Orthogonality and stability in large matrix iterative algorithms

Chris Paige* & Wolfgang Wülling†

*Computer Science, McGill University,
with financial support from NSERC (Canada).

†W2 Actuarial & Math Services Ltd., Blackpool, Lancashire, England, FY4 5PN

Sparse Days at CERFACS, Toulouse, France, June 25–26, 2012

Some Notation (reals only).

Singular Values: $\sigma_i(A)$. Norms: $\|v\| \equiv \|v\|_2 \triangleq \sqrt{v^T v}$,

$$\|A\|_2 \triangleq \sigma_{\max}(A), \qquad \|A\|_F^2 \triangleq \mathrm{trace}(A^T A).$$

$\mathrm{sut}(A) \equiv$ the STRICTLY UPPER TRIANGULAR part of A.

$\mathrm{slt}(A) \equiv$ the STRICTLY LOWER TRIANGULAR part of A.

ϵ the computer floating point precision.

$O(\epsilon)\|A\| \approx 0$, $O(\epsilon) \simeq 0$, SUT≡Strictly Upper Triangular

Backward Rounding Error Analysis (BREA)

Reliable Numerical Algorithms

E.g. Householder QR is a Backward Stable Algorithm (BSA),

so there exists an orthogonal $\hat{Q}$ such that for $B = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$:

$$B + E = \hat{Q} \begin{bmatrix} R^c \\ 0 \end{bmatrix}, \quad \|E\| \le O(\epsilon)\|B\|, \qquad \|Q^c - \hat{Q}\| \le O(\epsilon),$$

Q^c & R^c the computed Q & R,

Abbreviation:

$$\boxed{B \approx \hat{Q} \begin{bmatrix} R^c \\ 0 \end{bmatrix}, \quad Q^c \simeq \hat{Q}, \quad \hat{Q}^T \hat{Q} = I.}$$

$O(\epsilon)\|A\| \approx 0$, $O(\epsilon) \simeq 0$, *SUT=Strictly Upper Triangular*

Backward Rounding Error Analysis (BREA).

Because of Wilkinson's BREA theory, we KNOW we have Backward Stable **orthogonal transformation algorithms**,

e.g. the QR algorithm—while the matrix is not too big.

For many large sparse matrix problems we turn to

Vector orthogonalization algorithms, e.g. CG.

These are not BSAs, Wilkinson's BREA theory doesn't apply.

But many of these algorithms still 'work'.

We seek a Matrix based Rounding Error Theory

to describe their subtle numerical behaviours*.

Vector Orthogonalization Algorithms

'Vector Orthogonalization' Algorithms

orthogonalize each **new** vector
against previous **supposedly** orthogonal vectors.

For example:

Modified Gram–Schmidt (MGS),
Arnoldi's method for the unsymmetric eigenproblem,
Saad & Schultz: MGS-GMRES for unsymmetric $Ax = b$,

Lanczos: Tridiagonalization of square A,
Lanczos, & Hestenes & Stiefel: Conjugate gradients (CG) $Ax = b$,
Golub & Kahan 'Vector Orthogn.' Bidiagonalization of general B.

ALL can lose orthogonality using **finite precision computations**!

Assume each "orthogonal" vector has unit length: $\|v_j\|_2 = 1$.

An ideal measure of loss of orthogonality

Given $V \in \mathbb{R}^{n\times k}$ with $V^TV = U^T + I + U, \quad U \triangleq \mathrm{sut}(V^TV)$.

