|\sqrt{M^{-1}}r\|} \geq \cos\theta \tag{2.15}$$

where $\tan\frac{\theta}{2} = \kappa_2 \cdot \tan\frac{\psi}{2}$ and $\kappa_2 = \sqrt{\mathrm{cond}(M)}$, provided $0 \leq \kappa_2 \tan\frac{\psi}{2} < 1$.

Now lemma 2 gives:

$$\frac{(p,q)^2}{(p,Wp)\cdot(q,W^{-1}q)} \geq \left(\frac{2}{\left(\kappa_1 \frac{1+\sin\theta}{1-\sin\theta}\right)^{\frac{1}{2}} + \left(\kappa_1 \frac{1+\sin\theta}{1-\sin\theta}\right)^{-\frac{1}{2}}} \right)^2 \tag{2.16}$$

Where $\kappa_1 = \mathrm{cond}(W) = \mathrm{cond}(\sqrt{M^{-1}} \cdot A \cdot \sqrt{M^{-1}}) = \mathrm{cond}(M^{-1}A)$.

Trigonometric manipulation gives:

$$\frac{1+\sin\theta}{1-\sin\theta} = \left(\frac{1+\tan\frac{\theta}{2}}{1-\tan\frac{\theta}{2}}\right)^2 = \left(\frac{1+\kappa_2\tan\frac{\psi}{2}}{1-\kappa_2\tan\frac{\psi}{2}}\right)^2 \tag{2.17}$$

It is easy to verify that

$$1 - \left(\frac{2}{\kappa^{\frac{1}{2}} + \kappa^{-\frac{1}{2}}}\right)^2 = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2$$

So we finally get the result that the bound for the ratio in (2.4) is:

$$\sqrt{R} \leq \frac{\kappa' - 1}{\kappa' + 1} \text{ ; where } \kappa' = \kappa_1 \cdot \left(\frac{1+\kappa_2\tan\frac{\psi}{2}}{1-\kappa_2\tan\frac{\psi}{2}}\right)^2 \tag{2.18}$$

Now we can state the main theorem of this paper:

Theorem 5 (Convergence rate inexact preconditioned steepest descent)

Let the system $Ax = b$ be solved by the iterative process (2.1).
Let $v_k = r_k + q_k$, and assume that $\frac{|(v_k, r_k)|}{\|v_k\| \cdot \|r_k\|} \geq \cos\psi$.

Then $\left(\frac{(r_{k+1}, A^{-1}r_{k+1})}{(r_k, A^{-1}r_k)}\right)^{\frac{1}{2}} \leq \frac{\kappa' - 1}{\kappa' + 1}$

where $\kappa' = \kappa_1 \cdot \left(\frac{1 + \kappa_2 \tan\frac{\psi}{2}}{1 - \kappa_2 \tan\frac{\psi}{2}}\right)^2$; $\kappa_1 = \text{cond}(M^{-1}A)$ and $\kappa_2 = \sqrt{\text{cond}(M)}$

provided $0 \leq \kappa_2 \tan\frac{\psi}{2} < 1$

2.3 Proof that the lower bound in theorem 5 is optimal.

The result in theorem 5 is obtained by two consecutive transformations:

i) The transformation in (2.13) and (2.15) where both r and v are transformed by $\sqrt{M^{-1}}$.

ii) The transformation in (2.14) and (2.16) where p and q are transformed by $\sqrt{W}$ and $\sqrt{W^{-1}}$ respectively.

We show that it is possible to get equality in both these transformations:

Lemma 6:

If $M = \begin{pmatrix} \kappa_2 & 0 \\ 0 & \kappa_2^{-1} \end{pmatrix}$; $r = \begin{pmatrix} \cos\frac{\psi}{2} \\ \sin\frac{\psi}{2} \end{pmatrix}$; $v = \begin{pmatrix} \cos\frac{\psi}{2} \\ -\sin\frac{\psi}{2} \end{pmatrix}$

we get equality in (2.15).

