University of Maryland College Park

Institute for Advanced Computer Studies TR–2004–17
Department of Computer Science TR–4572

# Error Analysis of the Quasi-Gram–Schmidt Algorithm*

G. W. Stewart†

March 2004

ABSTRACT

Let the $n \times p$ $(n \geq p)$ matrix $X$ have the QR factorization $X = QR$, where $R$ is an upper triangular matrix of order $p$ and $Q$ is orthonormal. This widely used decomposition has the drawback that $Q$ is not generally sparse even when $X$ is. One cure is to discard $Q$ retaining only $X$ and $R$. Products like $a = Q^{\mathrm{T}}y = R^{-\mathrm{T}}X^{\mathrm{T}}y$ can then be formed by computing $b = X^{\mathrm{T}}y$ and solving the system $R^{\mathrm{T}}a = b$. This approach can be used to modify the Gram–Schmidt algorithm for computing $Q$ and $R$ to compute $R$ without forming $Q$ or altering $X$. Unfortunately, this quasi-Gram–Schmidt algorithm can produce inaccurate results. In this paper it is shown that with reorthogonalization the inaccuracies are bounded under certain natural conditions.

*This report is available by anonymous ftp from `thales.cs.umd.edu` in the directory `pub/reports` or on the web at `http://www.cs.umd.edu/~stewart/`.

†Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (`stewart@cs.umd.edu`). This work was supported in part by the National Science Foundation under grant CCR0204084.

# Error Analysis of the Quasi-Gram–Schmidt Algorithm

G. W. Stewart

ABSTRACT

Let the $n\times p$ ($n \geq p$) matrix $X$ have the QR factorization $X = QR$, where $R$ is an upper triangular matrix of order $p$ and $Q$ is orthonormal. This widely used decomposition has the drawback that $Q$ is not generally sparse even when $X$ is. One cure is to discard $Q$ retaining only $X$ and $R$. Products like $a = Q^{\mathrm{T}}y = R^{-\mathrm{T}}X^{\mathrm{T}}y$ can then be formed by computing $b = X^{\mathrm{T}}y$ and solving the system $R^{\mathrm{T}}a = b$. This approach can be used to modify the Gram–Schmidt algorithm for computing $Q$ and $R$ to compute $R$ without forming $Q$ or altering $X$. Unfortunately, this quasi-Gram–Schmidt algorithm can produce inaccurate results. In this paper it is shown that with reorthogonalization the inaccuracies are bounded under certain natural conditions.

## 1. Introduction

This paper is concerned with the error analysis of an orthogonalization technique, called the quasi-Gram–Schmidt method — or for short, the QGS method. In a previous paper [5] the author has shown that this method can be combined with column pivoting to yield low-rank approximations to sparse matrices.

To set the background let $X$ be an $n\times p$ matrix with $n \geq p$. Then $X$ has the QR factorization

$$X = QR,$$

where $Q^{\mathrm{T}}Q = I$ and $R$ is upper triangular with positive diagonal elements. If $X$ has full column rank — as we will assume from now on — then this QR factorization is unique.

An important way of computing a QR factorization is the Gram–Schmidt algorithm with reorthogonalization. The heart of the method is an algorithm for updating a QR factorization. Suppose we know the QR factorization of $X$ and wish to compute the QR factorization of $(X\ \ x)$, where $x$ is independent of the columns of $X$. This can be done as follows.

$$\begin{array}{ll} 1. & r = Q^{\mathrm{T}}x \\ 2. & u = x - Qr \\ 3. & \rho = \|u\| \\ 4. & q = u/\rho \end{array} \qquad (1.1)$$

Here the norm $\|\cdot\|$ is the vector 2-norm. We call this process a Gram–Schmidt step. It is easy to show that if $q$, $r$, and $\rho$ are computed exactly then

$$(X\ x) = (Q\ q)\begin{pmatrix} R & r \\ 0 & \rho \end{pmatrix}$$

is the QR factorization of $(X\ x)$. The Gram–Schmidt algorithm consists of applying Gram–Schmidt steps successively to the columns of a matrix to obtain its QR factorization.

When the Gram–Schmidt step is executed in finite-precision floating-point arithmetic, the vector $q$ may not be orthogonal to the columns of $Q$. Specifically, when there is cancellation in statement 2 of (1.1), the trailing digits in the computed value of $u$ will be inaccurate, and $u$ will deviate from orthogonalization in proportion to the degree of cancellation.