Instead of using $\|U\|_2$,

if $\boxed{S \triangleq (I+U)^{-1}U}$, $k \times k$, strictly upper triangular, then

$\|S\|_2$ is a **great** measure of loss of orthogonality in V:

$$0 \le \|S\|_2 \le 1.$$

$$S = 0 \quad \Leftrightarrow \quad V^TV = I,$$

$$\|S\|_2 = 1 \quad \Leftrightarrow \quad V \text{ rank deficient.}$$

In fact

# of usvs of S = column rank deficiency of V.

Gram-Schmidt Orthogonalization & Augmented Backward Stability

MGS

In theory MGS produces V and upper triangular R so that:

$$B = VR, \qquad V^T V = I.$$

For many **large sparse matrix problems**:

MGS + Krylov sequences $\{b, Ab, A^2 b, \ldots\}$ leads to, e.g.:

Arnoldi's method for unsymmetric $Ax = x\lambda$, which leads to:

Saad & Schultz's MGS-GMRES for unsymmetric $Ax = b$.

Important to understand MGS.

Measuring loss of orthogonality in MGS

Remember:

k steps of MGS for $B_k = V_k R_k$ gave computed V_k

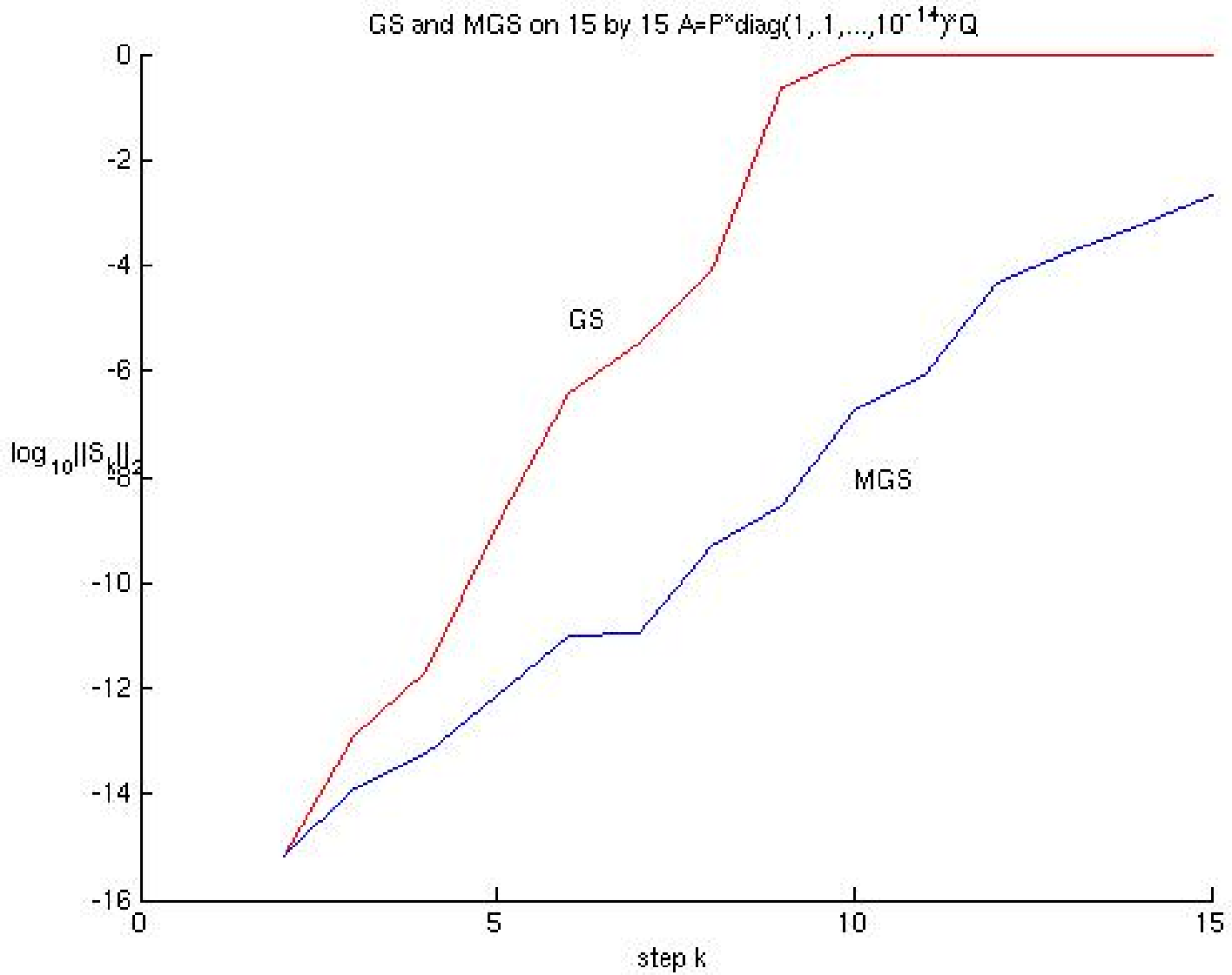
where $V_k^T V_k = U_k^T + I + U_k, \qquad U_k \triangleq \mathrm{sut}(V_k^T V_k).$