Proof:

$$q = \sqrt{M^{-1}} \cdot r = \begin{pmatrix} (\kappa_2)^{-\frac{1}{2}} \cos\frac{\psi}{2} \\ (\kappa_2)^{\frac{1}{2}} \sin\frac{\psi}{2} \end{pmatrix} \; ; \quad \text{and } p = \sqrt{M^{-1}} \cdot v = \begin{pmatrix} (\kappa_2)^{-\frac{1}{2}} \cos\frac{\psi}{2} \\ -(\kappa_2)^{\frac{1}{2}} \sin\frac{\psi}{2} \end{pmatrix}$$

so $\tan\frac{\theta}{2} = \kappa_2 \tan\frac{\psi}{2}$; where θ is the angle between q and p.

♦

Now we have $q = \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix}$ and $p = \begin{pmatrix} \cos\frac{\theta}{2} \\ -\sin\frac{\theta}{2} \end{pmatrix}$ so we must find a W that gives equality in (2.16):

Lemma 7:

$$\text{If } W = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} (\kappa_1)^{\frac{1}{2}} & 0 \\ 0 & (\kappa_1)^{-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \; ; \; q = \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix} \; ; \; p = \begin{pmatrix} \cos\frac{\theta}{2} \\ -\sin\frac{\theta}{2} \end{pmatrix}$$

and if $\theta < \frac{\pi}{2}$, then we get equality in (2.16).

Proof:

The reason why this lemma holds is that q and p are placed symmetrically around $\mathbf{x}_1 + \mathbf{x}_2$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the largest and the smallest eigenvectors of W. The lemma is easiest shown by first rotating the coordinate system by 45°. Then the result is obtained by noting that the left hand side of (2.16) is the cosine of the angle between $\sqrt{W} \cdot p$ and $\sqrt{W^{-1}} \cdot q$. We also need some trigonometric identities. We omit the details.

♦

Both W and M are p.d. symmetric, so the same must be true for $A = \sqrt{M} \cdot W \cdot \sqrt{M}$. $\mathrm{Cond}(W) = \kappa_1$ and $\mathrm{cond}(M) = \kappa_2^2$, so given κ_1, κ_2 and ψ, we can always construct an example with equality in theorem 5. Hence we have the following result:

Theorem 8:

> The lower bound in theorem 5 is the best possible result based only on the information: cond(M), cond($M^{-1}A$) and the angle between v and r.

The lower bound is however quite pessimistic, because to obtain it we need a special alignment between the eigenvectors of M and **A:** The largest and the smallest eigenvectors of both M and W must lie in the same plane, and they must be rotated 45^o with respect to each other. To get a better result, we must use more information, but it is not easy to see how this can be done. If we for example knew that:

$$\frac{(v^T M^{-1} r)^2}{(v^T M^{-1} v)(r^T M^{-1} r)} \geq \cos^2\theta \,, \tag{2.19}$$

we could have substituted $\kappa_2 \tan\frac{\psi}{2}$ with $\tan\frac{\theta}{2}$ in theorem 5. This would give a much more realistic lower bound, but to find θ we must know $M^{-1}r$, i.e. we must solve the preconditioner exactly.

Furthermore, for practical purposes it is more interesting to have a lower bound based on δ rather than ψ. Although an obtainable upper bound on ψ is easy to get from δ, this makes the theoretical lower convergence rate even more remote from what we will see in real situations.

3 Numerical experiments

The model problem studied is the following:

$$-u_{xx} = 1; \; u(0) = u(1) = 0 \tag{2.20}$$

Discretized with $h = \frac{1}{n+1}$ the system of equations becomes:

$$\begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} (n+1)^{-2} \\ \vdots \\ \vdots \\ (n+1)^{-2} \end{bmatrix} \tag{2.21}$$

Whenever nothing else is said, the preconditioner M is taken to be the identity matrix.

In each stage of the iteration, the residual is added a random vector, scaled to have norm $\delta\|r_k\|$:

$$z_k = r_k + q_k \cdot \delta \frac{\|r_k\|}{\|q_k\|} \tag{2.22}$$

q_k is a random vector with uniform distribution in the n-cube, i.e. each component of q is uniformly distributed in [-1,1].

The initial vector is chosen randomly in such a way that the initial residual r_0 is a random vector with uniform distribution in the n-cube.

