DIAGONAL SCALING AS PRECONDITIONER FOR PRECONDITIONED BICGSTAB METHODS TO SOLVE LINEAR EQUATIONS FROM SADDLE-POINT PROBLEMS

AGAH D. GARNADI¹

In this note we will discuss the construction of a diagonal preconditioner for linear equation :

$$\hat{C}x = b$$

where:

$$\hat{C} = \begin{pmatrix} M & A \\ A^T & 0 \end{pmatrix}$$

is an $(n + m) \times (n + m)$ non-singular symmetric indefinite matrix. This form of linear equations arise from Linear Programming as the Karush-Kuhn-Tucker necessary condition and augmented least squares problem [ADR89] also from finite element discretization of linear elliptic equation using mixed finite element formulation as discussed in authors thesis [Ga93]. This problem is well known as the saddle-point problem.

We may solved saddle-point problem in an equivalent form by permuting the block such that the matrix will be of the form :

$$C = \begin{pmatrix} A^T & 0 \\ M & A \end{pmatrix}$$

So the preconditioned BiCGSTAB and its variant can be used without modification.

This note is arranged as follows. The first section concerns a particular type of diagonal preconditioner in relation with the condition number of C . In the second section, we discuss the distribution of eigenvalues of C and its relationship with extreme singular-values of M and A . The result of the second section will be used to estimate the rate of convergence of BiCGSTAB algorithm when used to solve the saddle-point problem, this is discussed in the third section. The result of this section will be used in the fourth section to build a preconditioner to accelerate the iterative method. Finally, the last section gave the range of diagonal preconditioner for saddle-point problem based on the discussion previously.

1. Conditioning by Scaling of Saddle-point Problem. In this section, we consider conditioning of a specific form of saddle-point problem, where M is identity matrix.

Let :

$$C = \begin{pmatrix} I_n & A \\ A^T & 0 \end{pmatrix}$$

be a non-singular matrix, where I_n is $n \times n$ identity and A is $n \times m$ matrix with $m \leq n$ and $\text{rank}(A) = m$. This form of matrix arises from the augmented least

¹ Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor, INDONESIA

squares problem [ADR89]. The inverse of C can be expressed in terms of generalized inverse of A :

$$A^+ = (A^T A)^{-1} A^T$$

and the orthogonal projection matrix on the null space of A^T ,

$$P = I_n - AA^+.$$

Then we can write:

$$C^{(-1)} = \begin{pmatrix} P & (A^+)^T \\ A^+ & -(A^T A)^{-1} \end{pmatrix}$$

The spectral condition number of C is :

$$\kappa(C) = \|C\|_2 \|C^{-1}\|_2$$

is strictly related to σ_1 and σ_m , maximum and minimum singular-values of A respectively. Bjorck [Bj67] showed this relationship is given by:

$$\kappa(C) = \frac{\frac{1}{2}(1 + \sqrt{(1 + 4\sigma_1^2)})}{\min(1, \frac{1}{2}(\sqrt{(1 + 4\sigma_m^2)} - 1))}$$

So that $\kappa(C)$ is much larger than the condition number of A , $\kappa(A) = \frac{\sigma_1}{\sigma_m}$ when σ_1 is much less than 1. In order to decrease the condition number of C , Bjork[Bj67] propose scaling as follows:

$$\begin{pmatrix} I_n & 0 \\ 0 & \alpha^{-1} I_m \end{pmatrix} \begin{pmatrix} I_n & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \alpha I_n & 0 \\ 0 & I_m \end{pmatrix} = \begin{pmatrix} \alpha I_n & A \\ A^T & 0 \end{pmatrix}$$

where α is a real positive number. The condition number of :

$$C_\alpha = \begin{pmatrix} \alpha I_n & A \\ A^T & 0 \end{pmatrix}$$

$\kappa(C_\alpha)$ is given by:

$$\kappa(C_\alpha) = \frac{\frac{1}{2}(\alpha + \sqrt{(\alpha^2 + 4\sigma_1^2)})}{\min(\alpha, \frac{1}{2}(\sqrt{(\alpha^2 + 4\sigma_m^2)} - \alpha))}$$

and attains its minimum of $\frac{1}{2}(1 + \sqrt{(1 + 8\kappa(A)^2)})$ for $\alpha = \frac{\sigma_m}{\sqrt{2}}$