If $\boxed{S_k \triangleq (I + U_k)^{-1} U_k}$ then

$\|S_k\|_2$ is our measure of loss of orthogonality in V_k.

$\|S_k\|_2$: Orthogy. Loss in $B_k = V_k R_k$ in GS & MGS

GS and MGS on $B_m \in \mathbb{R}^{15\times 15}$, singular values $1, 10^{-1}, \ldots, 10^{-14}$



$O(\epsilon)\|A\| \approx 0$, $O(\epsilon) \simeq 0$, *SUT=Strictly Upper Triangular*

Obtaining orthonormal Q_k from $n \times k$ $V \equiv V_k$

Given $V \in \mathbb{R}^{n\times k}$ with $V^T V = U^T + I + U$, $\quad U \triangleq \mathrm{sut}(V^T V)$,

if $S \triangleq (I+U)^{-1}U$ & $Q_k \triangleq \begin{bmatrix} S \\ V(I-S) \end{bmatrix}$,

then $\boxed{Q_k^T Q_k = I}$.

$(n+k)\times k$ Q_k is an "orthonormal augmentation" of $n\times k$ $V \equiv V_k$.

Aside: Proof of Orthogonality.

Given $V \in \mathbb{R}^{n\times k}$ with $V^TV = U^T + I + U$, $\quad U \equiv \mathrm{sut}(V^TV)$,

if $S \equiv (I+U)^{-1}U$ & $Q_k \equiv \begin{bmatrix} S \\ V(I-S) \end{bmatrix}$, then $\boxed{Q_k^TQ_k = I}$.

Proof:

$-(I+U)S = -U = I - (I+U)$, so $\boxed{(I+U)(I-S) = I}$.

$$\begin{aligned} Q_k^TQ_k &= S^TS + (I-S)^TV^TV(I-S) \\ &= S^TS + (I-S)^T(-I+I+U^T+I+U)(I-S) \\ &= S^TS - (I-S)^T(I-S) + I - S + I - S^T \\ &= S^TS - I + S + S^T - S^TS + I - S + I - S^T = I. \end{aligned}$$

Clearly $\|S\|_2 \le 1$. $\quad \|S\|_2 = 1 \Leftrightarrow \mathrm{rank}(V) < k$, $\quad$ (via the CSD).

Augmented BS of MGS for the QR of $B_k \in \mathbb{R}^{n\times k}$.

For $B_k = VR$, MGS gives computed V & R where from a REA:

$$\begin{bmatrix} 0 \\ B_k \end{bmatrix} + \begin{bmatrix} E \\ F \end{bmatrix} = Q_k R \equiv \begin{bmatrix} S \\ V(I-S) \end{bmatrix} R, \quad \left\| \begin{matrix} E \\ F \end{matrix} \right\|_2 \leq O(\epsilon)\|B\|_2, \quad Q_k^T Q_k = I,$$

so that R is Backward Stable for the QR factorization of $\begin{bmatrix} 0 \\ B_k \end{bmatrix}$, an Augmented problem! Thus "ABS" of MGS.

