# Deflations Preserving Relative Accuracy

W. Kahan, Prof. Emeritus
Mathematics Dept., and E.E. & Computer Science Dept. #1776
University of California, Berkeley CA 94720-1776
[ wkahan at eecs d0t berkeley d0t edu ]

**Contents:** **Page**



Posted at <www.eecs.berkeley.edu/~wkahan/ma221/Deflate.pdf>

# Deflations Preserving Relative Accuracy

## §0: Abstract

Deflation turns a matrix eigenproblem into two of smaller dimensions by annihilating a block of off-diagonal elements. When does deflation perturb at worst the last significant digit or two of each of an Hermitian matrix's eigenvalues no matter how widely their magnitudes spread? We seek practicable answers to this question, particularly for tridiagonals, analogous to answers for bidiagonals' singular values found by Ren-Cang Li in 1994. How deflation affects singular vectors and eigenvectors is assessed too, as is the exploitation of spectral gaps when known.

## §1: Introduction

Let Hermitian $H := H' := \begin{bmatrix} M & B \\ B' & W \end{bmatrix}$ and $Y := Y' := \begin{bmatrix} M & O \\ O' & W \end{bmatrix}$ have ordered *Spectra* respectively

$$\mathbb{E}(H) = \{ \theta_1 \ge \theta_2 \ge \ldots \ge \theta_n \} \quad \text{and} \quad \mathbb{E}(Y) = \{ \eta_1 \ge \eta_2 \ge \ldots \ge \eta_n \} = \mathbb{E}(M) \cup \mathbb{E}(W)$$

wherein $\mathbb{E}(M) = \{ \mu_1 \ge \mu_2 \ge \ldots \ge \mu_m \}$ and $\mathbb{E}(W) = \{ \omega_1 \ge \omega_2 \ge \ldots \ge \omega_{n-m} \}$.

Here the union $\cup$ is the union of *Multisets* because some eigenvalues $\eta_j$ may be repeated.

Y comes from H via *Deflation*, which reduces a big n-by-n eigenvalue computation to two smaller ones, m-by-m and (n–m)-by-(n–m) (not both much smaller), computable faster. It is well known (see Li & Mathias [1999]) that every $|\theta_j - \eta_j| \le \|B\|$, the biggest singular value of B. This bounds *Absolute* errors induced by deflation. We seek bounds upon *Relative* errors $\log(\theta_j/\eta_j)$. There are obvious bounds like roughly $\|H^{-1}\|\cdot\|B\|$ and $\|Y^{-1}\|\cdot\|B\|$, but these turn out often unnecessarily both too big and too expensive to compute. In §5 we find smaller bounds like $\|M^{-1}\cdot B\|$ and $\|B\cdot W^{-1}\|$, though they are not always much smaller; and they may be practicable, if practicable at all, only when H and/or M or W are/is nearly diagonal.

Let upper-triangular matrices $S := \begin{bmatrix} D & E \\ O' & F \end{bmatrix}$ and $Z := \begin{bmatrix} D & O \\ O' & F \end{bmatrix}$ have ordered singular value sets respectively $\mathbb{S}(S) = \{ \sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_n \}$ and $\mathbb{S}(Z) = \{ \zeta_1 \ge \zeta_2 \ge \ldots \ge \zeta_n \}$, all nonnegative. Again, Z comes from S via deflation, and every $|\sigma_j - \zeta_j| \le \|E\|$. This bounds absolute errors induced by deflation; we seek bounds upon relative errors $\log(\sigma_j/\zeta_j)$. The obvious bounds like roughly $\|S^{-1}\|\cdot\|E\|$ and $\|Z^{-1}\|\cdot\|E\|$ turn out often unnecessarily both too big and too expensive to compute. Smaller bounds like $\|D^{-1}\cdot E\|$ and $\|E\cdot F^{-1}\|$ were found by Ren-Cang Li [1994] to improve bounds exhibited by Demmel & Kahan [1990] (p. 878) only for bidiagonals S; these may also be the only matrices for which Li's bounds rederived hereunder in §4 are practicable.

Besides perturbing eigenvalues and singular values, deflation rotates eigenvectors and singular vectors through angles assessed roughly in §6 and §7. Overestimates for these angles involve our bounds like $\|M^{-1}\cdot B\|$ and $\|D^{-1}\cdot E\|$ upon the relative perturbations to eigenvalues and

singular values, and involve also *Relative* gaps between eigenvalues $\eta_j$ and between singular values. Since these gaps are usually unknown when deflation occurs, the angles' overestimates serve mainly to allay fears that deflations preserving relative accuracies of `values will damage `vectors much more than must most likely be tolerated no matter how `vectors are computed.

Different gaps figure in §8's *Quadratic* relative error-bounds like $\|D^{-1}\cdot E\|^2/\text{gap}$ for singular values and like $\|M^{-1}\cdot B\|^2/\text{gap}$ for eigenvalues. When estimates available for the relative gaps underestimate them at worst mildly, these quadratic bounds can be so much smaller than bounds derived in §4 and §5 as to allow advantageously deflations otherwise disallowed, though such deflations preserving the relative accuracies of `values may impair `vectors intolerably.

Not every matrix computation always produces results of relative accuracy at least about as high as is deserved taking the data's uncertainty into account. A recent survey of such computations is Z. Drmač's §46 in L. Hogben's [2007] ***Handbook***. Among those computational methods that preserve relative accuracy, only a few are candidates for deflations that do likewise. After such a method has been applied to our data, how can we corroborate its results' claims to high relative accuracy? An answer to this question in §9 is the first application of our error-bounds.

Most deflations occur in certain iterations that act upon condensed matrices like tridiagonals and bidiagonals. The sooner a deflation the better, because it reduces both the cost of each iteration and the number of them, but this entails a conflict between reduced and augmented costs: To decide when deflation will not perturb desired results intolerably, iterations that alter M, B, D, E, *etc*. must be augmented by recomputations of bounds like $\|M^{-1}\cdot B\|$ and $\|D^{-1}\cdot E\|$ and their comparisons with tolerances. Ideally the augmentations should add little to the iterations' cost. To this end, relevant by-products of the iterations should be exploited wherever possible, and their innermost loops should be burdened at worst slightly, since deflations preserving relative accuracy can occur at most very infrequently compared with passes around the inner loop. See Parlett *et al*. [1994, 2000, 2012] for lengthy assessments of typical trade-offs of the likelihood of permissible deflations versus tests for them like some explored in §10 here in the context of his dqds iteration.

## §2: A Tiny Tolerance $\tau << 1$

Suppose a tiny positive *Tolerance* $\tau$ is given, and is so tiny that $\tau^2$ is quite negligible so that different approximations like $\tau \approx 1 - e^{-\tau} \approx -\log(1 - \tau) \approx \tau/(1 \pm \tau)$ need not be distinguished. This will simplify the discussion in so far as inequalities like $\tau > |\log(\theta/\eta)|$, $\tau > |(\theta - \eta)/\theta|$, $\tau > |(\theta - \eta)/\eta|$, … need not be distinguished when the tolerance $\tau$ is an upper bound upon tolerable relative errors with which $\mathbb{E}(Y)$ approximates $\mathbb{E}(H)$. Then we shall find in §5 that those errors are surely tolerable whenever *both* $\|M^{-1}\cdot B\| < \tau$ *and* $\|B\cdot W^{-1}\| < \tau$. These conditions resemble Li's conditions in §4 for $\mathbb{S}(Z)$ to approximate $\mathbb{S}(S)$ tolerably, except his have "*either* $\|D^{-1}\cdot E\| < 2\tau$ *or* $\|E\cdot F^{-1}\| < 2\tau$" in place of "*both … and …*".

## §3: Ostrowski's Inequalities for Congruent Hermitian Matrices

Y and H are *Congruent* if $Y = C^{-1\prime}\cdot H\cdot C^{-1}$. Alexandre Ostrowski's now classical inequalities assert that $1/\|C^{-1}\|^2 \le \theta_j/\eta_j \le \|C\|^2$ for every j (except $0/0 := 1$); see C-K. Li & R. Mathias [1999] for an elegant proof. Also if $Z = S\cdot C^{-1}$ or if $Z = C^{-1}\cdot S$, then $1/\|C^{-1}\| \le \sigma_j/\zeta_j \le \|C\|$ for every j (except $0/0 := 1$) follows. Most matrices C used below will resemble this one:

$C^{\pm1} := \begin{bmatrix} I & \pm U \\ O' & I \end{bmatrix}$ wherein U may be rectangular, in which case zero matrix O' has the same shape as the transpose of U, and the two identity matrices I have different dimensions. Since ||C|| is unchanged by unitary or real orthogonal pre- or post-multiplication, U in C may be replaced by a (rectangular) diagonal matrix of U's singular values, ||U|| among them, to let us deduce easily that $\|C^{\pm1}\| = \left\|\begin{bmatrix} 1 & \|U\| \\ 0 & 1 \end{bmatrix}\right\| = \|U\|/2 + \sqrt{(1 + \|U\|^2/4)} = \exp\big(\operatorname{arcsinh}(\|U\|/2)\big)$.

## §4: Derivation of Ren-Cang Li's Bounds for Singular Values

Obtained first in 1994, their proof was simplified in §3.2 of Parlett & Marques [2000], and will be streamlined a little hereunder. Recall upper-triangles S and Z and their singular values:

$$S := \begin{bmatrix} D & E \\ O' & F \end{bmatrix}, \quad \mathbb{S}(S) = \{\, \sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_n \,\}, \quad Z := \begin{bmatrix} D & O \\ O' & F \end{bmatrix}, \quad \mathbb{S}(Z) = \{\, \zeta_1 \ge \zeta_2 \ge \dots \ge \zeta_n \,\}.$$

Choose $C := \begin{bmatrix} I & D^{-1}\cdot E \\ O' & I \end{bmatrix}$ to get $Z = S\cdot C^{-1}$, $\|C^{\pm1}\| = \exp\big(\operatorname{arcsinh}(\|D^{-1}\cdot E\|/2)\big)$, and via §3 find that every relative error $\left|\log(\sigma_j/\zeta_j)\right| < \|D^{-1}\cdot E\|/2$. Similarly every $\left|\log(\sigma_j/\zeta_j)\right| < \|E\cdot F^{-1}\|/2$.

> **Conclusion**: If ***either*** $\|D^{-1}\cdot E\| < 2\tau$ ***or*** $\|E\cdot F^{-1}\| < 2\tau$ then every $\left|\log(\sigma_j/\zeta_j)\right| < \tau$.

The conclusion holds also if D and F are arbitrary squares instead of upper-triangles; further generalization to rectangles is immediate but immaterial here. More important, if some $\zeta_j = 0$ so one of $\|D^{-1}\cdot E\|$ and $\|E\cdot F^{-1}\|$ exists but not the other, the conclusion persists with $0/0 := 1$.

**Example:**

Let n-by-n $S := \operatorname{bidiag}\begin{bmatrix} s & s & \dots & s & s & e \\ 1 & 1 & \dots & 1 & 1 & f \end{bmatrix} = \begin{bmatrix} D & \mathbf{e} \\ \mathbf{o}' & f \end{bmatrix}$ in which the pair $\begin{bmatrix} s \\ 1 \end{bmatrix}$ is absent from only the first and last columns, and $s > f >> 1 > e > 0$. When is **e** is so small that replacing it by **o** deflates S without incurring relative errors worse than $\tau$ in singular values? The least singular value of S is very close to that of D: $\sigma_n \approx (s^2 - 1)/\sqrt{(s^{2n} - n\cdot s^2 + n - 1)}$. The largest singular values of S are not far from those of D: $\sigma_1 \approx s + 1$. This puts $f$ amidst $\mathbb{S}(D)$, so no *spectral gap* (*cf.* §8) is available compared with which to deem $e^2$ negligible. Yet R-C. Li's criterion implies that **e** is negligible if $e < 2\tau\cdot f$ although this can exceed $\sigma_n$ hugely. No other relative-accuracy-preserving criterion known to me would permit this example to be so deflated.