The cure for this problem is to repeat the orthogonalization as follows.

$$\begin{array}{ll}
1. & r_1 = Q^{\mathrm{T}}x \\
2. & u_1 = x - Qr_1 \\
3. & r_2 = Q^{\mathrm{T}}u_1 \\
4. & u_2 = u_1 - Qr_2 \\
5. & r = r_1 + r_2 \\
6. & \rho = \|u_2\| \\
7. & q = u_2/\rho
\end{array} \tag{1.2}$$

Typically this algorithm exhibits the following behavior. The norm of $u_2$ is less than the norm of $x$ by a factor proportional to the degree of dependence of $x$ on $X$. The norm of $u_2$, on the other hand, is not much smaller than the norm of $u$, which guarantees the orthogonality of $q$ to the column space of $X$. For more on the properties of Gram–Schmidt with reorthogonalization, see [2, Section 2.2.4] or [4, Section 3.1.4]).[1]

A widely used variant of the QR decomposition is the pivoted QR-decomposition in which

$$XP = QR,$$

where $P$ is a permutation matrix the reorders the columns of $X$. The standard algorithm for computing this decomposition [4, Section 5.2.1] tends to put strongly independent columns at the beginning of $XP$ and is useful in generating low-rank approximations to $X$. In [5] the author has shown how to adapt the Gram–Schmidt algorithm to compute this decomposition. In this paper we will treat only the unpivoted algorithm, since the underlying Gram–Schmidt step is the same in both algorithms.

[1]Atypically, $u_2$ can fail to be orthogonal to the column space of $x$, in which case the reorthogonalization process must be repeated.

When $X$ is large and sparse, the Gram–Schmidt algorithm is unsatisfactory. The reason is that as the algorithm progresses, the new columns of $Q$ become less and less sparse. A cure for this problem [5] is to recognize that, since $Q = XR^{-1}$, it is only necessary to retain the matrix $R$. Products, such as $r = Q^{\mathrm{T}}x$ can be computed by first computing $a = X^{\mathrm{T}}x$ and then solving the system $R^{\mathrm{T}}r = a$. These considerations lead to the following quasi-Gram–Schmidt (QGS) step with reorthogonalization

$$
\begin{array}{rl}
1. & a_1 = X^{\mathrm{T}}x \\
2. & \text{Solve the system } R^{\mathrm{T}}r_1 = a_1 \\
3. & \text{Solve the system } Rb_1 = r_1 \\
4. & u_1 = x - Xb_1 \\
5. & a_2 = X^{\mathrm{T}}u_1 \\
6. & \text{Solve the system } R^{\mathrm{T}}r_2 = a_2 \\
7. & \text{Solve the system } Rb_2 = r_2 \\
8. & u_2 = u_1 - Xb_2 \\
9. & r = r_1 + r_2 \\
10. & \rho = \|u_2\|
\end{array}
\tag{1.3}
$$

It is important to note that we do not compute $q = u_2/\rho$. Instead we *define* $q$ to be the last column of

$$
(X\ \ x)\begin{pmatrix} R & r \\ 0 & \rho \end{pmatrix}^{-1} = (X\ \ x)\begin{pmatrix} R^{-1} & -\rho^{-1}R^{-1}r \\ 0 & \rho^{-1} \end{pmatrix}; \tag{1.4}
$$

that is,

$$
q = \rho^{-1}x - \rho^{-1}XR^{-1}r = \rho^{-1}(x - Qr). \tag{1.5}
$$

There is good reason to be wary of this algorithm. If $R$ is ill-conditioned, the systems in statements 2, 3, 6, and 7 may be solved inaccurately, and these inaccuracies could compromise the orthogonality of $Q = XR^{-1}$. In fact, something like this must happen. For suppose we are given the correctly rounded $R$ from the QR factorization of $X$. Then $R = R + E$, where

$$
\|E\| \le \gamma\|R\|\epsilon_{\mathrm{M}}. \tag{1.6}
$$

Here $\epsilon_{\mathrm{M}}$ is the rounding unit (about $2.2{\cdot}10^{-16}$ in IEEE double-precision floating-point arithmetic), $\|\cdot\|$ is the spectral norm, and $\gamma$ is a constant depending on the dimensions of $X$. Suppose we compute an approximation $\tilde{Q}$ to $Q$ using $R + E$ but with no further error. Noting that up to terms of order $\epsilon_{\mathrm{M}}^2$

$$
(R + E)^{-1} \cong R^{-1} - R^{-1}ER^{-1},
$$

we find that

$$
\tilde{Q} = X(R + E)^{-1} \cong Q - QER^{-1}.
$$

Hence

$$
\tilde{Q}^{\mathrm{T}}\tilde{Q} \cong Q^{\mathrm{T}}Q - Q^{\mathrm{T}}QER^{-1} - R^{-\mathrm{T}}EQ^{\mathrm{T}}Q.
$$

Hence if we define

$$W = Q^{\mathrm{T}}Q - I$$

and take

$$\omega = \|W\|$$

as a measure the of nonorthonormality of $Q$, then

$$\tilde{\omega} \lesssim \omega + 2\|Q^{\mathrm{T}}Q\|\|R^{-1}\|\|E\| \leq \omega + 2\gamma\kappa(R)\epsilon_{\mathrm{M}}, \tag{1.7}$$

where $\kappa(R) = \|R\|\|R^{-1}\|$. Thus, however small the value of $\omega$, the rounding of $R$ is likely to increase it to something on the order of $\gamma\kappa(R)\epsilon_{\mathrm{M}}$.