How loss of orthogonality occurs:

$$\sigma_i(R) \approx \sigma_i(B_k), \qquad E = SR, \qquad S = ER^{-1},$$
$$\|S\|_2 \leq \|E\|_2 \|R^{-1}\|_2 \leq O(\epsilon)\|B_k\|_2 \|R^{-1}\|_2 \approx \kappa_2(B_k) O(\epsilon).$$

Björck & Paige (1992); Paige, Rozložník & Strakoš (2006); following Charles Sheffield's ∼1967 augmented matrix suggestion.

$O(\epsilon)\|A\| \approx 0$, $O(\epsilon) \simeq 0$, *SUT=Strictly Upper Triangular*

$\|S_k\|_2 \leq \kappa_2(B_k)\epsilon$ for MGS example

On the graph in the middle of the next slide,

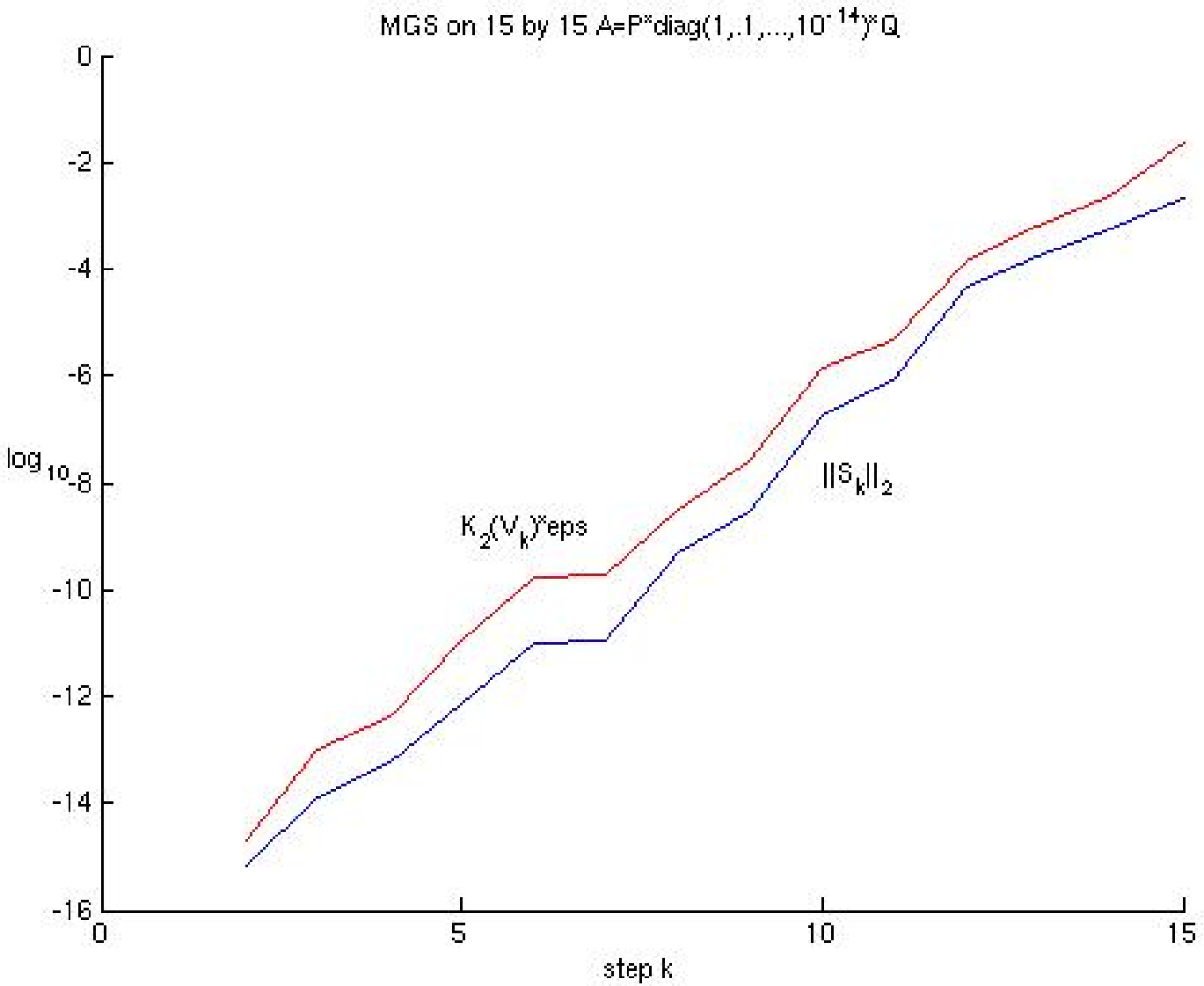
$$K_2(V_k)*\text{eps}$$

should be

$$\kappa_2(B_k)\epsilon$$

$\|S_k\|_2 \leq \kappa_2(B_k)\epsilon$ for MGS example

on $B_n \in \mathbb{R}^{15\times 15}$, singular values $1, 10^{-1}, \ldots, 10^{-14}$



$O(\epsilon)\|A\| \approx 0$, $O(\epsilon) \simeq 0$, *SUT=Strictly Upper Triangular*

Un-augment the result:

$$\begin{bmatrix} 0 \\ B_k \end{bmatrix} \approx Q_k R \equiv \begin{bmatrix} S \\ V(I-S) \end{bmatrix} R, \qquad Q_k^T Q_k = I.$$

By using the SVD of S, if $\text{rank}(B_k) = k$:

can show there exists $n \times k$ $\widehat{Q}_k$ such that:

$$\boxed{B_k \approx \widehat{Q}_k R, \qquad \widehat{Q}_k^T \widehat{Q}_k = I,}$$

so MGS is BS for computing R from B_k.

Björck & Paige (1992).

$\widehat{Q}_k$ is the closest orthogonal matrix to $V(I-S)$.

Nick Higham (2002).

MGS-GMRES for large sparse $Ax = b$

Convergence of Saad & Schultz MGS-GMRES