Note that every original saddle-point problems can be transformed into the augmented least squares problem. Since M is symmetric positive definite, then M has Cholesky factorization : $M = LL^T$. Hence :

$$\begin{pmatrix} L^{-1} & 0 \\ 0 & \alpha^{-1} I_m \end{pmatrix} \begin{pmatrix} M & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \alpha L^{-T} & 0 \\ 0 & I_m \end{pmatrix} = \begin{pmatrix} \alpha I_n & L^{-1} A \\ (L^{-1} A)^T & 0 \end{pmatrix}$$

2. Eigenvalue Distribution of Sadddle-point Problem. Since the original linear system arising from saddle point problems is symmetric indefinite and non-singular, then the spectrum contains both positive and negative eigenvalues. In this section we give a result concerning the spectrum of C and its relation to properties

of the sub-matrices M and A. This result will be used to analyze the construction of diagonal preconditioner of linear system.

Let M have eigenvalues $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n > 0$ Then :

$$\mu_n \|x\|^2 \leq \langle Mx, x \rangle \leq \mu_1 \|x\|^2, \text{ for } x \in R^n$$

Let A have singular-values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m > 0$

$$\sigma_m \|y\| \leq \|Ay\| \leq \sigma_1 \|y\|, \text{ for } y \in R^m$$

and

$$\sigma_m \|x\| \leq \|A^T x\| \leq \sigma_1 \|x\|, \text{ for } x \in N(A^T)^\perp$$

The following lemma gives a connection between the spectrum of C and eigenvalues of M and singular-values of A.

Lemma[RW92]

Let $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n > 0$ be eigenvalues of M and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m > 0$ be the singular-values of A, and let Λ denote the spectrum of C. Then:

$$\Lambda \subset I = I^+ \cup I^-$$

where

$$I^- = [\frac{1}{2}(\mu_n - \sqrt{(\mu_n^2 + 4\sigma_1^2)}), \frac{1}{2}(\mu_1 - \sqrt{(\mu_1^2 + 4\sigma_m^2)})]$$

and

$$I^+ = [\mu_n, \frac{1}{2}(\mu_1 + \sqrt{(\mu_1^2 + 4\sigma_1^2)})]$$

Proof

Let $\lambda \in \Lambda$ and let (x, y) be the corresponding eigenvector, i.e:

$$A^T x = \lambda y$$

$$Mx + Ay = \lambda x$$

Since $\text{rank}(A) = m$, if $x = 0$ then $y = 0$. So it is impossible that this pair is an eigenvector of C. So $x \neq 0$. Now we divide the proof into two parts. First we compute the bound for the positive eigenvalue, then the negative eigenvalue.

Let λ be a positive eigenvalue. By taking an inner product of the second equation with x and combine it with an inner product of the first equation with y we obtain:

$$\langle Mx, x \rangle + \lambda \|y\|^2 = \lambda \|x\|^2$$

applying bounds of eigenvalue of M we get :

$$(\mu_n - \lambda) \|x\|^2 \leq -\lambda \|y\|^2 \leq 0$$

Since $\|x\|$ is positive, we have $\lambda \geq \mu_n$. To get an upper bound, we use the first equation and inner product of the second equation with x to get:

$$\langle Mx, x \rangle + \frac{1}{\lambda} \|B^T x\|^2 = \lambda \|x\|^2$$

By the bound of inner product (i) and (iii) we get an implication of an inequality :

$$(\lambda^2 - \mu_1 \lambda - \sigma_1^2) \|x\|^2 \leq 0$$

Since $\|x\|$ is positive, we can conclude that : $\lambda \leq \frac{1}{2}(\mu_1 + \sqrt{(\mu_1^2 + 4\sigma_1^2)})$. So we have derived the bounds of positive eigenvalues.