By <u>average convergence rate</u> between iteration <u>m and n</u>, we mean:

$$\text{cnvrte}(m,n) = -\frac{1}{2(n-m)} \ln\left(\frac{(r_n, A^{-1}r_n)}{(r_m, A^{-1}r_m)}\right) \tag{2.23}$$

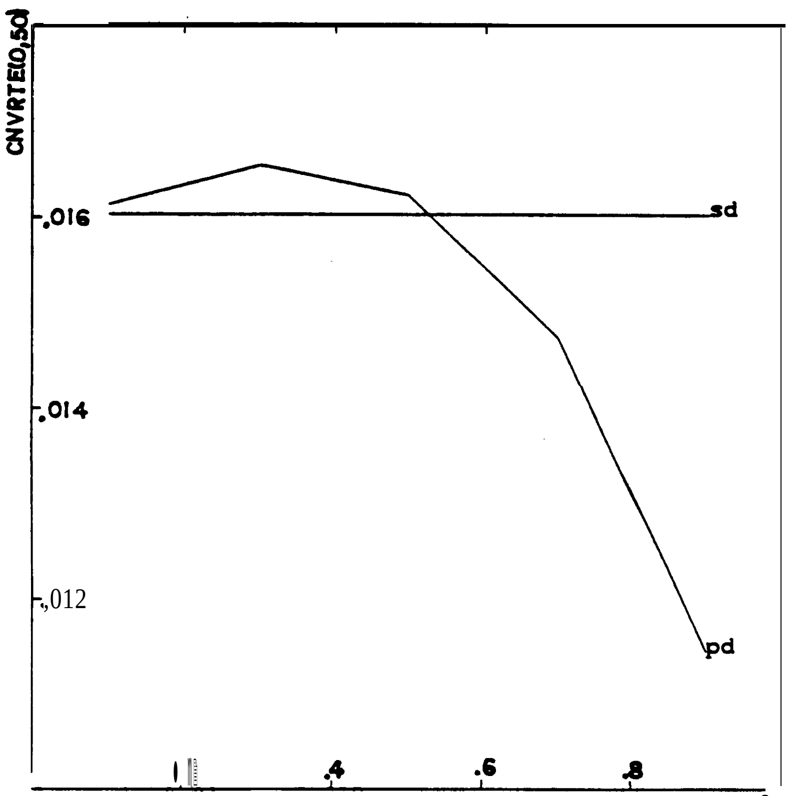


Fig. 6
Average convergence rate over the first 50 iterations for n = 20.

Sd = steepest descent (δ = 0)

Pd = inexact sd, perturbated with different values of 6.

Notice that for small values of 8, pd performs slightly better than sd! This effect is seen in many of the experiments, although usually not as much as here.