MGS-GMRES gives a BSS x_k to $Ax = b$ in $k \leq n$ steps if

$\sigma_{\min}(A) \gg n^2 \epsilon \|A\|_F$. Paige, Rozložník & Strakoš (2006).

Thoughts

- We now know how MGS works numerically:
  - It provides a BSS R,
  - It leads to BSS for LLS problems,
  - It can be used as the basis for other algorithms.
- We now know how full MGS-GMRES works numerically:
  - Full MGS-GMRES gives a BSS in $\leq n$ steps,
  - Full MGS-GMRES does not need re-orthogonalization,
  - Storage and operations per step increase with k.
    Use Truncated MGS-GMRES? Restarts? . . .
- We have tools for examining Arnoldi's algorithm.
- This was the easy part—implicit orthogonalization algorithms are much more difficult to analyze.
- $\begin{bmatrix} S \\ V(I-S) \end{bmatrix}$ is very useful, S is amazing.
- We can apply these ideas to many algorithms.

The Major Steps in the Analysis

Step 1: $V_k \triangleq [v_1, \ldots, v_k]$, the normalized, computed vectors.

Step 2: $U_k \triangleq \mathrm{sut}(V_k^T V_k)$, $\quad S_k \triangleq (I + U_k)^{-1} U_k$,

$S_k \;\rightarrow$ measures of loss of orthogonality & independence.

Step 3: Exists orthonormal $Q_k \triangleq \begin{bmatrix} S_k \\ V_k(I - S_k) \end{bmatrix}$, $\quad Q_k^T Q_k = I_k$.

Step 4: Find the augmented system. E.g. MGS of $n \times k$ B_k:

$$\begin{bmatrix} 0 \\ B_k \end{bmatrix} \approx Q_k R_c \equiv \begin{bmatrix} S_k \\ V_k(I - S_k) \end{bmatrix} R_c, \quad Q_k^T Q_k = I_k, \qquad \text{REA}.$$

Step 5: Un-augment this augmented system.