## §5: Derivations of Bounds for Eigenvalues

Recall $H := H' := \begin{bmatrix} M & B \\ B' & W \end{bmatrix}$ and $Y := Y' := \begin{bmatrix} M & O \\ O' & W \end{bmatrix}$ and their ordered *Spectra* respectively

$$\mathbb{E}(H) = \{ \theta_1 \geq \theta_2 \geq \ldots \geq \theta_n \} \quad \text{and} \quad \mathbb{E}(Y) = \{ \eta_1 \geq \eta_2 \geq \ldots \geq \eta_n \} = \mathbb{E}(M) \cup \mathbb{E}(W)$$

wherein $\mathbb{E}(M) = \{ \mu_1 \geq \mu_2 \geq \ldots \geq \mu_m \}$ and $\mathbb{E}(W) = \{ \omega_1 \geq \omega_2 \geq \ldots \geq \omega_{n-m} \}$.

We shall construct three versions of §3's C designed to connect first a subset of $\mathbb{E}(H)$ with $\mathbb{E}(M)$, then some of $\mathbb{E}(H)$ with $\mathbb{E}(W)$, and then all of $\mathbb{E}(H)$ with $\mathbb{E}(Y) = \mathbb{E}(M) \cup \mathbb{E}(W)$.

First try $C := \begin{bmatrix} I & M^{-1} \cdot B \\ O' & I \end{bmatrix}$ to get $C^{-1\prime} \cdot H \cdot C^{-1} = \begin{bmatrix} M & O \\ O' & \overline{W} \end{bmatrix}$ with $\overline{W} := W - B' \cdot M^{-1} \cdot B$. In this case §3 provides $\|C\|^{\pm 2} = \|C^{-1}\|^{\pm 2} = \exp(\pm 2 \cdot \operatorname{arcsinh}(\|M^{-1} \cdot B\|/2))$. This implies that some subset of m eigenvalues $\theta_j$ in $\mathbb{E}(H)$ are approximated by $\mathbb{E}(M)$ within factors no farther from 1 than are $\exp(\pm 2 \cdot \operatorname{arcsinh}(\|M^{-1} \cdot B\|/2))$. Consequently …

> The relative errors in $\mathbb{E}(M)$ are all smaller than threshold $\tau$ whenever $\|M^{-1} \cdot B\| < \tau$.

However, in the absence of a similar constraint upon $\|B \cdot W^{-1}\|$ too we cannot infer constraints like $\tau$ upon relative errors in $\mathbb{E}(W)$; for an extreme example take $W := O$.

Second try $C' := \begin{bmatrix} I & B \cdot W^{-1} \\ O' & I \end{bmatrix}$ to get $C^{-1\prime} \cdot H \cdot C^{-1} = \begin{bmatrix} \overline{M} & O \\ O' & W \end{bmatrix}$ with $\overline{M} := M - B \cdot W^{-1} \cdot B'$.
In this case $\|C\|^{\pm 2} = \|C^{-1}\|^{\pm 2} = \exp(\pm 2 \cdot \operatorname{arcsinh}(\|B \cdot W^{-1}\|/2))$. This implies that some subset of n–m eigenvalues $\theta_j$ in $\mathbb{E}(H)$ are approximated by $\mathbb{E}(W)$ within factors no farther from 1 than are $\exp(\pm 2 \cdot \operatorname{arcsinh}(\|B \cdot W^{-1}\|/2))$. Consequently …

> The relative errors in $\mathbb{E}(W)$ are all smaller than threshold $\tau$ whenever $\|B \cdot W^{-1}\| < \tau$.

When *both* $\|M^{-1} \cdot B\| < \tau$ *and* $\|B \cdot W^{-1}\| < \tau$, obviously *each* $\eta_j$ in $\mathbb{E}(Y) = \mathbb{E}(M) \cup \mathbb{E}(W)$ approximates *some* $\theta_i$ in $\mathbb{E}(H)$ with relative error no worse than $\tau$. However we have not yet deduced what we wish, namely that *every* $\theta_i$ in $\mathbb{E}(H)$ is approximated by its $\eta_i$ with relative error no worse than $\tau$. So far, our reasoning has yet to preclude that some eigenvalue in $\mathbb{E}(H)$ is approximated within $\tau$ twice, once by an eigenvalue in $\mathbb{E}(M)$ and again by an eigenvalue in $\mathbb{E}(W)$, leaving some other eigenvalue in $\mathbb{E}(H)$ approximated that closely by none in $\mathbb{E}(Y)$.

To preclude that mishap we shall find that $\mathbb{E}(\overline{M})$ approximates $\mathbb{E}(M)$ with relative errors no worse than $\tau^2$ whenever *both* $\|M^{-1} \cdot B\| < \tau$ *and* $\|B \cdot W^{-1}\| < \tau$. Then a matrix K satisfying $\overline{M} = M - B \cdot W^{-1} \cdot B' = (I - K)' \cdot M \cdot (I - K)$ and $\|K\| < \tau^2/2 + O(\tau^4)$ will be constructed, whence Ostrowski's inequality will imply the desired finding that $\mathbb{E}(\overline{M}) \approx \mathbb{E}(M)$ near enough, and also $C := \begin{bmatrix} I - K & O \\ W^{-1} \cdot B' & I \end{bmatrix}$ will have $C^{-1\prime} \cdot H \cdot C^{-1} = Y$ exactly and $\|C^{\pm 1}\|^2 < 1 + \tau + \tau^2 + O(\tau^4)$.

The construction of $K$ begins with the definitions of $G := M^{-1}\cdot B\cdot(B\cdot W^{-1})'/4$ and the matrix-valued function $f(X) := ( X\cdot M^{-1}\cdot X' + B\cdot W^{-1}\cdot B' )/2 = f(X)'$. Starting from $X_0 := O$, iterate $X_{k+1} := f(X_k)$ for $k = 0, 1, 2, 3, \ldots$ in turn. This iteration converges quickly to a fixed-point

$X = f(X) = 2M\cdot(G + G^2 + 2\cdot G^3 + 5\cdot G^4 + 14\cdot G^5 + 42\cdot G^6 + 132\cdot G^7 + 429\cdot G^8 + 1430\cdot G^9 + \ldots)$.

The coefficients in the bracketed series are the coefficients, all integers, of the Taylor series of $x(g) := (1 - \sqrt{(1 - 4g)})/2 = x(g)^2 + g$ around $g = 0$. Since $\|G\| < \tau^2/4 << 1/4$, the series for $X$ converges very fast to $X \approx 2M\cdot G$. Then $K := M^{-1}\cdot X \approx 2G$ turns out to behave as desired, so

$C := \begin{bmatrix} I - K & O \\ W^{-1}\cdot B' & I \end{bmatrix}$ has $H = C'\cdot Y\cdot C$ exactly, and $\|C^{\pm 1}\|^2 < 1 + \tau + O(\tau^2)$, whence our …

**Conclusion:** Whenever ***both*** $\|M^{-1}\cdot B\| < \tau$ ***and*** $\|B\cdot W^{-1}\| < \tau$ then, as claimed, $\mathbb{E}(Y)$ approximates $\mathbb{E}(H)$ with relative errors no worse than $\tau + O(\tau^2)$.

**Examples:** $A := \begin{bmatrix} \tau & 0 & 0 & \tau^2 \\ 0 & 1 & \tau & 0 \\ 0 & \tau & 1 & 0 \\ \tau^2 & 0 & 0 & \tau \end{bmatrix}$ has eigenvalues $1 \pm \tau$ and $(1 \pm \tau)\cdot\tau$ that change to $1, 1, \tau, \tau$ after off-diagonal elements are annihilated. $\|M^{-1}\cdot B\| = \tau << \|M^{-1}\|\cdot\|B\| = 1$ so our new relative error-bounds can come close to best-possible.

However $V := \begin{bmatrix} \tau & 0 & 0 & \tau^2 \\ 0 & 1 & \tau & 0 \\ 0 & \tau & -1 & 0 \\ \tau^2 & 0 & 0 & -\tau \end{bmatrix}$ has eigenvalues $\pm\sqrt{(1 + \tau^2)}$ and $\pm\tau\cdot\sqrt{(1 + \tau^2)}$ that change to $\pm 1, \pm\tau$ after off-diagonal elements are annihilated. $\|M^{-1}\cdot B\| = \|B\cdot W^{-1}\| = \tau$ so our new error-bounds too are capable of extreme pessimism.

## §6: The Quality of Computed Eigenvectors

Besides affecting eigenvalues and singular values, deflation affects eigenvectors and singular vectors. These can be affected drastically, rotated through angles as big as $\pi/4$ in the case of example $A$ above, unless the eigenvalues $\eta_j$ of §5's $Y$ are separated by *relative* gaps adequately wide compared with threshold $\tau$. This is the case for example $V$ above; deflation rotates its eigenvectors through angles like $\tau$. In the absence of hypotheses about spectral gaps, what little can be inferred about the accuracies of eigenvectors computed after our deflation is that their *Residuals* are *Relatively* small in the senses discussed hereunder. …

After $H := \begin{bmatrix} M & B \\ B' & W \end{bmatrix}$ deflates to $Y := \begin{bmatrix} M & O \\ O' & W \end{bmatrix}$ and (part of) its spectrum $\mathbb{E}(Y)$ is accepted as a computed approximation to (part of) $\mathbb{E}(H)$, corresponding eigenvectors of $Y$ will be accepted as computed approximations to corresponding eigenvectors of $H$. Let $\mathbf{y}$ be a normalized eigenvector of $Y$ and $\eta$ its eigenvalue, so $Y\cdot\mathbf{y} = \eta\cdot\mathbf{y}$ and $\|\mathbf{y}\| = 1$. Residual $\mathbf{r} := H\cdot\mathbf{y} - \eta\cdot\mathbf{y}$ indicates how nearly $\mathbf{y}$ approximates an eigenvector $\mathbf{h}$ of $H$ belonging to its eigenvalue $\theta$ approximated by $\eta$. We find $\|\mathbf{r}\| < \tau\cdot|\eta|$ when $\|M^{-1}\cdot B\| < \tau$ and/or $\|B\cdot W^{-1}\| < \tau$ as follows:

Since $\eta \in \mathbb{E}(Y) = \mathbb{E}(M) \cup \mathbb{E}(W)$, $\eta \in \mathbb{E}(M)$ or $\eta \in \mathbb{E}(W)$ or both. For definiteness suppose $\eta = \mu \in \mathbb{E}(M)$, since the alternative can be handled analogously, and let $\mathbf{u}$ be the normalized

eigenvector of $M$ belonging to $\mu$ so that $M\cdot\mathbf{u} = \mu\cdot\mathbf{u}$ and $\|\mathbf{u}\| = 1$. Then $Y$'s eigenvector $\mathbf{y} = \begin{bmatrix}\mathbf{u}\\ \mathbf{o}\end{bmatrix}$, and residual $\mathbf{r} = H\cdot\mathbf{y} - \eta\cdot\mathbf{y} = \begin{bmatrix}\mathbf{o}\\ B'\cdot\mathbf{u}\end{bmatrix}$ has $\|\mathbf{r}\| = \|B'\cdot\mathbf{u}\| = \|(M^{-1}\cdot B)'\cdot\mu\cdot\mathbf{u}\| \le |\mu|\cdot\|M^{-1}\cdot B\|$.

This is why $\|\mathbf{r}\| < \tau\cdot|\eta|$ when $\eta \in \mathbb{E}(M)$ and $\|M^{-1}\cdot B\| < \tau$. Similarly $\|\mathbf{r}\| < \tau\cdot|\eta|$ when $\eta \in \mathbb{E}(W)$ and $\|B\cdot W^{-1}\| < \tau$. Those are the ways in which residual $\mathbf{r}$ is *Relatively* small.

