Preconditioning for Normal Equations and Least Squares

Andy Wathen
Oxford University, UK

Solution of

$$\boldsymbol{Bx} = \boldsymbol{b}$$

where $\boldsymbol{B}$ is not symmetric (non-self-adjoint), and elimination is infeasible $\Rightarrow$ iterative methods.

Two main possibilities:

- solve $\boldsymbol{Bx} = \boldsymbol{b}$ with a non-symmetric iterative method such as GMRES
- solve the Normal Equations $\boldsymbol{B^T Bx} = \boldsymbol{B^T b}$ with a symmetric iterative method such as Conjugate Gradients (LSQR)

In either case generally need *preconditioning* to get acceptably fast convergence

Preconditioning:

- GMRES : a matrix(operator) approximation P such that

$$P^{-1}Bx = P^{-1}b \quad \text{or} \quad BP^{-1}y = b, x = P^{-1}y$$

  much easier to solve
- LSQR : a matrix approximation for B^TB

Practical considerations:

- GMRES : no descriptive convergence theory $\Rightarrow$ heuristics for preconditioning
- LSQR : convergence bounded by eigenvalues of the (preconditioned) normal equations, but how to approximate B^TB (since have B and generally do not want to compute B^TB)?

One obvious idea: find a good (spectral) approximation P to B, then use $P^T P$ as preconditioner for $B^T B$

Unfortunately—as deserves to be much more widely known—P can be an excellent preconditioner for B even in the symmetric case, whereas $P^T P$ can be arbitrarily poor as a preconditioner for $B^T B$. (*Braess & Peisker (1986)*)

One obvious idea: find a good (spectral) approximation P to B, then use $P^T P$ as preconditioner for $B^T B$

Unfortunately—as deserves to be much more widely known—P can be an excellent preconditioner for B even in the symmetric case, whereas $P^T P$ can be arbitrarily poor as a preconditioner for $B^T B$. (*Braess & Peisker (1986)*)

The matrix squaring problem

Example 1

$$B = \begin{bmatrix} b_0 & & & \\ & b_1 & & \\ & & \ddots & \\ & & & b_n \end{bmatrix}, \quad P = \begin{bmatrix} & & & b_0 \\ & & b_1 & \\ & \iddots & & \\ b_n & & & \end{bmatrix},$$

so that

$$P^{-1} = \begin{bmatrix} & & & b_n^{-1} \\ & & b_{n-1}^{-1} & \\ & \iddots & & \\ b_0^{-1} & & & \end{bmatrix}, P^{-1}B = \begin{bmatrix} & & & 1 \\ & & 1 & \\ & \iddots & & \\ 1 & & & \end{bmatrix} := Y$$

where clearly $Y^2 = I$ so $\sigma(P^{-1}B) = \{\pm 1\}$. ($P^{-1}B$ is real symmetric, hence diagonalisable). Thus GMRES with preconditioner P will converge (terminate) in ≤ 2 iterations.

P is thus an excellent preconditioner for B.

However:

$$(P^T P)^{-1} B^T B = \begin{bmatrix} (b_0/b_n)^2 & & & \\ & (b_1/b_{n-1})^2 & & \\ & & \ddots & \\ & & & (b_n/b_0)^2 \end{bmatrix}$$

so that the eigenvalues of the preconditioned normal equations can be arbitrarily badly distributed. For example, taking $b_k = 10^k$ gives

$$\sigma((P^T P)^{-1} B^T B) = \{10^{2n-4k}, k = 0, 1, \ldots n\},$$

thus widely spread eigenvalues and condition number 10^{4n}.

$P^T P$ is a very bad preconditioner for $B^T B$.

Example 2: both B & P SPD

If $A \in \mathbb{R}^{m \times n}$ is any matrix of full column rank $n \leq m$, $C = QA$ for any orthogonal matrix $Q \in \mathbb{R}^{m \times m}$ and $\ell = m + n$, then

$$B = \begin{bmatrix} 2I & C \\ C^T & 2A^TA \end{bmatrix} \in \mathbb{R}^{\ell \times \ell}, P = \begin{bmatrix} 2I & 0 \\ 0 & 2A^TA \end{bmatrix} \in \mathbb{R}^{\ell \times \ell}$$

and $$\frac{1}{2} z^T P z \leq z^T B z \leq \frac{3}{2} z^T P z \quad \text{for all } z,$$

so that $\sigma(P^{-1}B) \subset [\frac{1}{2}, \frac{3}{2}]$, $\kappa \leq 3$

so P is an excellent preconditioner for B

CG convergence:

$$\|z - z_k\|_B \leq 2 \underbrace{\left(\frac{\sqrt{3}-1}{\sqrt{3}+1} \right)^k}_{\approx (0.27)^k} \|z - z_0\|_B$$

However:

$$z=\begin{bmatrix} x \\ 0 \end{bmatrix} \Rightarrow \frac{z^T B^2 z}{z^T P^2 z} = \frac{4x^T x + x^T Q A A^T Q^T x}{9x^T x} \leq \lambda_{\max}(P^{-2}B^2)$$

$$z=\begin{bmatrix} 0 \\ y \end{bmatrix} \Rightarrow \frac{z^T B^2 z}{z^T P^2 z} = \frac{y^T A^T A y + 4y^T (A^T A)^2 y}{9y^T (A^T A)^2 y} \geq \lambda_{\min}(P^{-2}B^2)$$

$$\Rightarrow \kappa = \frac{\lambda_{\max}(P^{-2}B^2)}{\lambda_{\min}(P^{-2}B^2)} \geq \frac{4 + \sigma_{\max}(A)^2}{4 + 1/\sigma_{\max}(A)^2}$$

where $\sigma_{\max}(A)$ is the largest singular value of A, arbitrarily large!