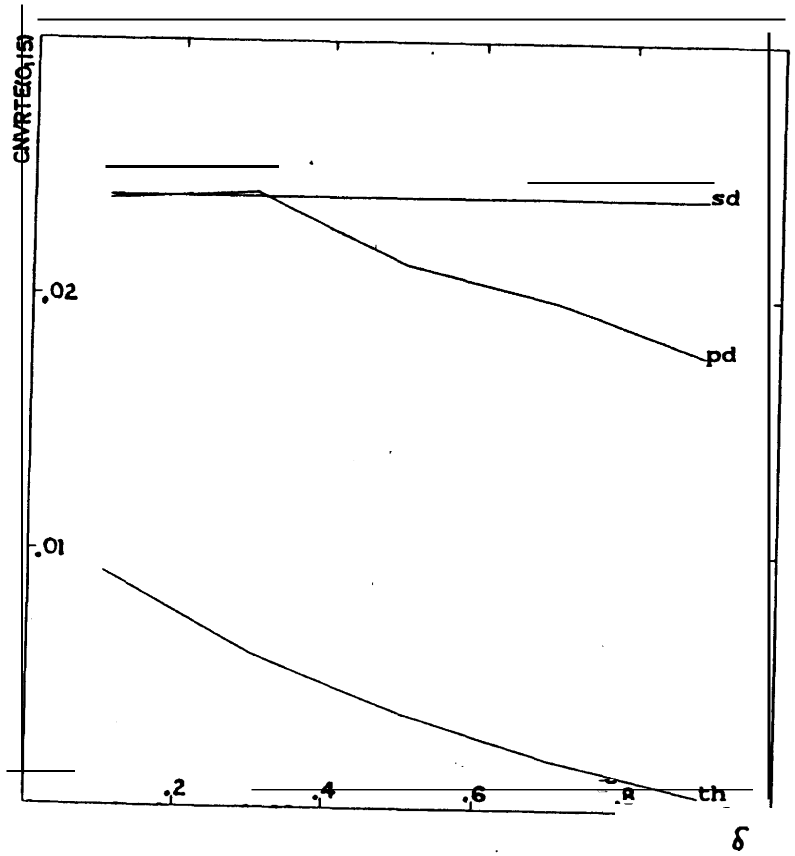


Fig. 2
Average convergence rate over the first 15 iterations for n = 20.

Sd and pd as above.
Th = theoretical lower bound for convergence rate from theorem 5.

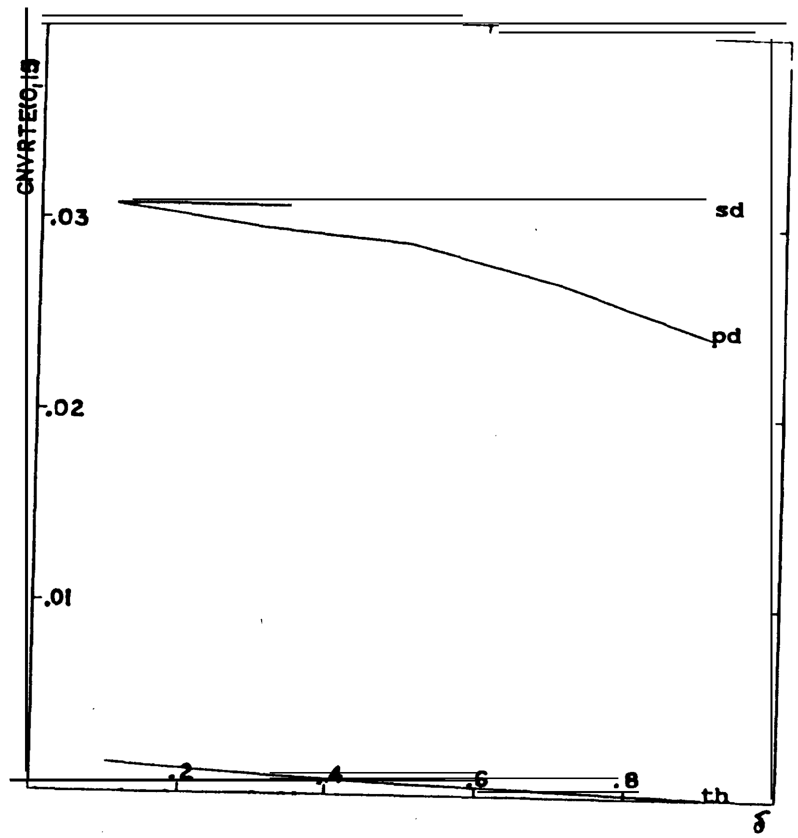
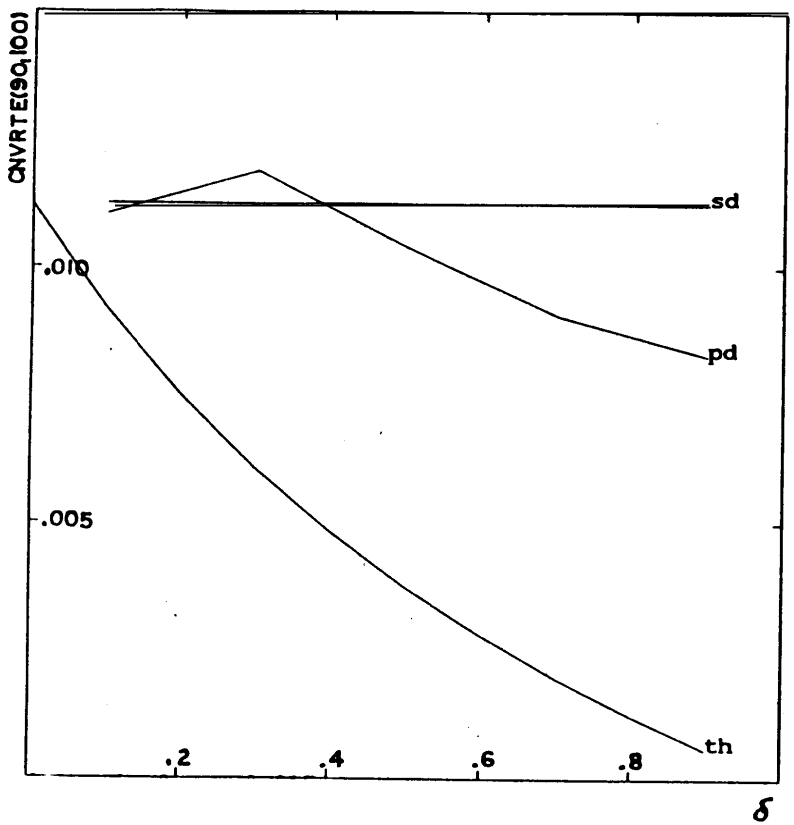