The above analysis shows that there is a lower limit on the orthogonality the QGS algorithm can be expected to attain. In [5], the author gave numerical examples that suggested that the algorithm would come close to this limiting accuracy, provided a reorthogonalization step is included [as it is in (1.3)]. The purpose of this note is show by an informal rounding-error analysis that this is indeed the case.

This paper is organized as follows. In the next section we discuss the effects of scaling on the condition of $R$ and on sizes of the quantities in the QGS step. In Section 3 we give some numerical examples of the behavior of the QGS method to provide the reader with a view of the goal of our informal analysis. In Section 4 we show how the CGS step behaves in the absence of rounding error. This analysis is related to material in a paper by Hoffmann [3]. The analysis proper is given in Section 5 followed by a discussion and an appendix.

As above $\epsilon_{\mathrm{M}}$ will denote the rounding unit and $\|\cdot\|$ the spectral norm. We will ignore second and higher order terms in $\epsilon_{\mathrm{M}}$ and indicate their omission by the substitution of '$\cong$' for '$=$' and '$\lesssim$' for '$\leq$'. We will use a generic constant $\langle\gamma\rangle$ to make minor adjustments in our bounds or to restore '$=$' or '$\leq$'. Note that two appearances of $\langle\gamma\rangle$ in an expression need not represent the same number.

We will denote the column space of $X$ by $\mathcal{R}(X)$ and its orthogonal complement by $\mathcal{R}(X)^{\perp}$. For any vector $v$, we will write

$$v = v_X + v_{\perp},$$

where $v_X$ is the orthogonal projection of $v$ onto $\mathcal{R}(X)$ and $v_{\perp}$ is the orthogonal projection of $v$ onto $\mathcal{R}(X)^{\perp}$. We will measure the orthogonality of $v$ to $\mathcal{R}(X)$ by the ratio

$$\tau(v) = \frac{\|v_X\|}{\|v_{\perp}\|}. \tag{1.8}$$

Note that $\tau(v)$ is the the tangent of the angle between $v_{\perp}$ and $\mathcal{R}(X)^{\perp}$. Thus a small value indicates that $v$ has only a small component along $\mathcal{R}(X)$. Conversely, a large value indicates that $v$ has only a small component in $\mathcal{R}(X)^{\perp}$.

The QGS algorithm without reorthogonalization is related to the method of seminormal equations for solving the least squares problem

$$\|x - Xr\|_2^2 = \min .$$

With reorthogonalization it is related to Björck's method [1] of corrected seminormal equations. However, both of these methods assume that the matrix $R$ is the R-factor from a nearby matrix $X + E$, where $\|E\|$ is of the order $\|X\|\epsilon_{\mathrm{M}}$. Björck gives an extensive error analysis; however, it does not seem adaptable to our problem of tracking the orthogonality of the implicit Q-factor as the QGS method progresses.

## 2. Scaling and magnitude

Before beginning the analysis proper, we must dispose of some questions of scaling and magnitude. Let $D = \mathrm{diag}(d_1, \ldots, d_p)$ be any diagonal matrix with positive diagonal elements. Then

$$Q = XR^{-1} = (XD)(RD)^{-1}.$$

Thus, without affecting $Q$ we can scale the columns of $X$ and $R$ by arbitrary factors. In particular, by taking $d_1, \ldots, d_{p-1} = 1$ and $d_p$ sufficiently small, we can make $\kappa(R)$ as large as we want. This means that the bound (1.7) on $\tilde{\omega}$ can be made arbitrarily large. Thus for this bound to be informative, we must find a scaling that minimizes $\kappa(R)$. Although this problem is unsolved, a remarkable theorem of van der Sluice [7] states that the conditions number of $R$ is approximately minimized when the columns of $R$ all have the same norm.

It is also important to have a sense of the sizes of the quantities involved in our bounds. Since $\|Q\|^2 = \|Q^{\mathrm{T}}Q\| = \|I + W\|$, we have

$$\sqrt{1-\omega} \le \|Q\| \le \sqrt{1+\omega}$$

Moreover, $X^{\mathrm{T}}X = R^{\mathrm{T}}(Q^{\mathrm{T}}Q)R = R^{\mathrm{T}}(I + W)R$. Hence,

$$\sqrt{1-\omega}\,\|R\| \le \|X\| \le \sqrt{1+\omega}\,\|R\|.$$

Thus if $\omega$ is less than, say, 0.1, $\|X\|$, $\|Q\|$, and $\|R\|$ are approximately equal. Similar results hold for the individual columns of $Q$, $X$, and $R$.