E.g. MGS: $\exists$ $n \times k$ $\widehat{Q}_k$ such that $B_k \approx \widehat{Q}_k R_c$, $\quad \widehat{Q}_k^T \widehat{Q}_k = I_k$.*

Orthogonality and the Lanczos process

Implicit orthogn.: the Lanczos process (1950)

Given $A = A^T \in \mathbb{R}^{n\times n}$ and $v_1 \in \mathbb{R}^n$, $\quad v_1^T v_1 = 1$,
compute an orthonormal sequence $v_1, v_2, \ldots, v_{k+1}$

via a 3 term recurrence:

$$v_{j+1}\beta_{j+1} := Av_j - v_j\alpha_j - v_{j-1}\beta_j,$$

implicit orthogonalization, makes life MUCH tougher!

Then with

$$V_k \equiv [v_1, v_2, \ldots, v_k], \qquad T_k \equiv \begin{bmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \cdot & \\ & \cdot & \cdot & \beta_k \\ & & \beta_k & \alpha_k \end{bmatrix},$$

$$\boxed{AV_k = V_k T_k + v_{k+1}\beta_{k+1}e_k^T = V_{k+1}T_{k+1,k},} \qquad \boxed{V_{k+1}^T V_{k+1} = I.}$$

Useful for large sparse $Ax = x\lambda$, $\quad Ax = b$.

‘Recursive augmented stability’ of the Lanczos process

RAS of the computational Lanczos process

Ideally, given $A = A^T \in \mathbb{R}^{n\times n}$ and $v_1 \in \mathbb{R}^n$, $v_1^T v_1 = 1$,

$$V_k \triangleq [v_1, v_2, \ldots, v_k], \qquad T_k{}^T = T_k \triangleq \text{tridiag}(\beta_i, \alpha_i, \beta_{i+1}),$$

$$(1)\text{: } AV_k = V_k T_k + v_{k+1}\beta_{k+1} e_k^T, \qquad (2)\text{: } V_{k+1}{}^T V_{k+1} = I.$$

Computed T_k & V_k from the Lanczos algorithm satisfy:

$$(1)\text{: } \boxed{\left(\begin{bmatrix} T_k & 0 \\ 0 & A \end{bmatrix} + H_k\right) Q_k = Q_k T_k + q_{k+1}\beta_{k+1} e_k^T}$$

$$H_k = H_k^T, \quad \|H_k\|_2 \approx 0,$$

$$(2)\text{: } [Q_k \mid q_{k+1}]^T [Q_k \mid q_{k+1}] = I_{k+1}. \qquad \textbf{RAS:}$$

k steps of an exact Lanczos process for an AUGMENTED matrix.

The same old Q_k.

Normalized **computed** V_k : $V_k^T V_k = U_k^T + I + U_k$,

$$U_k \triangleq \mathrm{sut}(V_k^T V_k), \qquad S_k \triangleq (I+U_k)^{-1} U_k,$$

$$Q_k \triangleq \begin{bmatrix} S_k \\ V_k(I-S_k) \end{bmatrix}, \qquad Q_k^T Q_k = I_k.$$

Development of T_k and Q_k

The augmented problem seems weird:

$$\left(\begin{bmatrix} T_k & 0 \\ 0 & A \end{bmatrix} + H_k\right) Q_k = Q_k T_k + q_{k+1}\beta_{k+1} e_k^T, \quad Q_k \triangleq \begin{bmatrix} S_k \\ V_k(I - S_k) \end{bmatrix}.$$

Let $\bar{Q}_k$ be Q_k less its zero kth row, then $\bar{Q}_{k+1} = [Q_k, q_{k+1}]$, and the augmented problem becomes

$$\begin{bmatrix} T_{k,k-1} & 0 \\ 0 & A \end{bmatrix} \bar{Q}_k \approx \bar{Q}_{k+1} T_{k+1,k}, \quad \bar{Q}_k^T \bar{Q}_k = I_k, \quad \bar{Q}_{k+1}^T \bar{Q}_{k+1} = I_{k+1}.$$

Showing how

$$T_{k,k-1} \quad \rightarrow \quad T_{k+1,k} \qquad \& \qquad \bar{Q}_k \quad \rightarrow \quad \bar{Q}_{k+1}.$$

"Recursive", from implicit nature. Strange, but not quite so weird.