Next, we consider λ be negative eigenvalue of C . The lower bound of I^- can be derived following the proof for upper bound of I^+ . Hence it is only suffice to derive the upperbound for the negative eigenvalue. To Obtain this, Let $x = v + w$, where $v \in N(A^T)^\perp$ and $w \in N(A^T)$ By the first equation and th inner product of the second equation by v we get :

$$\langle Mw, v \rangle = \lambda \|x\|^2 - \langle Mv, v \rangle - \frac{1}{\lambda} \|B^T v\|^2$$

Applying the bound of eigenvalues of M and singular-values of A we obtain:

$$\langle Mw, v \rangle \geq (\lambda - \mu_1 - \frac{\sigma_m^2}{\lambda}) \|v\|^2$$

To proceed we must bound $\langle Mw, v \rangle$. This is obtained by calculating inner product of (ii) with w and using eigenvalues bound of M . Since $w \in N(B^T)$ and $\lambda - \mu_n$ is negative we get :

$$\langle Mw, w \rangle \leq (\lambda - \mu_n) \|w\| \leq 0$$

By the symmetry of M and combines it with the previous inequality, we obtain:

$$0 \leq (\lambda^2 - \mu_1 \lambda - \sigma_m^2) \|v\|$$

If $v = 0$, then $y = 0$ and (ii) reduce to $Mw = \lambda w$. But this contradict to the fact that λ is negative, hence it follows that $\lambda^2 - \mu_1 \lambda - \sigma^2 \geq 0$. Hence :

$$\lambda \leq \frac{1}{2}(\mu_1 - \sqrt{(\mu_1^2 + 4\sigma_1^2)})$$

. q.e.d.

3. Rate Convergence Estimate of BiCGSTAB for Saddle-point Problem. Assume that $I = I^- \cup I^+$ as stated in previous Lemma. From the choice of parameter ω in BiCGSTAB methods, we can derive a convergence estimate of BiCGSTAB for saddle-point problem. Recall that we choose ω such that $r^{(k)} = b - Cx^{(k)}$ is minimum in 2-norms, then :

$$\begin{aligned} \|r^{(k)}\|^2 &= \|b - Cx^{(k)}\|^2 \\ &\leq \|b - Cu\|^2, \text{ for } u \in K(r^{(0)}, C, k) \end{aligned}$$

Using spectral theory [RW92] we obtain :

$$\begin{aligned} \|r^{(k)}\| &\leq \delta_k(\Lambda(C)) \|b\| \\ &\leq \delta_k(I) \|b\| \end{aligned}$$

where :

$$\delta_k(\Lambda(C)) = \inf_{p \in \Pi_k^1} \sup_{\lambda \in \Lambda} |p(\lambda)|$$

Observe that the convergence can be estimated from the quantity $\delta_k(\Lambda(C))$. Unfortunately, it is hard to obtain the optimal polynomial to bound the above estimates [RW92], unless the eigenvalues of C are positive. Which in this case the bounds is similar to the bound of Conjugate Gradient Method. However, some properties can be derived. Observe that if the polynomials in Π_k^1 is scaled properly we obtain that $\delta_k(\alpha I) \leq \delta_k(I)$ for any $\alpha \neq 0$. also if $\tilde{I} \subset I$, then obviously $\delta_k(\tilde{I}) \leq \delta_k(I)$. So The BiCGSTAB convergence rates can be reduced if we can shrink the intervals I .