The appearance of $|\eta|$ in the inequality "$\|\mathbf{r}\| < \tau\cdot|\eta|$" is what justifies the term "*Relatively*". It embraces widely disparate eigenvalue magnitudes $|\eta|$ and is important because it explains why our deflation that preserves eigenvalues' relative accuracy also preserves eigenvectors belonging to *Relatively* well-separated eigenvalues. Here are some of the explanation's details:

The *Absolute Spectral Gap* $\gamma$ separating $\hat\eta \in \mathbb{E}(Y)$ from the rest of $\mathbb{E}(Y)$ is defined thus:

$$\gamma := \min\{\, |\eta - \hat\eta| \text{ over all } \eta \in \mathbb{E}(Y) \text{ with } \eta \ne \hat\eta \,\}.$$

Let a *Relative Spectral Gap* $\rho$ separating $\hat\eta \in \mathbb{E}(Y)$ from the rest of $\mathbb{E}(Y)$ be defined thus:

$$\rho := \gamma/|\hat\eta| = \min\{\, |\eta/\hat\eta - 1| \text{ over all } \eta \in \mathbb{E}(Y) \text{ with } \eta \ne \hat\eta \,\}.$$

Two technicalities intrude here. First, for simplicity's sake $\hat\eta$ is assumed a simple eigenvalue of $Y$ whose corresponding normalized eigenvector $\mathbf{y}$ is rotated slightly from the normalized eigenvector $\mathbf{h}$ of $H$ belonging to its simple eigenvalue $\theta$ within $\hat\eta\cdot e^{\pm\tau}$. The angle $\angle(\mathbf{y}, \mathbf{h})$ of that slight rotation is in question here. Second, we shall be concerned with $\hat\eta$ and $\mathbf{y}$ only when gap $\rho >> \tau$ and threshold $\tau << 1$; otherwise $\angle(\mathbf{y}, \mathbf{h})$ can be much bigger than "slight".

According to ch. 11.7 of Parlett's [1998] book and works cited therein, deflation rotates $\mathbf{h}$ through an angle $\angle(\mathbf{y}, \mathbf{h})$ no bigger than about $\|\mathbf{r}\|/\gamma$ when it is small. Above we found that $\|\mathbf{r}\| < \tau\cdot|\hat\eta|$ when our deflation preserves relative accuracy. This implies what was claimed:

Our deflation rotates eigenvector $\mathbf{h}$ through an angle $\angle(\mathbf{y}, \mathbf{h})$ no bigger than about $\tau/\rho$.

The foregoing angle overestimate can be generalized by substituting for $\hat\eta$ a relatively tight cluster of eigenvalues of $Y$ separated from the rest by a sufficiently big relative gap $\rho$. Then the eigenvectors of $Y$ belonging to that eigenvalue cluster $\hat\eta$ span an invariant subspace of $Y$. An analogous invariant subspace of $H$ is spanned by its eigenvectors belonging to the clustered eigenvalues of $H$ falling within $\hat\eta\cdot e^{\pm\tau}$. Then our deflation rotates one invariant subspace onto the other through angles (*cf.* §11.7.1 *etc.* of Parlett [1998]) again no bigger than about $\tau/\rho$.

**Conclusion:** §5's deflations that preserve eigenvalues' relative accuracy also preserve eigenvectors about as well as relative spectral gaps like $\rho$ allow.

## §7: The Quality of Computed Singular Vectors

Let $\phi$ be a singular value of $F$ in §4 with normalized singular vectors $\mathbf{u}$ and $\mathbf{v}$ that satisfy $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$, $F\cdot\mathbf{u} = \phi\cdot\mathbf{v}$ and $\mathbf{v}'\cdot F = \phi\cdot\mathbf{u}'$. After $S$ in §4 has been deflated to $Z$, so $\phi$ in $\mathbb{S}(Z)$ has been accepted as a computed approximation to a singular value of $S$, corresponding

singular vectors $\begin{bmatrix}\mathbf{o}\\ \mathbf{u}\end{bmatrix}$ and $\begin{bmatrix}\mathbf{o}\\ \mathbf{v}\end{bmatrix}$ of Z will be accepted as computed approximations to singular vectors of S. Their residuals are $\mathbf{r} := S\cdot\begin{bmatrix}\mathbf{o}\\ \mathbf{u}\end{bmatrix} - \phi\cdot\begin{bmatrix}\mathbf{o}\\ \mathbf{v}\end{bmatrix} = \begin{bmatrix}E\cdot\mathbf{u}\\ \mathbf{o}\end{bmatrix}$ and $\begin{bmatrix}\mathbf{o}\\ \mathbf{v}\end{bmatrix}'\cdot S - \phi\cdot\begin{bmatrix}\mathbf{o}\\ \mathbf{u}\end{bmatrix}' = \mathbf{o}'$. Now $\|\mathbf{r}\| = \|E\cdot\mathbf{u}\| = \phi\cdot\|E\cdot F^{-1}\cdot\mathbf{v}\| \le \phi\cdot\|E\cdot F^{-1}\| < 2\tau\cdot\phi$ when $\|E\cdot F^{-1}\| < 2\tau$, which is one of R-C. Li's deflation conditions in §4 sufficient to keep relative errors in $\mathbb{S}(Z)$ below threshold $\tau$. This one condition implies *Relatively* small residuals ($\|\mathbf{r}\|/\phi < 2\tau$) for approximate singular vectors of S computed from singular vectors of F in Z, so deflation rotates each of these vectors through angles no bigger than about $2\tau/\rho$ for an appropriate *Relative Gap* $\rho >> \tau$, as in §6, but now between adjacent singular values in $\mathbb{S}(Z)$.

What happens to approximate singular vectors of S computed from singular vectors $\mathbf{x}$ and $\mathbf{y}$ of D in Z when $\|E\cdot F^{-1}\| < 2\tau << 1 << \|D^{-1}\cdot E\|$ ? Now $D\cdot\mathbf{x} = \delta\cdot\mathbf{y}$, $\mathbf{y}'\cdot D = \delta\cdot\mathbf{x}'$, $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, $S\cdot\begin{bmatrix}\mathbf{x}\\ \mathbf{o}\end{bmatrix} - \delta\cdot\begin{bmatrix}\mathbf{y}\\ \mathbf{o}\end{bmatrix} = \mathbf{o}$ and $\mathbf{r}' := \begin{bmatrix}\mathbf{y}\\ \mathbf{o}\end{bmatrix}'\cdot S - \delta\cdot\begin{bmatrix}\mathbf{x}\\ \mathbf{o}\end{bmatrix}' = \begin{bmatrix}\mathbf{o}' & \mathbf{y}'\cdot E\end{bmatrix}$; so now $\|\mathbf{r}'\| = \|\mathbf{y}'\cdot E\| = \delta\cdot\|\mathbf{x}'\cdot D^{-1}\cdot E\|$. But now no reason exists to expect $\|\mathbf{r}'\|/\delta$ to be small. Try $S := \begin{bmatrix}\tau^2 & \tau\\ 0 & 1\end{bmatrix}$ and $Z := \begin{bmatrix}\tau^2 & 0\\ 0 & 1\end{bmatrix}$ for example; their $\|\mathbf{r}'\|/\delta = 1/\tau >> 1$, yet deflation turns singular vectors through angles like $2\tau/\rho$ or smaller anyway. Still to be explained is why this always happens when just one of R-C. Li's deflation conditions in §4, namely $\|E\cdot F^{-1}\| < 2\tau << 1$, is satisfied but not the other; say $\|D^{-1}\cdot E\| >> 1$. Let's see:

The singular vectors of a matrix are just the singular vectors of its inverse swapped. Here $S^{-1} = \begin{bmatrix} D^{-1} & -D^{-1}\cdot E\cdot F^{-1}\\ O' & F^{-1}\end{bmatrix}$; its singular vectors near $\begin{bmatrix}\mathbf{x}\\ \mathbf{o}\end{bmatrix}$ and $\begin{bmatrix}\mathbf{y}\\ \mathbf{o}\end{bmatrix}$ belonging to its singular value near $\delta^{-1}$ have residuals $S^{-1}\cdot\begin{bmatrix}\mathbf{y}\\ \mathbf{o}\end{bmatrix} - \delta^{-1}\cdot\begin{bmatrix}\mathbf{x}\\ \mathbf{o}\end{bmatrix} = \mathbf{o}$ and a new $\mathbf{r}' := \begin{bmatrix}\mathbf{x}\\ \mathbf{o}\end{bmatrix}'\cdot S^{-1} - \delta^{-1}\cdot\begin{bmatrix}\mathbf{y}\\ \mathbf{o}\end{bmatrix}'$; now $\mathbf{r}' = [\mathbf{o}' \quad -\mathbf{x}'\cdot D^{-1}\cdot E\cdot F^{-1}]$ has a relatively small $\|\mathbf{r}'\|/\delta^{-1} = \|\mathbf{y}'\cdot E\cdot F^{-1}\| \le \|E\cdot F^{-1}\| < 2\tau$. The relevant relative spectral gap among singular values $\zeta^{-1}$ of $Z^{-1}$ is $\overline{\rho} := \min_{\zeta\ne\delta}|\zeta^{-1}/\delta^{-1} - 1|$, which turns out to be related to the relevant relative spectral gap $\rho := \min_{\zeta\ne\delta}|\zeta/\delta - 1|$ among the singular values $\zeta$ of Z thus: $\overline{\rho} \ge \rho/(\rho + 1)$ and $\rho \ge \overline{\rho}/(\overline{\rho} + 1)$. If either of $\overline{\rho}$ or $\rho$ is too tiny, the other must be too tiny too. Both relative spectral gaps produce roughly similar over-estimates, big or small, of angles like $2\tau/\rho$ of rotations of singular vectors:

> **Conclusion:** §4's deflation that preserves singular values' relative accuracy also preserves singular vectors about as well as relative spectral gaps like $\rho$ allow.

## §8: Quadratic Relative Error-Bounds and Spectral Gaps

Recall that $H := H' := \begin{bmatrix} M & B \\ B' & W \end{bmatrix}$ and $Y := Y' := \begin{bmatrix} M & O \\ O' & W \end{bmatrix}$ have ordered spectra respectively

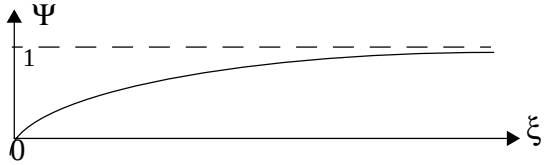
$$\mathbb{E}(H) = \{ \theta_1 \geq \theta_2 \geq \ldots \geq \theta_n \} \quad \text{and} \quad \mathbb{E}(Y) = \{ \eta_1 \geq \eta_2 \geq \ldots \geq \eta_n \} = \mathbb{E}(M) \cup \mathbb{E}(W)$$
$$\text{wherein} \quad \mathbb{E}(M) = \{ \mu_1 \geq \mu_2 \geq \ldots \geq \mu_m \} \quad \text{and} \quad \mathbb{E}(W) = \{ \omega_1 \geq \omega_2 \geq \ldots \geq \omega_{n-m} \} .$$

Our error-bounds upon differences between $\mathbb{E}(H)$ and $\mathbb{E}(Y)$ have been roughly proportional to $B$ so far. When $B$ is small enough, smaller *Quadratic* bounds roughly proportional to $B'\cdot B$ may be available provided $\mathbb{E}(M)$ and $\mathbb{E}(W)$ are separated by sufficiently big and known *Gaps*. Quadratic bounds come with a price: complicated proofs and hypotheses rarely applicable.

In §6 the rotations of eigenvectors by deflation involved gaps $\gamma$ and $\rho$ within $\mathbb{E}(Y)$; those must be distinguished from gaps $\bar\gamma$ and $\Gamma$ defined hereunder to separate $\mathbb{E}(M)$ from $\mathbb{E}(W)$:

The ***Absolute Spectral Gap*** $\bar\gamma$ separating $\eta \in \mathbb{E}(Y)$ from $\mathbb{E}(M)$ or $\mathbb{E}(W)$ is defined thus:
If $\eta \in \mathbb{E}(M)$ then $\bar\gamma(\eta) := \min\{ |\omega - \eta| \text{ over all } \omega \in \mathbb{E}(W) \}$, else
if $\eta \in \mathbb{E}(W)$ then $\bar\gamma(\eta) := \min\{ |\mu - \eta| \text{ over all } \mu \in \mathbb{E}(M) \}$.
Let a ***Relative Spectral Gap*** $\Gamma$ separating $\eta \in \mathbb{E}(Y)$ from $\mathbb{E}(M)$ or $\mathbb{E}(W)$ be defined thus:
If $\eta \in \mathbb{E}(M) \cap \mathbb{E}(W)$ then $\Gamma(\eta) := \bar\gamma(\eta) = 0$; else $\Gamma(\eta) := \bar\gamma(\eta)/|\eta|$.