Fig. 3
Average convergence rate over the first 15 iterations for n = 50.

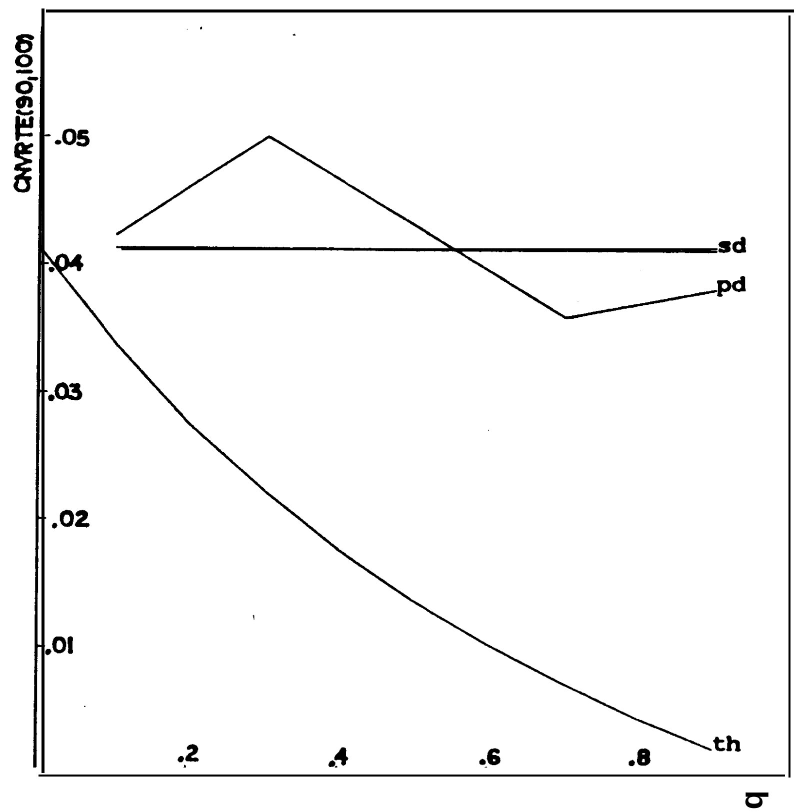
Sd, pd and th as above.

Both sd and pd is performing vastly better than their theoretical lower limit. This is partly because this is initial performance. The convergence rate drops considerably at later stages in the iteration.



<u>Fig. 4</u>
Asymptotic performance for n = 20.
Convergence rate between iteration 90 and 100.

Sd is now down to approximately its worst case. Pd is still performing a lot better than in the worst case.



<u>Fig. 5</u>
Asymptotic performance for n = 10. Convergence rate between iteration 90 and 100.

Notice that for small δ pd is still performing considerably better than sd.