By scaling the interval I as indicated previously such that the lower bound of I^+ is unity, we can assume that I^+ and I^- are given by:

$$I^- = [\frac{1}{2}(1 - \sqrt{(1 + 4\rho^2\kappa(A)^2)}), \frac{1}{2}(\kappa(M)^2 - \sqrt{(\kappa(M)^2 + 4\rho^2)})),$$

and

$$I^+ = [1, \frac{1}{2}(\kappa(M)^2 + \sqrt{(\kappa(M)^2 + 4\rho^2\kappa(A)^2)})],$$

where $\rho = \rho(A, M)$ denotes relative scaling between the smallest singular values of A and singular values of M :

$$\rho(A, M) = \frac{\sigma_m}{\mu_n}$$

The expression I^- and I^+ as given above indicates that convergence rates of BiCGSTAB can be estimated by three parameters $\kappa(A)$, $\kappa(M)$, and $\rho(A, M)$. Note that the upper limit of I^+ will increase as function of ρ when $\kappa(A)$ and $\kappa(M)$ are fixed. On the other hand, the upper limit of I^- will approach zero as $\rho \rightarrow 0$. Since the polynomials in Π_k^1 must satisfy $p(0) = 1$, obviously this is not desirable in order to make $\delta_k(I)$ small. Therefore the scaling factor $\rho(A, M)$ should be neither too small nor too large, i.e. the matrices A and M should be properly balanced.

However, if the scaling factor $\rho(A, M)$ is fixed, it is clear that a reduction one of $\kappa(A)$ or $\kappa(M)$ will lead to shrinking of the interval I^- and I^+ given above. Therefore, if the matrices A and M are well conditioned and properly balanced, the BiCGSTAB applied to saddle point problems will converge fast.

4. Preconditioning of Saddle-Point Problem. From the previous section we know that the rate of convergence of BiCGSTAB for saddle point problems in some sense is governed by three parameters: $\kappa(A)$, $\kappa(M)$, and $\rho(A, M)$. There is no clear connection between these parameters with $\kappa(C)$ the condition number of C .

As previously discussed, the rate of convergence of BiCGSTAB for saddle point problem can be accelerated if we can shrink the interval I^+ and I^- . This shrinking of intervals can be achieved by reducing the condition number of A or M while kept the relative scaling $\rho(A, M)$ fixed. Therefore the preconditioned systems will be designed such that the conditioning of at least one of the matrices is reduced, while the relative scaling between them is kept approximately fixed.

We will use a block diagonal preconditioner for saddle point problems. This type of preconditioner has been proposed by Fortin [Fo89] and explored further by Rusten

& Winther [RW92]. The advantages of block diagonal preconditioner is that the preconditioned systems have the same structures with the original linear systems with the property that convergence rate of the iterative method is significantly improved.

Let $L \in R^{n \times n}$ and $R \in R^{m \times m}$ be non-singular matrices, and let $K_1 = \text{diag}(R^T, L)$ and $K_2 = \text{diag}(L^T, R)$. Then the matrix $K_1^{-1}CK_2^{-1}$ is given by:

$$K_1^{-1}CK_2^{-1} = \begin{pmatrix} (L^{-1}AR^{-1})^T & 0 \\ (L^{-1}ML^{-1}) & L^{-1}AR^{-1} \end{pmatrix}$$

Hence the block structure of original linear systems still preserved after preconditioning.

If L and R are such that :

$$\kappa(L^{-1}ML^{-1}) \ll \kappa(M)$$

and/or

$$\kappa(L^{-1}AR^{-1}) \ll \kappa(A)$$

while

$$\rho(L^{-1}AR^{-1}, L^{-1}ML^{-1}) \approx \rho(A, M)$$

then according to discussion in previous section, the preconditioned BiCGSTAB will usually converge faster.