Let $\Psi(\xi) := \tan\big(\frac{1}{2}\arctan(2\xi)\big) = \tanh\big(\frac{1}{2}\operatorname{arcsinh}(2\xi)\big) = 2\xi\big/\big(1 + \sqrt{(1 + 4\xi^2)}\big)$; among its properties only these will be needed: $0 < d\Psi(\xi)/d\xi \leq 1$;
$\Psi(\xi)/\xi \nearrow 1$ as $\xi \searrow 0$; $\Psi(\xi) \nearrow 1$ as $\xi \nearrow \infty$.
These properties suffice to confirm that
$\Psi(\xi/\gamma)\cdot\xi \leq \min\{ \xi, \xi^2/\gamma \}$ if $\xi > 0$ and $\gamma \geq 0$,
which will be used implicitly and repeatedly.




Optimal quadratic *absolute* error-bounds for eigenvalues come from C-K. Li & R-C. Li [2005]:

$$|\theta_j - \eta_j| \leq \Psi(\|B\|/\bar\gamma(\eta_j))\cdot\|B\| \qquad \mathbb{AB}$$
$$< \min\{ \|B\|, \|B\|^2/\bar\gamma(\eta_j) \} \quad \text{when } \|B\| > 0 \text{ and } \bar\gamma(\eta_j) > 0 .$$

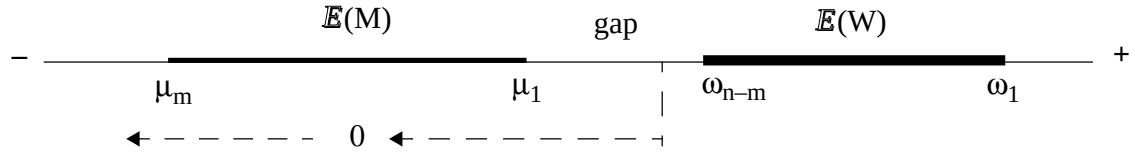
Those *absolute* error-bounds $\mathbb{AB}$ imply immediately these quadratic *relative* error-bounds:

$$|\theta_j/\eta_j - 1| \leq \Psi(\|B/\eta_j\|/\Gamma(\eta_j))\cdot\|B/\eta_j\| . \qquad \mathbb{RAB}$$

These bounds tend to pessimism partly because they are so general, allowing $\mathbb{E}(M)$ and $\mathbb{E}(W)$ to mingle like red and black cards in a shuffled deck, and partly because they use $\|B/\eta_j\|$, in which $\eta_j$ is unlikely to be known when it is needed. To replace $\|B/\eta_j\|$ by something perhaps smaller and maybe cheaper to compute, how much generality must we relinquish? We start by relinquishing mingling; we shall not let the narrowest interval containing $\mathbb{E}(M)$ overlap the narrowest interval containing $\mathbb{E}(W)$. Moreover we shall seek quadratic relative error-bounds only for positive eigenvalues, first the largest of them, then the least.

In conversations with Ren-Cang Li in May 2012 he altered the proof of $\mathbb{AB}$ to get this claim:

Suppose W has all the largest eigenvalues of Y; say every $\eta_j = \omega_j > 0$ for $1 \le j \le n-m$, and $\eta_{n-m} = \omega_{n-m} \ge \eta_{i+n-m} = \mu_i$ for $1 \le i \le m$. Then relative gaps $\Gamma(\omega_j) = 1 - \mu_1/\omega_j$, and
$0 \le \theta_j/\omega_j - 1 \le \Psi(\|B\cdot W^{-1}\|/\Gamma(\omega_j))\cdot\|B\cdot W^{-1}\|$ for $1 \le j \le n-m$. $\mathbb{RBW}$




**Proof:** It starts with $j := 1$; assume $\theta_1 \ne \omega_1$ to leave something to prove. $\theta_1\cdot I - H$ must be positive semidefinite (actually singular), so $\theta_1\cdot I - W$ and $\theta_1\cdot I - M$ must be positive definite. A congruence analogous to the first in §5 implies that $\theta_1\cdot I - W - B'\cdot(\theta_1\cdot I - M)^{-1}\cdot B$ must be positive semidefinite, and then so is $\theta_1\cdot W^{-2} - W^{-1} - (B\cdot W^{-1})'\cdot(\theta_1\cdot I - M)^{-1}\cdot B\cdot W^{-1}$. It must annihilate some column $\mathbf{v}$ normalized so $\mathbf{v}'\cdot\mathbf{v} = 1$, whence follows that

$$\mathbf{v}'\cdot(\theta_1\cdot W^{-2} - W^{-1})\cdot\mathbf{v} = \mathbf{v}'\cdot(B\cdot W^{-1})'\cdot(\theta_1\cdot I - M)^{-1}\cdot B\cdot W^{-1}\cdot\mathbf{v}.$$

The least eigenvalue of $\theta_1\cdot W^{-2} - W^{-1}$ turns out to be $\theta_1/\omega_1^2 - 1/\omega_1$ because $\theta_1/\omega^2 - 1/\omega$ is monotone decreasing on the interval $0 < \omega_{n-m} \le \omega \le \omega_1 < \theta_1$. Therefor the last equation's left-hand side satisfies $\theta_1/\omega_1^2 - 1/\omega_1 \le \mathbf{v}'\cdot(\theta_1\cdot W^{-2} - W^{-1})\cdot\mathbf{v}$. The equation's right-hand side satisfies $\mathbf{v}'\cdot(B\cdot W^{-1})'\cdot(\theta_1\cdot I - M)^{-1}\cdot B\cdot W^{-1}\cdot\mathbf{v} \le \|B\cdot W^{-1}\|^2/(\theta_1 - \mu_1)$. Together the last two inequalities imply $\theta_1/\omega_1 - 1 \le \|B\cdot W^{-1}\|^2/(\theta_1/\omega_1 - 1 + 1 - \mu_1/\omega_1)$, whence $\mathbb{RBW}$ soon follows for $j = 1$.

The rest of the proof goes by induction on n; suppose $\mathbb{RBW}$ is true when H has dimension m+1, m+2, … and n–1, but now H has dimension n. No generality is lost by assuming that W is diagonal since this can be achieved by an orthogonal or unitary similarity that alters no eigenvalue nor norm.. Obtain $\overline{H}$ from H by striking out its row #(m+1) and its column #(m+1), thus reducing W to $\overline{W} := \mathrm{diag}[\omega_2, \omega_3, \ldots, \omega_{n-m}]$ and B to $\overline{B}$ lacking the first column of B, so $\overline{B}\cdot\overline{W}^{-1}$ lacks only the first column of $B\cdot W^{-1}$ and $\|\overline{B}\cdot\overline{W}^{-1}\| \le \|B\cdot W^{-1}\|$. Now $\Gamma(\omega_j)$ is unchanged, and $\mathbb{RBW}$ implies that $\{\overline{\theta}_1, \overline{\theta}_2, \ldots, \overline{\theta}_{n-1}\} = \mathbb{E}(\overline{H})$ satisfies

$0 \le \overline{\theta}_{j-1}/\omega_j - 1 \le \Psi(\|\overline{B}\cdot\overline{W}^{-1}\|/\Gamma(\omega_j))\cdot\|\overline{B}\cdot\overline{W}^{-1}\| \le \Psi(\|B\cdot W^{-1}\|/\Gamma(\omega_j))\cdot\|B\cdot W^{-1}\|$ for $2 \le j \le n-m$.

Repeated appeals to *Cauchy's Interlace Theorem* implying $\omega_j \le \theta_j \le \overline{\theta}_{j-1}$ finish the proof. []

( Cauchy's Interlace Theorem occupies Ch. 10.1 of Parlett's [1998] text. )

$\mathbb{RAB}$ said $\theta_j/\omega_j - 1 \le \Psi(\|B/\omega_j\|/\Gamma(\omega_j))\cdot\|B/\omega_j\|$ under hypotheses assumed for $\mathbb{RBW}$, whose bound may be larger than $\mathbb{RAB}$'s for some small j (big $\omega_j$) but is probably less for bigger j (smaller $\omega_j$). In both error-bounds the critical quantity is $\Gamma(\omega_{n-m}) = 1 - \mu_1/\omega_{n-m}$, one of two smallest relative gaps between $\mathbb{E}$(M) and $\mathbb{E}$(W). Gaps can be hard to (under)estimate usefully. Otherwise neither $\mathbb{RBW}$ nor $\mathbb{RAB}$ imposes requirements upon the signs in $\mathbb{E}$(M); some or all of them may be negative. $\mathbb{RBW}$ outdoes $\mathbb{RAB}$ when $\|B\cdot W^{-1}\|/\|B/\omega_{n-m}\|$ is small.

A slightly different alteration of $\mathbb{AB}$'s proof leads to this claim about lesser eigenvalues:

Suppose H and Y are both positive definite, and the least m eigenvalues $\eta_j$ of Y all come from $\mathbb{E}(M)$, so $W - \mu_1 \cdot I$ must be nonnegative definite. (These hypotheses imply that B is small enough that both $M - B \cdot W^{-1} \cdot B'$ and $W - B' \cdot M^{-1} \cdot B$ are positive definite too.) Then

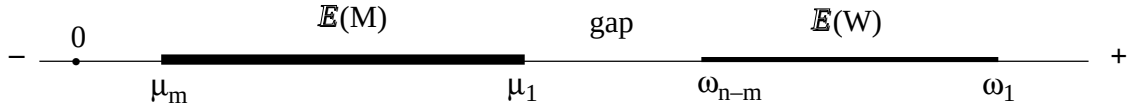
$$0 \le 1 - \theta_{n-m+j}/\mu_j \le \Psi(\|M^{-1/2} \cdot B/\sqrt{\mu_j}\|/\Gamma(\mu_j)) \cdot \|M^{-1/2} \cdot B/\sqrt{\mu_j}\| \quad \text{for } 1 \le j \le m . \qquad \mathbb{R}\sqrt{\mathbb{M}}\mathbb{B}$$

This bound seems unlikely to be useful unless $m = 1$, whereupon it reduces to $\mathbb{RAB}$ above.

A substantial alteration of $\mathbb{AB}$'s proof yields claim $\mathbb{RMB}$ hereunder about lesser eigenvalues:

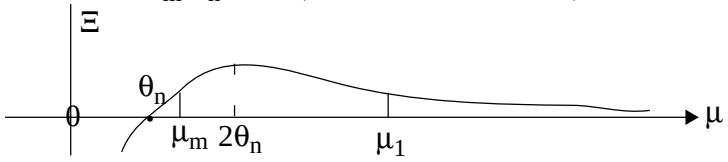
Suppose Y is positive definite and its least m eigenvalues $\eta_j$ all come from $\mathbb{E}(M)$; thus $W - \mu_1 \cdot I$ is positive definite and $\omega_{n-m} > \eta_{n-m+j} = \mu_j > 0$ for $1 \le j \le m$. Further suppose that $\|M^{-1} \cdot B\| < 1/\sqrt{\left((\mu_1/\mu_m)^2 + \mu_1/\mu_m\right)} \le 1/\sqrt{2}$. Then every $\theta_i > 0$ and $\Gamma(\mu_j) = \omega_{n-m}/\mu_j - 1$ and

$$0 \le 1 - \theta_{n-m+j}/\mu_j < \Psi(\|M^{-1} \cdot B\|/\Gamma(\mu_j)) \cdot \|M^{-1} \cdot B\| \quad \text{for } 1 \le j \le m . \qquad \mathbb{RMB}$$




**Proof:** In §5, $\overline{W} := W - B' \cdot M^{-1} \cdot B = (W - \mu_1 \cdot I) + \left(\mu_1 \cdot I - (M^{-1} \cdot B)' \cdot M \cdot (M^{-1} \cdot B)\right)$ is positive definite because of §3 and $\|M^{-1} \cdot B\| < 1/\sqrt{2}$, so H is positive definite, whence every $\theta_i > 0$.

Now induction starts with $j = m$; assume $\theta_n \ne \mu_m$ to leave something to prove. The proof will need $\upsilon := \min\{ (\mu - \theta_n)/\mu^2 \text{ over } \mu_m \le \mu \le \mu_1 \}$; let us see how $\upsilon = (\mu_m - \theta_n)/{\mu_m}^2$ follows from our supposition about $\|M^{-1} \cdot B\|$. Cauchy's Interlace Theorem implies that $\mu_m \ge \theta_n$, so $\upsilon > 0$; and from §5 comes $\mu_m/\theta_n \le \exp\left(2 \cdot \operatorname{arcsinh}(\|M^{-1} \cdot B\|/2)\right) < 2$ because $\|M^{-1} \cdot B\| < 1/\sqrt{2}$.