Augmented Biorthonormality

Suppose we want biorthonormal V & W, i.e. $W^T V = I$.

With vector orthogonalization this might be far from true.

But for computed V & W with $w_i^T v_i = 1, \quad i = 1, 2, \ldots$

$$U \triangleq \operatorname{sut}(W^T V), \qquad S \triangleq (I + U)^{-1} U,$$
$$L \triangleq \operatorname{slt}(W^T V), \qquad R \triangleq (I + L^T)^{-1} L^T,$$
$$Q \triangleq \begin{bmatrix} S \\ V(I - S) \end{bmatrix}, \qquad P \triangleq \begin{bmatrix} R \\ W(I - R) \end{bmatrix}.$$

Then $P^T Q = I$, biorthonormal augmented matrices!

Paige (2009), from a question by Ron Morgan, Zeuthen 2008.

The Unsymmetric Lanczos Process

Using this concept of **augmented biorthonormality**
shows that running k steps of the Lanczos process
on **unsymmetric** $A \in \mathbb{R}^{n \times n}$ in the presence of rounding errors
is equivalent to running k steps of an **exact** Lanczos process
on a perturbation E of the **augmented** matrix $\mathrm{diag}(T_k, A)$.

(Paige, Panayotov, & Zemke, in preparation, extending the analysis by Zhaojun Bai, 1994).

The flaw in the unsymmetric Lanczos process:

$\|E\|_2$ can be **large** if the process approaches breakdown.

Use "look ahead"? (Bill Gragg, Beresford Parlett, et al.).

Some possible Applications of the Analysis

Lanczos Process: Symmetric A.

- The eigenproblem: via the Lanczos process.
- Solution of equations $Ax = b$, $A = A^T$:
  - Positive definite A: CG, Lanczos process, etc.
  - Indefinite A: e.g. MINRES, SYMMLQ, (Mike Saunders & me), etc.
  - All above & even singular A: MINRES-QLP, (Sou-Cheng Choi, Mike Saunders & me).

Lanczos Process on Unsymmetric $n \times n$ A.

Use this to analyze:

- The eigenproblem: unsym. Lanczos process & variants.
- Solution of unsym. equations: Roland Freund's QMR, etc.

General $n \times m$ B.

- Golub & Kahan's **"vector" bidiagonalization** for the SVD.

  (Theory: Lanczos process on **symmetric** $\begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}$, $\begin{bmatrix} u_1 \\ 0 \end{bmatrix}$.)

  And with $u_1 = b/\|b\|_2$ for $Bx \approx b$ can solve:

  - Solution of equations & Least squares
    CGLS (Hestenes & Stiefel), LSQR (Mike Saunders & me),
    LSMR (David Fong & Mike Saunders),
  - Total Least Squares (Åke Björck),
  - Scaled TLS (Zdeněk Strakoš & me),
  - Core problems (Zdeněk Strakoš & me).

Background to this "Augmented" approach:

C. Sheffield, comment to Gene Golub, circa 1967.

C. C. Paige, Ph.D. thesis, London University 1971.

A. Greenbaum, Ph.D. thesis, UC, Berkeley 1981, LAIA 1989.

Å. Björck and C.C. Paige, SIMAX 1992, BIT 1994.

J.L. Barlow, N. Bosner and Z. Drmač, LAIA 2005.

C. C. Paige, M. Rozložník, and Z. Strakoš, SIMAX 2006.

C. C. Paige, *A useful form of unitary matrix obtained from any sequence of unit 2-norm n-vectors*.
SIMAX, 31 (2009), pp. 565–583.

C. C. Paige, *An Augmented Stability Result for the Lanczos Hermitian Matrix Tridiagonalization Process*.
SIMAX, 31 (2010), pp. 2347–2359.