The choice of preconditioning matrices L and R is obviously problem dependent. Since M is symmetric positive definite and the preconditioning of such systems is extensively studied [Gu84], we often know how to find a matrix L with required property that $\kappa(L^{-1}ML^{-1})$ is sufficiently small. In such cases the remaining problem is therefore partly to find a matrix R such that $(L^{-1}AR^{-1})$ is well conditioned, i.e R should be chosen such that $(L^{-1}AR^{-1})$ is approximately a unitary matrix. Observe that this is equivalent to the requirement:

$$A^T(LL^T)A \approx R^TR$$

Hence R should be chosen such that R^TR is a good preconditioner for the positive definite matrix $A^T(LL^T)A$.

5. Choice of Diagonal Preconditioner. In this section we proposed three block preconditioner for saddle-point problem, in a case M a diagonal matrix, as a result of discussion of previous sections.

The first preconditioner M_1 is constructed by taking $L = \alpha \text{diag}(m_{11}^{-\frac{1}{2}}, \dots, m_{nn}^{-\frac{1}{2}})$, where m_{ii} is diagonal entry of mass matrix and α is scaling parameter to conditioned C . We choose $R = I$. Since the mass matrix is generated using lumped mass integration, we obtain M is diagonal, so L is the Cholesky factor of M for $\alpha = 0$. The best choice of α have to be approximately close to the smallest singular values of $(L^{-1}A)$. Using heuristic arguments, Arioli et. al. [ADR89] gave estimates $\alpha = \frac{\max |l_{ik}^{-1}a_{kj}|}{1000}$.

The second preconditioner M_2 is constructed by taking $L = I_n$ and $R = \text{diag}((a_{1j}a_{j1})^{\frac{1}{2}}, \dots, (a_{nj}a_{jn})^{\frac{1}{2}})$. This preconditioner is to approximate A^TA by its diagonal.

The third preconditioner M_3 is constructed by taking $L = I_n$ and choose R a diagonal matrix from A^TA that is given by Schwartz[Sc85] in the authors thesis [Ga93].

Acknowledgement

I would like to thank AIDAB, the Ministry of Foreign Affairs and International Trade, that made me possible to study at ANU. I'd like to thank Stephen G. Roberts, for his patience, time and suggestion to improve this article.

Reference

- [Bj67] A. Bjork, 1967, Iterative Refinement of Linear Least-Squares Solutions, BIT, **7**, pp 257-278
- [Ga93] A.D. Garnadi, 1993, Iterative Methods for Mixed Finite Element Formulation of Biharmonic Equation, Graduate Diploma in Science, The Australian National University, Canberra
- [Gu84] I. Gustafsson, 1984, A Preconditioned Iterative Methods for The Solution of Biharmonic Equation., IMA J. Numer. An., **4**, pp 55-67
- [Gu89] I. Gustafsson, 1989, A Class of Preconditioned Conjugate Gradient Methods Applied to Finite Element Equation - A Survey on MIC methods, in O. Axelsson & Y. I. Kotilina, eds., Lect. Notes Math, **1457**, 55-67
- [Jo92] W. Joubert., 1992, Lanczos Methods for the NonSymmetric Systems of Linear Equations, SIAM J. Mat. Anal. Applic., **13**, pp 926-943
- [RW92] T. Rusten & R. Winther., 1992, The Preconditioned Iterative Methods for Saddle-Point Problems, SIAM J. Matrix Anal. Applic., **13**, pp 887-905
- [Sc88] H. R. Schwarz, 1988, Finite Element Methods, Academic Press.
- [Sl69] A. van der Sluis., 1969, Condition Numbers and Equilibration of Matrices, Num. Math., **14**, pp 14-23
- [So89] P. Sonneveld, 1989, CGS: A Fast Lanczos-Type Solver for Nonsymmetric Linear Systems, SIAM J. Sci. Stat. Comput., **10**, pp 36-52
- [Vo92] H. van der Vorst., 1992, BiCGSTAB: a Fast and Smoothly Converging Variant of BiCG for the Solution of NonSymmetric Linear Equations, SIAM J. Sci. Stat. Comput., **13**, pp 631-644