Since expression $\Xi(\mu) := (\mu - \theta_n)/\mu^2$ reaches its maximum $1/(4\theta_n)$ at $\mu = 2\theta > \mu_m$, we find $\upsilon = \min\{ (\mu_1 - \theta_n)/{\mu_1}^2, (\mu_m - \theta_n)/{\mu_m}^2 \} = (\mu_m - \theta_n)/{\mu_m}^2$ as claimed because, unless $\mu_m = \mu_1$, $\operatorname{sign}\left((\mu_1 - \theta_n)/{\mu_1}^2 - (\mu_m - \theta_n)/{\mu_m}^2\right)$ turns out the same as $\operatorname{sign}\left((1 + \mu_m/\mu_1) \cdot \theta_n/\mu_m - 1\right)$; again §5 supplies $\theta_n/\mu_m \ge \exp\left(-2 \cdot \operatorname{arcsinh}(\|M^{-1} \cdot B\|/2)\right) \ge 1/(1 + \mu_m/\mu_1)$ to confirm the claimed $\upsilon$.

Next, apply a congruence like the second in §5 to $H - \theta_n \cdot I$, which is positive semidefinite (actually singular), to infer that $M - \theta_n \cdot I - B \cdot (W - \theta_n \cdot I)^{-1} \cdot B'$ is positive semidefinite, and then another congruence to find $M^{-1} - \theta_n \cdot M^{-2} - (M^{-1} \cdot B) \cdot (W - \theta_n \cdot I)^{-1} \cdot (M^{-1} \cdot B)'$ positive semidefinite. Its normalized null-vector $\mathbf{v}$ satisfies $\mathbf{v}' \cdot (M^{-1} - \theta_n \cdot M^{-2}) \cdot \mathbf{v} = \mathbf{v}' \cdot (M^{-1} \cdot B) \cdot (W - \theta_n \cdot I)^{-1} \cdot (M^{-1} \cdot B)' \cdot \mathbf{v}$. Combine this equation with $\mathbf{v}' \cdot \mathbf{v} = 1$ and the foregoing value $\upsilon$ of the least eigenvalue of

$M^{-1} - \theta_n \cdot M^{-2}$ to infer that $\upsilon = (\mu_m - \theta_n)/\mu_m^2 \le \|M^{-1}\cdot B\|^2/(\omega_{n-m} - \theta_n)$. From this inequality soon follows the desired result: $0 \le 1 - \theta_n/\mu_m < \Psi(\|M^{-1}\cdot B\|/\Gamma(\mu_m))\cdot\|M^{-1}\cdot B\|$ for $j = m$.

The proof continues by induction on $m$: The induction hypothesis asserts that $\mathbb{RMB}$ is true when $M$ has dimension $1, 2, \ldots, m-1$; but now $M$ has dimension $m$. No generality is lost by assuming that $M = \text{diagonal}[\mu_1, \mu_2, \ldots, \mu_{m-1}, \mu_m]$ since this can be achieved by a unitary or orthogonal similarity that changes no eigenvalue nor norm. Obtain $\overline{H}$ from $H$ by deleting its row #m and column #m, thus repacing $M$ by $\overline{M} := \text{diagonal}[\mu_1, \mu_2, \ldots, \mu_{m-2}, \mu_{m-1}]$ and $B$ by $\overline{B}$ without the last row of $B$, and $M^{-1}\cdot B$ likewise by $\overline{M}^{-1}\cdot\overline{B}$ with $\|\overline{M}^{-1}\cdot\overline{B}\| \le \|M^{-1}\cdot B\|$. Since $\|M^{-1}\cdot B\| < 1/\sqrt{\big((\mu_1/\mu_m)^2 + \mu_1/\mu_m\big)} \le 1/\sqrt{\big((\mu_1/\mu_{m-1})^2 + \mu_1/\mu_{m-1}\big)}$, $\overline{H}$ satisfies $\mathbb{RMB}$'s hypotheses, whence $0 \le 1 - \overline{\theta}_{n-m+j}/\mu_j < \Psi(\|M^{-1}\cdot B\|/\Gamma(\mu_j))\cdot\|M^{-1}\cdot B\|$ for $1 \le j \le m-1$ in which $\overline{\theta}_i$ comes from $\mathbb{E}(\overline{H}) = \{\, \overline{\theta}_1 \ge \overline{\theta}_2 \ge \ldots \ge \overline{\theta}_{n-1} \,\}$. Cauchy's Interlace Theorem puts $\overline{\theta}_i \le \theta_i$ for $1 \le i \le n-1$ and then $0 \le 1 - \theta_{n-m+j}/\mu_j \le 1 - \overline{\theta}_{n-m+j}/\mu_j < \Psi(\|M^{-1}\cdot B\|/\Gamma(\mu_j))\cdot\|M^{-1}\cdot B\|$ for $1 \le j \le m-1$ as well as for $j = m$, completing the proof of $\mathbb{RMB}$. []

Unlike our previous error estimates, $\mathbb{RMB}$ bounds relative errors in $\mathbb{E}(M)$ only if its variation $\mu_1/\mu_m$ is restrained by "$\|M^{-1}\cdot B\| < 1/\sqrt{\big((\mu_1/\mu_m)^2 + \mu_1/\mu_m\big)}$", though this restraint is usually satisfied already when a demand for high accuracy has delayed deflation until $\|M^{-1}\cdot B\|$ is tiny like $\sqrt{\tau}$. Otherwise some such restraint seems unavoidable because of examples like this one:

$$H_3 := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 8 & 8 \\ 0 & 8 & \frac{271}{30} \end{bmatrix}, \quad \theta_n = 1/2, \quad \mu_1 = 1, \quad \mu_m = 8, \quad \|M^{-1}\cdot B\| = 1, \quad \Gamma_1 = 241/30, \quad \text{but}$$
$$1 - \theta_n/\mu_1 = 1/2 \not\le 0.12261 \approx \Psi(\|M^{-1}\cdot B\|/\Gamma_1)\cdot\|M^{-1}\cdot B\|.$$

The example $H_2 := \begin{bmatrix} \cos(2\alpha) & \sin(2\alpha) \\ \sin(2\alpha) & -\cos(2\alpha) \end{bmatrix}$ has eigenvalues $\pm 1$ and eigenvectors $\begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$ that deflation rotates through an angle $\alpha$ while changing eigenvalues by $\pm 2\cdot\sin^2(\alpha)$, which may be negligible though $\alpha$ is not. This example, like example V in §5, reminds us that deflations allowed by negligible *quadratic* error-bounds for `values may rotate `vectors excessively.

**Quadratic Error-Bounds for Singular Values**

Recall upper-triangles $S$ and its deflation $Z$ and their ordered nonnegative singular values:

$$S := \begin{bmatrix} D & E \\ O' & F \end{bmatrix}, \quad \mathbb{S}(S) = \{\, \sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_n \,\}; \quad Z := \begin{bmatrix} D & O \\ O' & F \end{bmatrix}, \quad \mathbb{S}(Z) = \{\, \zeta_1 \ge \zeta_2 \ge \ldots \ge \zeta_n \,\}.$$

Here $\mathbb{S}(Z) = \mathbb{S}(D) \cup \mathbb{S}(F)$, wherein $\mathbb{S}(D) = \{\, \delta_1 \ge \delta_2 \ge \ldots \ge \delta_m \,\}$ will have to be distinguished from $\mathbb{S}(F) = \{\, \phi_1 \ge \phi_2 \ge \ldots \ge \phi_{n-m} \,\}$ by gaps defined in a way now familiar:

The ***Absolute Spectral Gap*** $\bar\gamma$ separating $\zeta \in \mathbb{S}(Z)$ from $\mathbb{S}(D)$ or $\mathbb{S}(F)$ is defined thus:

If $\zeta \in \mathbb{S}(D)$ then $\bar\gamma(\zeta) := \min\{\, |\phi - \zeta| \text{ over all } \phi \in \mathbb{S}(F)\,\}$, else

if $\zeta \in \mathbb{S}(F)$ then $\bar\gamma(\zeta) := \min\{\, |\delta - \zeta| \text{ over all } \delta \in \mathbb{S}(D)\,\}$.

Let a ***Relative Spectral Gap*** $\Gamma$ separating $\zeta \in \mathbb{S}(Z)$ from $\mathbb{S}(D)$ or $\mathbb{S}(F)$ be defined thus:

If $\zeta \in \mathbb{S}(D) \cap \mathbb{S}(F)$ then $\Gamma(\zeta) := \bar\gamma(\zeta) = 0$; else $\Gamma(\zeta) := \bar\gamma(\zeta)/\zeta$.

Yes, the gap-functions $\bar\gamma(\xi)$ and $\Gamma(\xi)$ are *overloaded* according to whether their argument $\xi$ comes from $\mathbb{E}(Y)$ or from $\mathbb{S}(Z)$; let's hope their context will preclude confusion.

Li & Li [2005] used $\mathbb{AB}$ to derive similar quadratic *absolute* error-bounds for the singular values of $S = \begin{bmatrix} D & E \\ O' & F \end{bmatrix}$ because they and their negatives are the eigenvalues of $\begin{bmatrix} O & D' & O & O \\ D & O & E & O \\ O' & E' & O & F' \\ O' & O' & F & O \end{bmatrix}$:

$$|\sigma_j - \zeta_j| \le \Psi(\|E\|/\bar\gamma(\zeta_j))\cdot\|E\| \qquad \mathbb{AE}$$
$$< \min\{\, \|E\|,\ \|E\|^2/\bar\gamma(\zeta_j)\,\} \quad \text{when } \|E\| > 0 \text{ and } \bar\gamma(\zeta_j) > 0 .$$

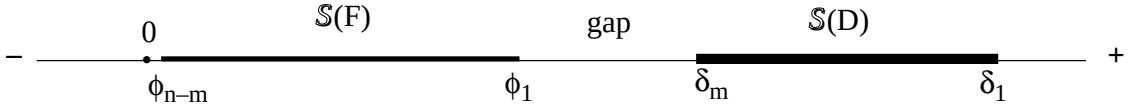
Those *absolute* error-bounds $\mathbb{AE}$ imply immediately these quadratic *relative* error-bounds:

$$|\sigma_j/\zeta_j - 1| \le \Psi(\|E/\zeta_j\|/\Gamma(\zeta_j))\cdot\|E/\zeta_j\| . \qquad \mathbb{RAE}$$

As we did for $\mathbb{RAB}$, we shall try to replace $\|E/\zeta_j\|$ by something perhaps smaller and cheaper to compute. To do so we shall again relinquish some of $\mathbb{RAE}$'s generality by assuming that a sufficiently wide relative gap separates the smallest interval containing $\mathbb{S}(D)$ from the smallest interval containing $\mathbb{S}(F)$. Here is an analog of R-C. Li's $\mathbb{RBW}$, but proved differently:

Suppose $\mathbb{S}(D)$ has all the largest singular values of $Z$, so every $\zeta_j = \delta_j$ for $1 \le j \le m$, and $\zeta_m = \delta_m > \zeta_{m+1} = \phi_1$. Let gaps $G_j := 1 - (\phi_1/\delta_j)^2 - \|D^{-1}\cdot E\|^2$ for $1 \le j \le m$. If $G_j > 0$ then

$$0 \le (\sigma_j/\delta_j)^2 - 1 \le \Psi(\|D^{-1}\cdot E\|/G_j)\cdot\|D^{-1}\cdot E\| . \qquad \mathbb{RDE}$$




**Proof:** It starts with $j = 1$; assume $\sigma_1 \ne \delta_1$ to leave something to prove. To simplify notation temporarily, drop the subscripts from $\sigma := \sigma_1$, $\delta := \delta_1$ and $\phi := \phi_1$. Then $\sigma^2$ is the largest eigenvalue of $S\cdot S' = \begin{bmatrix} D\cdot D' + E\cdot E' & E\cdot F' \\ F\cdot E' & F\cdot F' \end{bmatrix}$, so $\sigma^2 > \delta^2 > \phi^2$. No generality is lost by assuming temporarily that $D$ and $F$ are diagonals of their respective singular values. Then $\sigma^2\cdot I - S\cdot S'$ is positive semidefinite (and singular), and congruences now familiar establish the same for first $\sigma^2\cdot I - D^2 - E\cdot E' - E\cdot F'\cdot(\sigma^2\cdot I - F^2)^{-1}\cdot F\cdot E' = \sigma^2\cdot I - D^2 - \sigma^2\cdot E\cdot(\sigma^2\cdot I - F^2)^{-1}\cdot E'$ and then $\sigma^2\cdot D^{-2} - I - \sigma^2\cdot(D^{-1}\cdot E)\cdot(\sigma^2\cdot I - F^2)^{-1}\cdot(D^{-1}\cdot E)'$. Its unit null-vector $\mathbf{v}$ satisfies $\mathbf{v}'\cdot\mathbf{v} = 1$ and $(\sigma/\delta)^2 - 1 \le \mathbf{v}'\cdot(\sigma^2\cdot D^{-2} - I)\cdot\mathbf{v} = \sigma^2\cdot\mathbf{v}'\cdot(D^{-1}\cdot E)\cdot(\sigma^2\cdot I - F^2)^{-1}\cdot(D^{-1}\cdot E)'\cdot\mathbf{v} \le \sigma^2\cdot\|D^{-1}\cdot E\|^2/(\sigma^2 - \phi^2)$. This inequality implies that $(\sigma_1/\delta_1)^2 - 1 \le \Psi(\|D^{-1}\cdot E\|/G_1)\cdot\|D^{-1}\cdot E\|$, which is $\mathbb{RDE}$ for $j = 1$.