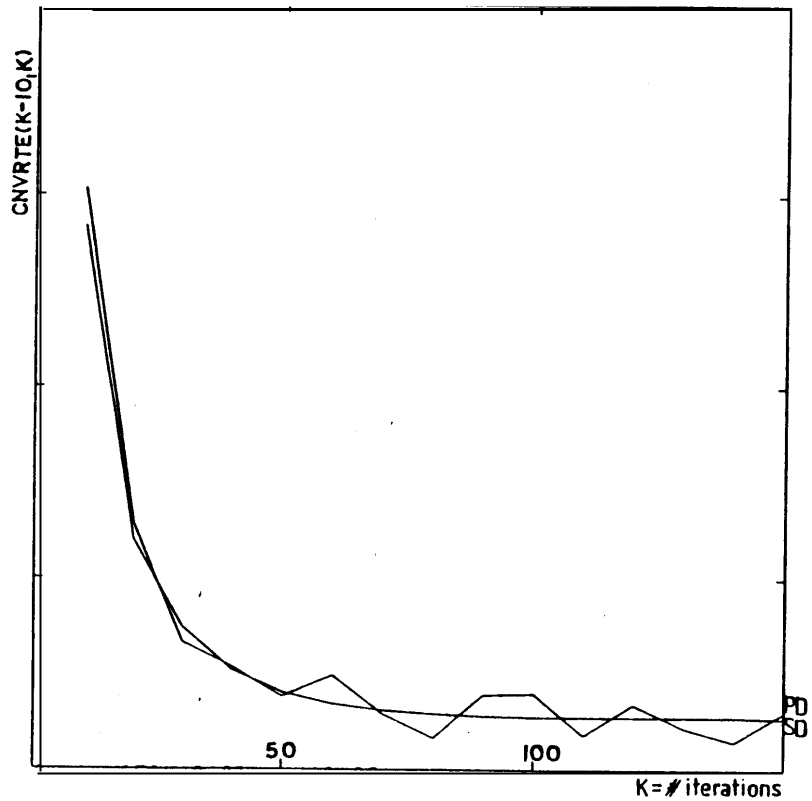


Fig. 6
Convergence rate as function of the number of iterations, k (average convergence rate between iteration k-10 and k). n = 20, $\delta = 0.2$