$P^T P = P^2$ arbitrarily bad preconditioner for $B^T B = B^2$.

Many non-symmetric examples, but GMRES convergence not clear.

Nachtigal, Reddy & Trefethen (1992) in comparison of different non-symmetric iterative methods already showed examples where LSQR is much worse than GMRES (and vice-versa)

When matrix squaring can not arise

Theorem (essentially due to *Gratton et al (2018)*)

If $\|I - BP^{-1}\| < \sqrt{2} - 1$, then
$\sigma((P^T P)^{-1} B^T B) \subset (0, 2)$.

Moreover if $\|I - BP^{-1}\| = \sqrt{2} - 1 - \epsilon$ then

$$\lambda_{\min}((P^T P)^{-1} B^T B) \geq \sqrt{2}\epsilon + \epsilon^2,$$

$$\lambda_{\max}((P^T P)^{-1} B^T B) \leq 2 - \sqrt{2}\epsilon - \epsilon^2 \quad \square$$

$\Rightarrow$ e.g. if $\epsilon = 0.1$ then CG for the preconditioned normal equations (and LSQR) iteration error must contract by 0.56 or better at every iteration in the natural norm.

Linear Least Squares

$$x = \mathbf{argmin} \|b - Ax\|_2, \quad A \in \mathbb{R}^{m \times n}, m > n$$

also usually use LSQR when factorisation infeasible: with right preconditioner, $P \in \mathbb{R}^{n \times n}$:

$$y = \mathbf{argmin} \|b - AP^{-1}y\|, \quad x = P^{-1}y,$$

LSQR mathematically equivalent to CG for the preconditioned normal equations

$$P^{-T}A^TAP^{-1}y = P^{-T}A^Tb$$

$$y = \mathbf{argmin} \|b - AP^{-1}y\|, \quad x = P^{-1}y,$$

$$\Leftrightarrow$$

$$P^{-T}A^{T}AP^{-1}y = P^{-T}A^{T}b$$

If can choose P so that

$$\kappa(AP^{-1}) = \sigma_{\mathbf{max}}(AP^{-1})/\sigma_{\mathbf{min}}(AP^{-1})$$

is small, then rapid LSQR convergence because squares of the singular values are the eigenvalues of the preconditioned normal matrix

$$(AP^{-1})^{T}AP^{-1} = P^{-T}A^{T}AP^{-1}$$

E.g. *Rokhlin & Tygert (2008)* randomized generation of P when $m \gg n$ ($\rightarrow$ blendenpik,...)

Least Squares Matrix Squaring problem: Example 3

$$A = \begin{bmatrix} E \\ B \end{bmatrix} \in \mathbb{R}^{2n \times n}, B, E = QP \in \mathbb{R}^{n \times n}, Q \text{ orthogonal}$$

$$AP^{-1} = \begin{bmatrix} Q \\ BP^{-1} \end{bmatrix} \text{ and } P^{-T}A^{T}AP^{-1} = I + P^{-T}B^{T}BP^{-1}$$

Now, if B, P are as in Examples 1,2, then eigenvalues of BP^{-1} can be nicely distributed (clustered) whereas eigenvalues of $P^{-T}B^{T}BP^{-1}$ can be badly distributed (widely spread).

Thus, though smallest eigenvalue ≥ 1, largest eigenvalue of preconditioned normal equations can be large $\Rightarrow$ CG for normal equations—and LSQR for the least squares problem—could converge slowly

Conclusion

- preconditioning for normal equations not generally easy
- maybe why GMRES is more widely used for linear equations
- preconditioning for least squares more difficult in general

References and Acknowledgement

Braess, D. & Peisker, P., 1986,
'On the numerical solution of the biharmonic equation and the role of squaring matrices for preconditioning',
IMA J. Numer. Anal. **6**, pp. 393–404.

Gratton, S., Gurol, S., Simon, E. & Toint, P. 2018,
'A note on preconditioning weighted linear least-squares, with consequences for weakly constrained variational data assimilation',
Quarterly Journal of the Royal Meteorological Society **144**, pp. 934–940.

Nachtigal, N., Reddy, S. & Trefethen L., 1992,
'How fast are nonsymmetric matrix iterations?',
SIAM J. Matrix Anal. Appl. **13**, pp. 778–795.

Rokhlin, V. & Tygert, M., 2008,
'A fast randomized algorithm for overdetermined linear least-squares regression',
Proc. Nat. Acad. Sci. **105**, pp. 13212–13217.

I thank Michael Saunders for getting me to think about preconditioning for LSQR