The proof continues by induction on the dimension of $D$. The induction hypothesis is that $\mathbb{RDE}$ is valid for dimensions 1, 2, … and m–1; but now $D$ has dimension $m$. Still assuming

D and F are diagonal, obtain $\overline{S}$ from S by striking off its first row and column. Doing so replaces D by $\overline{D} = \text{diagonal}[\delta_2, \delta_3, \ldots, \delta_{m-1}, \delta_m]$, E by $\overline{E}$ lacking the first row of E, and likewise $D^{-1}\cdot E$ by $\overline{D}^{-1}\cdot\overline{E}$ so that $\|\overline{D}^{-1}\cdot\overline{E}\| \le \|D^{-1}\cdot E\|$ and $\overline{G}_{j-1} := 1 - (\phi_1/\delta_j)^2 - \|\overline{D}^{-1}\cdot\overline{E}\|^2 \ge G_j$. And $\overline{S}\cdot\overline{S}'$ is just $S\cdot S'$ shorn of its first row and column, so Cauchy's Interlace Theorem tells us $\sigma_j^2 \le \sigma_{j-1}^2$ for $2 \le j \le m$ as well as $\delta_j^2 \le \sigma_j^2$ inferred from $S'\cdot S$. The induction hypothesis implies $0 \le (\sigma_j/\delta_j)^2 - 1 \le (\sigma_{j-1}/\delta_j)^2 - 1 \le \Psi(\|\overline{D}^{-1}\cdot\overline{E}\|/\overline{G}_{j-1})\cdot\|\overline{D}^{-1}\cdot\overline{E}\| \le \Psi(\|D^{-1}\cdot E\|/G_j)\cdot\|D^{-1}\cdot E\|$ for each $G_j > 0$ in $2 \le j \le m$, after which unravelling the diagonalizations of D and F finishes the proof of $\mathbb{RDE}$. []

$\mathbb{RDE}$'s requirement "$G_j > 0$" is an annoying complication, perhaps superfluous, almost surely immaterial because relative-accuracy-preserving deflations allowed by $\mathbb{RDE}$ will occur only when $\|D^{-1}\cdot E\|^2 < \tau << 1 - (\phi_1/\delta_m)^2$, whereupon every $G_j \approx 1 - (\phi_1/\delta_j)^2 > \Gamma(\delta_j)$.

To cope with the smallest singular values of S a trick used in §7 will be used again here. The deflation of $S^{-1} = \begin{bmatrix} D^{-1} & -D^{-1}\cdot E\cdot F^{-1} \\ O' & F^{-1} \end{bmatrix}$ produces $Z^{-1} = \begin{bmatrix} D^{-1} & O \\ O' & F^{-1} \end{bmatrix}$; their singular values are the reciprocals of the singular values of S and Z respectively. Applying $\mathbb{RDE}$ to $S^{-1}$ (with F increased infinitesimally perhaps) produces the inequalities in $\mathbb{REF}$ hereunder:

Suppose F has all the smallest singular values in $\mathbb{S}(Z)$, so every $\zeta_{m+j} = \phi_j < \zeta_m = \delta_m$ for $1 \le j \le n-m$. If $\phi_j = 0$ then $\sigma_{m+j} = 0$; otherwise if gap $\overline{G}_j := 1 - (\phi_j/\delta_m)^2 - \|D^{-1}\cdot E\|^2 > 0$ then

$$0 \le (\phi_j/\sigma_{m+j})^2 - 1 \le \Psi(\|D^{-1}\cdot E\|/\overline{G}_j)\cdot\|D^{-1}\cdot E\| \quad \text{for } 1 \le j \le n-m. \qquad \mathbb{REF}$$


The reappearance of $\|D^{-1}\cdot E\|$ in $\mathbb{REF}$ is not a typo. It emerges from the proof and reassures us that a deflation preserving relative accuracy in *all* the biggest singular values does about the same for the smallest, no matter how small they are, and *vice-versa*. It is reminiscent of §4.

However, unlike §4 and §7, quadratic error-bounds may permit deflations that alter singular values at worst tolerably while rotating singular vectors intolerably.

• • • • • • •

The four quadratic relative error-bounds $\mathbb{RBW}$, $\mathbb{RMB}$, $\mathbb{RDE}$ and $\mathbb{REF}$ proved above are believed to be new but not unprecedented. An antecedent is Theorem 5 on p. 881 of Demmel & Kahan [1990]. It is more complicated and weaker than the $\mathbb{R}\ldots$ bounds. Like all quadratic error-bounds, absolute and relative, ours require adequate underestimates of spectral gaps costly to compute except for special matrices, among them those dominated enough by their diagonals. Gap estimation will be discussed next.

**Estimating Spectral Gaps**

All four new $\mathbb{R}_{\cdots}$ bounds have the form "Relative Error $\le \Psi(\beta/\Gamma)\cdot\beta$" in which $\beta$ stands for a small (over)estimate of $\|D^{-1}\cdot E\|$ or $\|M^{-1}\cdot B\|$ or $\|B\cdot W^{-1}\|$, and $\Gamma$ stands for an (under)estimate of a relative gap. Applications of these bounds evade computation of the function $\Psi$ because the predicate "$\Psi(\beta/\Gamma)\cdot\beta < \tau$" simplifies to "$\beta^2 < (\tau+\Gamma)\cdot\tau$". To compute $\Gamma$ is not that simple.

Except in $\mathbb{RBW}$ sometimes, $\Gamma$ estimates the relative difference between the largest singular value of one submatrix and the larger least singular value of another submatrix. These involve norms: ( C's largest singular value ) = $\|C\|$ ; ( C's least singular value ) = $1/\|C^{-1}\|$ .

In Table 1 below, $\mathbb{RBW}$ and $\mathbb{RMB}$ are assumed used together, as are $\mathbb{RDE}$ and $\mathbb{REF}$.

**Table 1: Gaps that $\Gamma$ must (under)estimate**

| Bound | Minimum Relative Gap |
|---|---|
| $\mathbb{RBW}$ & $\mathbb{RMB}$ | $\Gamma(\omega_{n-m}) = 1 - \|M\|\cdot\|W^{-1}\|$ |
| | $\Gamma(\mu_1) = 1/(\|M\|\cdot\|W^{-1}\|) - 1$ |
| $\mathbb{RDE}$ & $\mathbb{REF}$ | $G_m = \overline{G}_1 = 1 - \|F\|^2\cdot\|D^{-1}\|^2 - \|D^{-1}\cdot E\|^2$ |

The table's formulas reveal why all the norms must be overestimated relatively tightly to yield usable underestimates $\Gamma > 0$. Explained hereunder is why relatively tight overestimates of $\|\ldots\|$ tend to be costly to compute; exceptions are matrices dominated enough by their diagonals.

Error-analyses frequently approximate $\|C\|$ by another norm cheaper to compute; an example is $\|C\|_\infty := \max_i \sum_j |c_{ij}|$, the biggest-row-sum norm. Then $\|C\| \le \|C\|_\infty \le \sqrt{m}\cdot\|C\|$ for m-by-m matrices C; and examples C exist making either inequality an equality. When m is big there are m-by-m triangular examples F for which both $\|F\|_\infty/\|F\|$ and $\|F'\|_\infty/\|F\|$ fall at most a few percent short of $\sqrt{m}$, so $\|\ldots\|_\infty$ is far from a tight approximation to $\|\ldots\|$ for arbitrary triangles. And there are big m-by-m positive definite examples M for which $\|M\|_\infty/\|M\|$ is only a little less than 2, which is rarely tight enough. So far as I know, tight estimates of $\|\ldots\|$ cost far more than $O(m^2)$ work unless they are obtained for special matrices like those known to have rank far less than their dimensions, or matrices dominated enough by their diagonals; or else the estimates are probabilistic.

Probabilistic estimates are generated by iterations that almost always converge rapidly to $\|\ldots\|$ from below. After a few iterations the iterate is expected to fall short of $\|\ldots\|$ by at most a few percent; then adding a few percent more is expected to overestimate $\|\ldots\|$ only slightly. For example, if M is positive (semi)definite then $\|M^{k+1}\cdot\mathbf{x}\|/\|M^k\cdot\mathbf{x}\| \nearrow \|M\|$ as $k \nearrow \infty$ unless $\mathbf{x}$ is a very unlikely choice. Another example, motivated by the formula "$\|F\| = \max_{\mathbf{x}\neq\mathbf{0}} \|F\cdot\mathbf{x}\|/\|\mathbf{x}\|$", generates a sequence of vectors $\mathbf{x}$ that follow the upward gradients of $\|F\cdot\mathbf{x}\|/\|\mathbf{x}\|$ from an initial choice of $\mathbf{x}$; convergence is fast from almost any initial choice. The trouble with probabilistic estimates is their lack of inexpensive ways to expose bad luck which, however unlikely, befalls every day a few at least of the billions of computed estimates.

Estimates of $\|W^{-1}\|$ and $\|D^{-1}\|$ incur another layer of uncertainty and expense. According to Demmel *et al*. [2001], there are reasons to expect any estimator of $\|C^{-1}\|$ substantially cheaper than the cost of computing $C^{-1}$ to over/underestimate $\|C^{-1}\|$ substantially for some matrices C. Nothing unexpected appears among the overestimators surveyed by N.J. Higham [1987]; for an m-by-m triangular D they all incur at least m divisions to overestimate $\|D^{-1}\|$, often grossly even if it is bidiagonal unless it is dominated enough by its diagonal.

C is deemed *Dominated by its Diagonal* row-wise when every $|c_{ii}| > \sum_{j\neq i} |c_{ij}|$; column-wise dominance is defined analogously. These dominances dominate more than necessary for cheap and fairly tight overestimates of $\|C\|$ and $\|C^{-1}\|$ whenever such are available. The formulas in Table 2 hereunder help produce cheaply the gap's underestimates needed in Table 1:

**Table 2: Tight Estimates of ||…|| for a Matrix Dominated Enough by its Diagonal**

| ℝ… bounds | Estimates that Straddle ‖…‖ |
|---|---|
| ℝBW & ℝMB | $\max_j\{m_{jj}\} \le \|M\| \le \|M\|_\infty = \max_j\{ \sum_k \|m_{jk}\| \}$ |
| | $1/\min_j\{w_{jj}\} \le \|W^{-1}\| \le 1/\min_j\{ w_{jj} - \sum_{k\neq j} \|w_{jk}\| \}$ if positive |
| ℝDE & ℝEF | $\max_j\{\|f_{jj}\|\} \le \|F\| \le \max\{ \max_j\{\sum_{k\ge j} \|f_{jk}\|\}, \max_j\{\sum_{k\le j} \|f_{kj}\|\} \}$ |
| | $1/\min_j\{\|d_{jj}\|\} \le \|D^{-1}\| \le 1/\min_j \{ \|d_{jj}\| - (\sum_{k<j} \|d_{kj}\| + \sum_{k>j} \|d_{jk}\|)/2 \}$ if positive |

The estimates for positive definite M and $W^{-1}$ come from Gershgorin's Circles Theorem, for which see §14.2 of Hogben [2007], or see p. 54 of Kahan [2012'] for a very condensed proof. The estimate for an upper triangle F comes from Gershgorin's Circles Theorem applied to $\begin{bmatrix} O & F \\ F' & O \end{bmatrix}$, whose eigenvalues are the singular values of F and their negatives. The estimate for an upper triangle $D^{-1}$ comes from Johnson [1989]. Table 2's estimates are most convenient when the matrices in question are tridiagonal or bidiagonal.