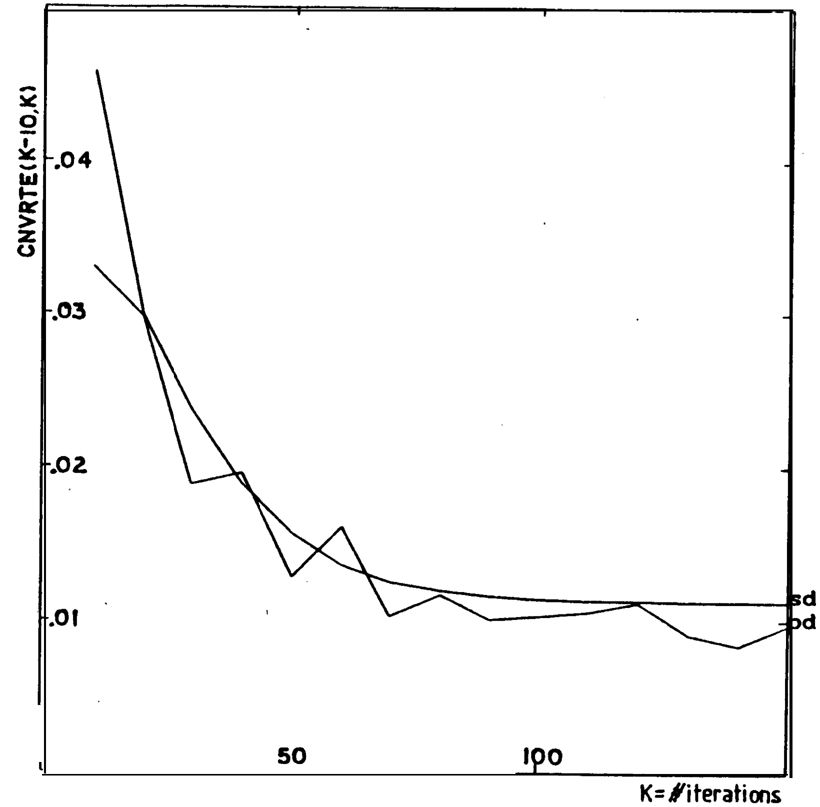


Fig. 7
The same as fig. 6, but here with $\delta = 0.7$

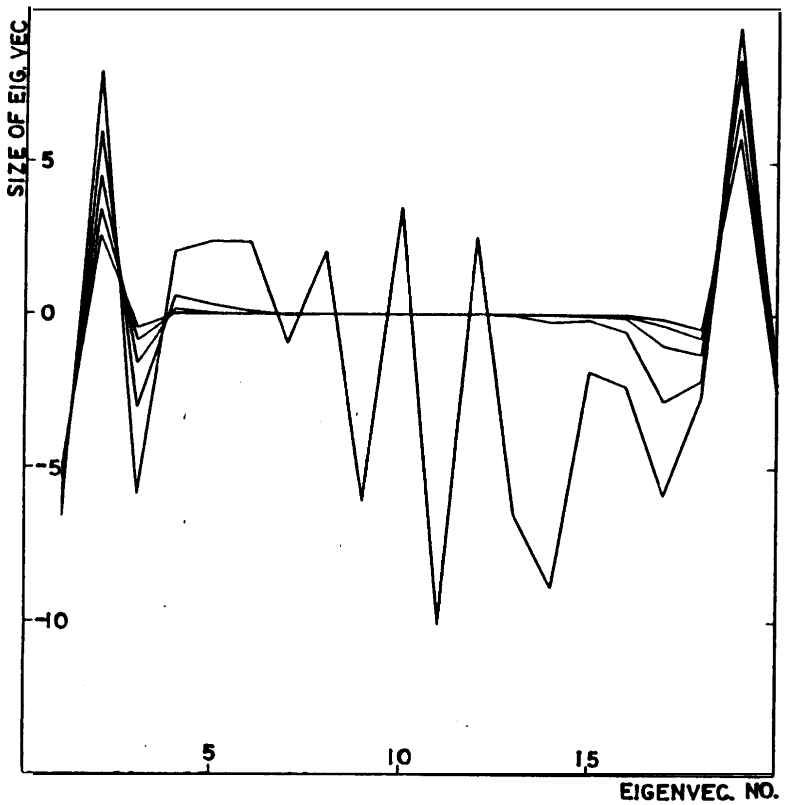


Fig. 8
Fourier spectrum for sd, n=20 after 0, 6,12,18 and 24 iterations. Slowest oscillating component to the left.

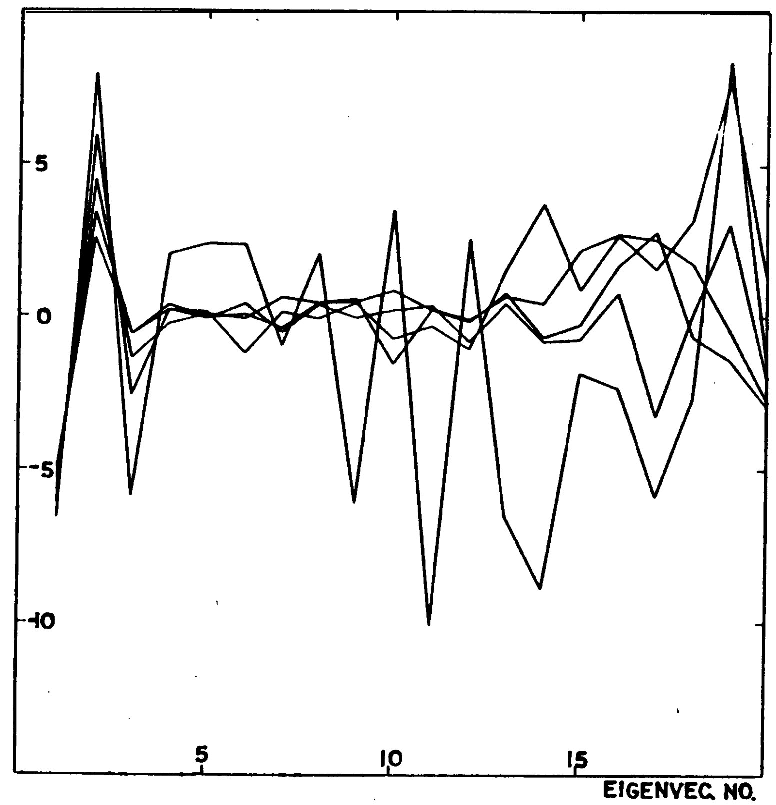


Fig. 9
Same as fig. 8, but here with pd, $\delta = 0.3$.

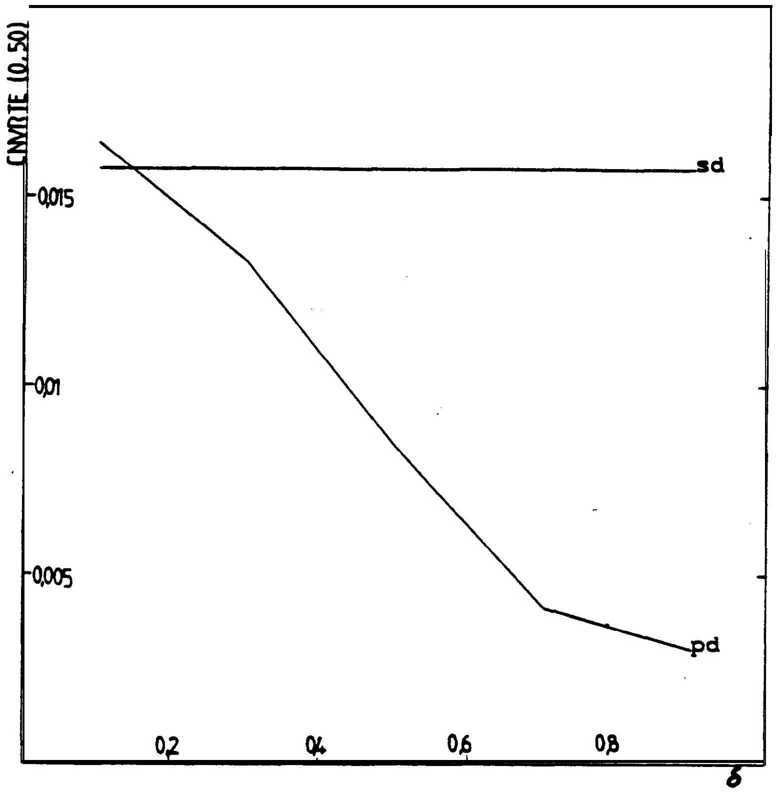


Fig.
The effect of preconditioning. Average convergence rate over the **first 50** iterations. n = 20. The coeffkient matrix A **and** the **preconditioner** M is:

$$A=\begin{bmatrix}1 & & & & \\ & 2 & & & \\ & & 3 & & \\ & & & \cdot & \\ & & & & n\end{bmatrix}\begin{bmatrix}2 & -1 & & & \\ -1 & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & -1 \\ & & & -1 & 2\end{bmatrix}\begin{bmatrix}1 & & & & \\ & 2 & & & \\ & & 3 & & \\ & & & \cdot & \\ & & & & n\end{bmatrix}$$

$$\text{and } \sqrt{M}=\begin{bmatrix}1 & & & & \\ & 2 & & & \\ & & 3 & & \\ & & & \cdot & \\ & & & & n\end{bmatrix}$$

In this example **theorem** 5 does not say anything about the **convergnce** rate when $\delta > 0.1$, because in this case

$$\kappa_2 \cdot \tan\frac{\psi}{2} > 1.$$

4 Acknowledgements

Professor Gene Golub; Stanford, suggested this project. It was his idea to analyze the effect of inexact preconditioning by using Kantorowich inequalities. I whish to thank him for his helpful comments, for his enthusiasm and for his friendliness during my visit to Stanford in the winter -87.

5 References

Bauer F.L, Householder A.S. : Some inequalities involving the euclidean condition number of a matrix. Numerische Mathematik 2, 308-311 (1961)

Bauer F.L : A further generalization of the Kantorovic inequality. Numerische Mathematik 3, 117-l 19 (1961).

Golub G.H., Overton M.L. : Convergence of Two-Stage Richardson Iterative Procedure for Solving Systems of Linear Equations. Springer Lecture Notes in Mathematics no. 912, Proceedings Dundee 198 1.

Golub G.H., Overton M.L. : The Convergence of Inexact Chebychev and Richardson Iterative Methods for Solving Linear Systems. Preprint (December 1986).

Householder A.S. : Lectures on Numerical Algebra. Mathematical Association of America (1972)