For complex matrices, estimates depending heavily upon elements' magnitudes can be sped up by saving a square root per element thus: First multiply by whatever diagonal unitary matrices make every diagonal element of S nonnegative; then overestimate $|\xi + \imath\cdot\eta| = \sqrt{(\xi^2 + \eta^2)}$ by $\max\{|\xi|, |\eta|\} + (\sqrt{2} - 1)\cdot\min\{|\xi|, |\eta|\}$, which exceeds $|\xi + \imath\cdot\eta|$ by less than 8.25% .

> Considering how expensive are worthwhile estimates of spectral gaps, and how rarely deflation is permitted by quadratic error-bounds, what good are they? Perhaps they serve here mostly to explain why the non-quadratic bounds of §4 and §5 are so often so pessimistic though best-possible without estimates of gaps.

The last words about quadratic relative error-bounds seem unlikely to have been written yet.

## §9: Application to Tests of Computed Eigenvalues' Relative Accuracies

The Rayleigh-Ritz method (ch. 11 of Parlett [1998]) will be used here to corroborate claims of high relative accuracy for some of the computed eigenvalues and eigenvectors of a given big n-by-n matrix $A = A'$. A subset of its approximated eigenvalues are arrayed as an m-by-m diagonal $\Lambda$, and corresponding approximated eigenvectors are the columns of n-by-m matrix $\overline{Q}$. Normally these columns are orthonormal or very nearly so. To clean them up, compute their residual $V := I - \overline{Q}'\cdot\overline{Q}$, preferably accumulating scalar products extra-precisely, and then replace $\overline{Q}$ by $Q := \overline{Q} + \overline{Q}\cdot V/2$ whose residual, $I - Q'\cdot Q = (3V^2 + V^3)/4 \pm$ roundoff, need not be computed if it is predictably tiny enough to drown in roundoff.

The next task is to replace $\Lambda$ by an improved approximation M and determine whether its m eigenvalues approximate some m of A's with high relative accuracy. To this end compute a temporary residual $\overline{R} := A\cdot Q - Q\cdot\Lambda$, preferably accumulating scalar products extra-precisely, and then $\overline{\Delta\Lambda} := Q'\cdot\overline{R}$, which would be symmetric but for roundoff. Next replace $\overline{\Delta\Lambda}$ by its symmetric part $\Delta\Lambda := (\overline{\Delta\Lambda} + \overline{\Delta\Lambda}')/2$ and then compute $M := \Lambda + \Delta\Lambda$ and $R := \overline{R} - Q\cdot\Delta\Lambda$. Now this final residual $R \approx (A\cdot Q - Q\cdot M \pm$ roundoff) and $Q'\cdot R \approx (O \pm$ roundoff). Ideally M should be so nearly diagonal that $R\cdot M^{-1}$ can be computed easily; finally $\|R\cdot M^{-1}\|$ is the desired upper bound upon the relative errors in the eigenvalues of M. These can be computed as corrections to $\Lambda$ by a pass of Jacobi's iteration (ch. 9 in Parlett [1998]) as described by Drmač in §46 of Hogben [2007]. That iteration's rotations should postmultiply Q to update its columns, which then approximate eigenvectors of A more closely than before.

Why is $\|R\cdot M^{-1}\|$ an upper bound upon the relative errors in $\mathbb{E}(M)$? An explanation ignoring roundoff begins with the augmentation of Q to a notional (not computed) n-by-n orthogonal matrix $[Q, P] = ([Q, P]')^{-1}$. Notional $H := [Q, P]'\cdot A\cdot[Q, P] = \begin{bmatrix} M & B \\ B' & W \end{bmatrix}$ has $[Q, P]'\cdot R = \begin{bmatrix} O \\ B' \end{bmatrix}$, $\mathbb{E}(H) = \mathbb{E}(A)$ and $\|M^{-1}\cdot B\| = \|R\cdot M^{-1}\|$, whence §5's first error-bound becomes applicable.

That error-bound is applicable also when H is the result of a method like Lanczos' (ch. 13 in Parlett [1998]) that reduces (part of) A to tridiagonal form. Now M is tridiagonal and only the lower leftmost element of B has a substantial magnitude. In this case the computation of $M^{-1}\cdot B$ and its norm costs time proportional to m, or at worst $m\cdot(n-m)$ if reorthogonalization applied to Q has scattered throughout B small elements not small enough to ignore.

In case only the first or last m eigenvalues of A are desired, it is prudent to compute several more than m to provide a reason to believe that the first or last m are among them. Doing so often exposes a gap between some eigenvalues and the rest, thus permitting quadratic error-bounds like those in §8 above to be applied.

## §10: Application to Deflations during Singular Value Computations by dqds

Here R-C. Li's criterion in §4 will be applied as he intended to the deflation of a bidiagonal matrix during dqds iterations (Parlett *et al*. [1994, 2000]) to compute its singular values. But first the dqds process will be described as briefly as possible, which is not briefly at all.

The process computes squared singular values of bidiagonal upper triangle $S$ as eigenvalues of symmetric tridiagonal $T := S \cdot S'$ (or $S' \cdot S$, whose eigenvalues are the same) without having to compute the elements of $S$ or $T$ explicitly. Instead arrays $\{q_j\}$ and $\{e_j\}$ are computed where

$$S \cdot S' = T = \text{tridiag} \begin{bmatrix} & \sqrt{q_2 \cdot e_1} & & \sqrt{q_3 \cdot e_2} & & \cdots & & \sqrt{q_{n-1} \cdot e_{n-2}} & & \sqrt{q_n \cdot e_{n-1}} & \\ q_1 + e_1 & & q_2 + e_2 & & q_3 + e_3 & \cdots & q_{n-1} + e_{n-1} & & q_n \\ & \sqrt{q_2 \cdot e_1} & & \sqrt{q_3 \cdot e_2} & & \cdots & & \sqrt{q_{n-1} \cdot e_{n-2}} & & \sqrt{q_n \cdot e_{n-1}} & \end{bmatrix} ;$$

$\{\pm\sqrt{\bar{q}_j}\}$ lies on the diagonal of $S$ and $\{\pm\sqrt{\bar{e}_j}\}$ on its superdiagonal. The signs do not matter.

There are two reasons for the dqds process to act upon $\{q_j\}$ and $\{e_j\}$ instead of $S$ or $T$. The obvious reason is that no square root will be needed in the innermost loop of the process. The unobvious reason is that rounding off explicitly computed diagonal elements of $T$ can ruin its tiniest eigenvalue(s). (This happens to §4's example $S$ when its dimension $n$ is big and $s$ in $S := \text{bidiag} \begin{bmatrix} s & s & \cdots & s & s & e \\ 1 & 1 & \cdots & 1 & 1 & f \end{bmatrix}$ is any big number whose $s^2$ incurs a rounding error.)

Starting from $S_0 := S$, the dqds iteration chooses nonnegative *Shifts* $ß_k$ and obtains $S_{k+1}$ as the Cholesky factor of $S_k \cdot S'_k - ß_k \cdot I = S'_{k+1} \cdot S_{k+1}$ for $k = 0, 1, 2, 3, \ldots$ in turn. This works only if $ß_k$ does not exceed the least squared singular value of $S_k$, but the closer the better. A good choice for shift $ß_k$ raises difficult questions that will not be discussed here; see Parlett *et al*. [1994, 2000] and other works in progress. The squared singular values of $S_0$ exceed those of $S_{k+1}$ respectively by $\sum ß_k := ß_0 + ß_1 + \ldots + ß_k$. Iteration drives $S_k$ towards a diagonal.

Typically the shifts dwindle until the last one or two become zero and/or deflation occurs.

**Deflation is the subject to be discussed at length here.**

To simplify the discussion the iteration's subscript $k$ will be dropped. One dqds($ß$) iteration maps $S$, represented by $\{q_j\}$ and $\{e_j\}$, to $\bar{S}$ represented by $\{\bar{q}_j\}$ and $\{\bar{e}_j\}$ and satisfying $\bar{S}' \cdot \bar{S} = S \cdot S' - ß \cdot I$; then $\sum ß$ is updated to $\sum \underline{ß} := \sum ß + ß$. The iteration's inner loop goes thus:

**dqds(ß)**

$d_1 := q_1 - ß$ ; dmin := $d_1$ ; jmin := 1 ;
for j = 1 to n–1 do {
  $\bar{q}_j := d_j + e_j$ ; $t := q_{j+1} / \bar{q}_j$ ; … The division overlaps the next …
    if ($\bar{q}_j \le 0$) then {break out to *Early Failure, q.v.*} ;
    if ($d_j \le$ dmin) then {jmin := j ; dmin := $d_j$} ;
  $\bar{e}_j := t \cdot e_j$ ; $d_{j+1} := t \cdot d_j - ß$ } ; … end of j-loop
$\bar{q}_n := d_n$ ;
if ($d_n < 0$) then {go to *Late Failure, q.v.*}
  elseif ($d_n \le$ dmin) then {jmin := n ; dmin := $d_n$} .

*Early Failure* here differs slightly from its definition by Parlett *et al*. [1994, 2000], who break out when $d_j < 0$ and $j < n$. Here the count of negative values $\bar{q}_j$ turns out equal to the count of eigenvalues of $S \cdot S'$ less than $ß$, so it is worth recording for future use when our failure occurs

at $j = n-1$. The count of values $d_j < 0$ is less informative. However, dqds(0) cannot fail, and then the last diagonal element of the Cholesky factor of the leading j-by-j principal submatrix of $S \cdot S'$ turns out to be $\sqrt{\bar{d}_j}$, whence the last diagonal element of the inverse of the leading j-by-j principal submatrix of $S$ turns out to be $1/\sqrt{\bar{d}_j}$, which will figure in R-C. Li's deflation criterion below. Then too, according to §6.1 on pp. 204-6 of Fernando & Parlett [1994], the $j^{th}$ diagonal enement of $(S \cdot S')^{-1}$ is $1/d_j$, which will figure in Ming Gu's criterion below.

*Early Failure* occurs when ß is much too big, bigger than the second-least squared singular value of $S$. *Late Failure* occurs when ß is a little too big, between the least and second-least squared singular values. In both cases $\{\bar{q}_j\}$ and $\{\bar{e}_j\}$ will be discarded, a new smaller shift ß will be chosen somehow, and dqds(ß) will be tried again. If it succeeds, dmin > 0 and jmin figure in the computation of an improved (smaller) upper bound for the least squared singular value of $\bar{S}$ to help choose the next shift $\bar{ß}$; and the updated $\sum\bar{ß}$ will affect revised tolerances like $\tau$ for permissible deflations that preserve relative accuracies among the final results.

Here "deflation" means that an off-diagonal element $\sqrt{\bar{e}_j}$ is deemed negligible and set to 0. Failure to deflate in a timely fashion can inhibit the iterations' convergence. Started with every $e_j > 0$, iteration drives array $\{e_j\}$ towards zeros and drives array $\{q_j + \sum ß\}$ towards squared singular values of $S_0$ in descending order, as if big singular values migrated upward and small ones downward. But their migration is obstructed when some relatively tiny non-negligible $e_j$ has small singular values above it and big below, a situation called "Disordered Data". This is why …

> Numerous tests and branches complicate the dqds process with attempts to exploit every permissible deflation without wasting too much time rejecting impermissible deflations.

The initial iteration, and the first after every deflation, is a dqds(0) modified to scrutinize each $e_j$ for a permissible deflation. Li's condition "$\|D^{-1} \cdot E\| < 2\tau$" in §4 becomes "$e_j < (4\tau^2) \cdot d_j$" to permit the annihilation of $e_j$. Li's condition "$\|E \cdot F^{-1}\| < 2\tau$" becomes "$\bar{e}_{n-1} < (4\tau^2) \cdot \bar{q}_n$" to permit the annihilation of $\bar{e}_{n-1}$. Both possibilities can arise also after a successful dqds(ß) with $ß > 0$ since then $d_j > 0$ is a decreasing function of ß. But those conditions are too stringent. Conditions less stringent are desired intensely in the hope that they permit earlier deflations.

The search for less stringent deflation conditions begins with the observation that, with each iteration, $\sum ß$ typically increases by ß to $\sum\bar{ß}$ while a modest overestimate of $\|S\|^2$, namely

$$\tilde{n}^2 := \left( \sqrt{(\max_j\{q_j\})} + \sqrt{(\max_i\{e_i\})} \right)^2, \qquad \text{(to be refreshed infrequently)}$$

decreases by a similar amount. Because $\tilde{n}^2 \geq \|S\|^2 = \|S_0\|^2 - \sum ß$, the squared singular values of $S_0$ yet to be computed must all lie between $\sum ß$ and $\tilde{n}^2 + \sum ß$. An absolute error smaller than $\tau \cdot \sigma$ in a not-yet-computed singular value $\sigma$ of $S_0$ induces a relative error no worse than $\tau$. If this is tolerable, then $\sqrt{\bar{e}_j}$ in $S$ turns out to be eligible for annihilation whenever

$$e_j < \tau^2 \cdot \left( \text{if } \tilde{n}^2 > \sum ß \text{ then } 4\sum ß \text{ else } \tilde{n}^2 \cdot (1 + \sum ß/\tilde{n}^2)^2 \right) \qquad \text{AbsEtest}$$

because doing so changes no singular value of $S$ by more than $\sqrt{\bar{e}_j}$, and thus no singular value

$\sigma$ of $S_0$ by more than about $\tau\cdot\sigma$. Similar reasoning replaces R-C. Li's conditions above for annihilating $e_j$ during dqds(ß) by

$$e_j < 4\tau^2\cdot(1 + \sum ß/\tilde{n}^2)^2\cdot d_j \quad \text{and} \quad e_{n-1} < 4\tau^2\cdot(1 + \sum ß/\tilde{n}^2)^2\cdot d_n\,. \qquad \mathbb{RelEtest}$$

Actually both $e_j$ and $\bar{e}_j := t\cdot e_j$ get set to 0 when $\mathbb{AbsEtest}$ or $\mathbb{RelEtest}$ is satisfied. No harm is done if dqds(ß) fails subsequently because, after $\{\bar{q}_j|$ and $\{\bar{e}_j\}$ are discarded and dqds(ß) restarted with a smaller ß, the recomputed $d_j$ will be bigger than before.

Thus relaxed, $\mathbb{RelEtest}$ is more likely than $\mathbb{AbsEtest}$ to be effective during early iterations. Later, after $\sum ß$ has grown and $\tilde{n}^2$ has dwindled below $\sum ß$, deflation is more likely to occur after $\mathbb{AbsEtest}$. Alas, "more likely" might not be very likely. The $\mathbb{E}$tests are most effective in the more common cases, when first $q_n$ and then $\bar{e}_{n-1}$ become negligible, but less effective in cases of Disordered Data, when ß becomes negligible long before any $e_j$ does.

Disordered Data can require intolerably too many dqds(0) iterations before any criterion above pemits deflation. Criteria more relaxed than those have long been sought among quadratic error-bounds like the ones exhibited in §8 above. For examples see p. 881's Theorem 5 in Demmel & Kahan [1990] and p. 216 of Fernando & Parlett [1994]. That quest has been futile so far.

All deflation-permitting criteria based upon such quadratic bounds have had to await the opening of an obviously adequate gap between the least singular value of a leading principal submatrix of the bidiagonal S, and the largest of the complementary trailing principal submatrix of S. An adequate gap can be unobvious because the dqds process stores only the squares of the elements of S in arrays $\{q_j\}$ and $\{e_j\}$. Their square roots would be needed for the $\mathbb{RDE}$ & $\mathbb{REF}$ formulas tabulated in §8; for $\mathbb{RBW}$ & $\mathbb{RMB}$ applied to T := S·S' or to S'·S the formulas would need square roots $\sqrt{(q_j\cdot e_{j-1})}$ or $\sqrt{(q_j\cdot e_j)}$. These square roots are unavoidable because there are examples S and T whose diagonal elements differ by not much less than their off-diagonal elements from extreme singular values and eigenvalues. But the inner loop of dqds cannot afford augmentation by a square root for two reasons: First, a square root on most computers nowadays costs at least as much as a few divisions; second, during early dqds iterations when deflations are needed most, they are so rarely permitted by quadratic tests that almost none get rewarded.

Consequently quadratic bounds can permit deflations during the dqds process only outside its inner loop if at all. At which sites might such a deflation be permitted, albeit rarely?

The least unlikely site is at $e_{n-1}$, followed by $e_{n-2}$. That site must be situated at an obviously adequate gap between the least of the larger singular values above the site and the larger of the lesser singular value(s) below. The gap is obvious when $\sqrt{(\min q_j)} - \sqrt{(\max e_j)}$ above that site exceeds $\sqrt{(\max q_j)} + \sqrt{(\max e_j)}$ below it; their excess is a lower bound for the spectral gap. It is adequate when it is big enough that $\mathbb{RDE}$ & $\mathbb{REF}$ predict tolerable perturbations. An obviously adequate gap need not appear until after too many dqds iterations and deflations permitted by non-quadratic criteria, especially if the least several singular values of S are clustered tightly. This happens when $S_0$ comprises a diagonal sum of blocks roughly resembling example S in §4 separated by relatively small off-diagonal elements $\sqrt{\bar{e}_j}$ too big to annihilate.

Whether the costs of tests for gaps needed by quadratic error-bounds are likely to be repaid by substantially earlier deflations remains to be seen.

**Ming Gu's New d-Deflation Criterion:**
Theorem 2 of §6.1, pp. 204-5 in Fernando & Parlett [1994], shows for each $d_j$ of dqds(0) that $1/d_j$ is diagonal element #j of $(S \cdot S')^{-1}$, and consequently the least squared singular value of S lies between dmin and dmin/n. And annihilating $d_{jmin}$ changes no singular value of S by more than $\sqrt{dmin}$. Ming Gu noticed first that a tiny $d_{jmin} > 0$ need not propagate to a similarly tiny $\bar{q}_n$ followed by an $e_{n-1}$ tiny enough for deflation. However an annihilated $d_{jmin}$ always propagates its zero to $\bar{q}_n$ whereupon, at the cost of at most one more dqds(0), deflation at a subsequent $e_{n-1} = 0$ always follows.

When and why can this kind of deflation be expected to occur?

In the absence of deflations, as $\sum ß$ increases monotonically to a limit equal to the least squared singular value $\sigma_n^2$ of $S_0$, the sequence of successful shifts ß tends to 0 though not always monotonically. With a good shift-selection strategy (a long story for another day), ß soon gets so tiny that it might as well be 0; and by then the least singular value of S must be very tiny too but unobviously so because of Disordered Data. Since no deflation has occurred yet, each subsequent dqds(0) includes a test for the condition

$$d_j < \tau^2 \cdot (\text{if } m^2 > \sum ß \text{ then } 4\sum ß \text{ else } m^2 \cdot (1 + \sum ß/m^2)^2)/n \qquad \text{AbsDtest}$$

just before the Etests above to decide whether $d_j$ or $e_j$ is eligible for annihilation. After the annihilation of $d_j$ occurs it is followed immediately by this upward movement of data:

for i = j to n–1 do { $\bar{q}_i := e_i$ ; $\bar{e}_i := q_{i+1}$ } ; $\bar{q}_n := 0$ .

This i-loop saves time that would otherwise be wasted on n–j–1 divisions and other arithmetic in the rest of this dqds(0) iteration. After one more dqds(0) iteration (or some other scheme) applied to arrays $\{\bar{q}_j\}$ and $\{\bar{e}_j\}$ spreads the zero at $\bar{q}_n$ to $e_{n-1}$, deflation occurs there.

Sheng-Guo Li has programmed and tested AbsDtest or something like it added to a LAPACK version of the dqds process. His results show no degradation of accuracy nor speed, but show speed-ups by factors as big as 8 for examples S on which the former LAPACK program spent extraordinarily long times. An account by S-G. Li, M. Gu and B.N. Parlett [2012] has been submitted for publication.

## §11: Conclusion

Like criteria for terminating an iteration, criteria for deflation have to be chosen by the error-analyst to avoid excessive computation without incurring excessive inaccuracy. Deflation may be permitted by more than one criterion at each of very many sites; the opportunities are too numerous for all criteria to be tested at all sites. Instead an economical subset must be found.

The quest continues.

## §12: Citations

J. Demmel, B. Diament & G. Malajovich [2001] "On the Complexity of Computing Error Bounds" pp. 101-125 in ***FOUNDATIONS OF COMPUTATIONAL MATH.*** **1**. Somewhat speculative.

J.W. Demmel & W. Kahan [1990] "Accurate Singular Values of Bidiagonal Matrices" pp. 873-912 in ***SIAM J. Sci. Stat. Comput***. **11** #3.

K.V. Fernando & B.N. Parlett [1994] "Accurate singular values and differential qd algorithms" pp. 191-229 in ***Numerische Mathematik*** **67**.

N.J. Higham [1987] "A Survey of Condition Number Estimation for Triangular Matrices" pp. 573-596 in *SIAM REVIEW* **29** #4. This actually surveys (over)estimators, of norms of inverses of arbitrary triangular matrices, that cost much less to compute than the inverses do.

Leslie Hogben [2007] ed. ***Handbook of Linear Algebra*** 1504 pp., Chapman & Hall/CRC; a huge encyclopedic survey of facts citing a vast literature for their proofs.

C.R. Johnson [1989] "A Gersgorin-type lower bound for the smallest singular value" pp. 1-7 in ***Linear Algebra & Its Applications*** **112**: $1/\|C^{-1}\| \geq \max\{\, 0,\ \min_j \{\, |c_{jj}| - \sum_{k\neq j} (|c_{kj}| + |c_{jk}|)/2 \,\} \,\}$ .

W. Kahan [2012'] "A Tutorial Overview of Vector and Matrix Norms" posted on my web page at www.eecs.berkeley.edu/~wkahan/MathH110/NormOvrv.pdf .

Chi-Kwong Li & Roy Mathias (1999) "The Lidskii-Mirsky-Wielandt Theorem — additive and multiplicative versions" pp. 377-413 in ***Numerische Mathematik*** **81**, surveys unitarily invariant matrix norms' relations with perturbed Hermitian matrices' spectra, with elegant proofs.

Chi-Kwong Li & Ren-Cang Li [2005] "A note on eigenvalues of perturbed Hermitian matrices", pp. 183-190 in ***Linear Algebra and its Applications*** **395**, exploits spectral gaps optimally.

Ren-Cang Li [1994] "On Deflating Bidiagonal Matrices" unpublished note, Mathematics Dept., Univ. of Calif. @ Berkeley, CA 94720. The formulation here was derived as a special case of more general and much more complicated relationships published later in five papers …

R-C. Li [1997] "Relative Perturbation Theory: (III) More Bounds On Eigenvalue Variation" pp. 337—345 in ***Linear Algebra and its Applications*** **266**.

R-C. Li [1998] "Relative Perturbation Theory: (I) Eigenvalue and Singular Value Variations" pp. 956—982 in ***SIAM J. Matrix Anal. Appl.*** **19**.

R-C. Li [1999] "Relative Perturbation Theory: (II) Eigenspace And Singular Space Variations" pp. 471—492 in ***SIAM J. Matrix Anal. Appl.*** **20**.

R-C. Li [2000] "Relative Perturbation Theory: (IV) sin2θ Theorems" pp. 45—60 in ***Linear Algebra and its Applications*** **311**.

R-C. Li [2000] "A Bound On The Solution To A Structured Sylvester Equation With An Application To Relative Perturbation Theory" pp. 471—492 in ***SIAM J. Matrix Anal. Appl***. **21**

S-G. Li, M. Gu & B.N. Parlett [2012] "A Modified DQDS Algorithm" submitted to ***SIAM J. Sci. Comp***,; currently posted at http://arxiv.org/abs/1209.5462 .

B.N. Parlett [1998] ***The Symmetric Eigenvalue Problem*** 426 pp., SIAM, Philadelphia

B.N. Parlett & O.A. Marques [2000] "An Implementation of the dqds Algorithm (Positive Case)" pp. 217-259 in ***Linear Algebra and its Applications*** **309**.

.....