In what follows we will assume that

$$\|X\| = 1 \quad \text{and} \quad \|x\| = 1.$$

This means that in expressions like $\|R\|\epsilon_{\mathrm{M}}$ we can absorb the $\|R\|$ in our generic constant $\langle\gamma\rangle$, with a resulting simplification in the bounds. We will also assume that the columns $x_j$ of $X$ all have the same norm. This means that the columns of $R$ all have approximately the same norm, which, as we have seen above, is required to make our results meaningful.

### 3. Three examples

To help the reader follow the analysis of the QGS method we present three numerical examples that illustrate the behavior of algorithm (1.3). In all three examples, four steps algorithm (1.3) are applied to a 50×5 matrix $X = (x_1 \ \cdots \ x_5)$ to orthogonalize its columns. Thus after the $k$th step, $QR$ is the quasi-QR factorization of $X = (x_1 \ \cdots \ x_{k+1})$. The matrices $X$ were generated in the form

$$X = USV^{\mathrm{T}}$$

where $U$ is an 50×5 random orthonormal matrix, $V$ is a 5×5 random orthogonal matrix and $S$ is a diagonal matrix containing the singular values of $X$. All computations were performed in IEEE standard arithmetic with a rounding unit of about $2.2{\cdot}10^{-16}$.

In the first example, there is a sharp relative gap of about $10^{-6}$ between the second and third singular values. The last two columns in the table contain the values of $\omega$ and $\hat{\alpha} = \|R^{-1}\|\epsilon_{\mathrm{M}}$ for the $Q$ and $R$ after the QGS step. As we have conjectured, they track each other nicely.

Moving across the first row, we find that the second column of $x$ has a reasonable component along the orthogonal complement of the first column as measured by $\tau_x$. One QGS step produces that component to almost full accuracy and the second orthogonalization is redundant. The new 50×2 matrix $Q$ is almost fully orthonormal, and the corresponding value of $\hat{\alpha}$ is near the rounding unit.

The second row tells a more interesting tale. The first two columns of $X$ approximately span the subspace spanned by the singular vectors corresponding to the two large singular values. This means that when $x_3$ is orthogonalized in must approximately lie in the space spanned by the remaining singular vectors. Since the corresponding singular values are small, $x_{\perp}^{(3)}$ must be small and hence $\tau_{x_3}$ is large. In fact, it is so large that it takes two QGS steps to make the result fully orthogonal. It is significant that $\tau_1 \cong \hat{\alpha}\tau_x$. It is also significant that although $b_2$ is fully orthogonal, some orthogonality is lost in the passage to $q$. This loss is proportional to the increase in $\hat{\alpha}$.

The third row illustrates yet another point. As in the first column we have $\tau_1 \cong \hat{\alpha}\tau_x$. But in the reorthogonalization step the reduction stagnates: $\tau_2$ is approximately equal to $\hat{\alpha}$. (The reduction also stagnates in the second row. But since $\hat{\alpha}$ is near the rounding unit, it cannot be seen whether the stagnation is due to the size of $\hat{\alpha}$ or to the fact that we cannot hope to orthogonalize beyond the rounding unit.)

The second example shows the same phenomenon in a gentler setting. The singular grade smoothly without gaps from 1 to about $10^{-7}$. The values of $\tau_x$ reflect this grading. In all cases $\tau_1 \cong \hat{\alpha}\tau_x$, but $\tau_2$ is never much less than $\hat{\alpha}$. At every step, some orthogonality is lost in the passage from $b_2$ to $q$ and the loss is approximately proportional to the increase in $\hat{\alpha}$.

Example 1: Singular values

| | | | | |
|---|---|---|---|---|
| 1.0e+00 | 7.2e−01 | 3.6e−07 | 1.0e−07 | 6.1e−08 |

| $\hat{\alpha}$ | $\tau_x$ | $\tau_1$ | $\tau_2$ | $\tau_q$ | $\omega$ | $\hat{\alpha}_{\mathrm{new}}$ |
|---|---|---|---|---|---|---|
| 2.2e−16 | 6.7e+00 | 6.4e−16 | 2.1e−17 | 5.7e−16 | 8.6e−16 | 3.8e−15 |
| 3.8e−15 | 2.1e+05 | 1.2e−10 | 4.8e−16 | 3.2e−11 | 5.3e−11 | 7.5e−10 |
| 7.5e−10 | 3.8e+06 | 6.4e−06 | 5.3e−11 | 4.4e−10 | 4.2e−10 | 2.4e−09 |
| 2.4e−09 | 6.0e+06 | 2.4e−03 | 5.7e−11 | 4.2e−10 | 6.2e−10 | 3.7e−09 |

Example 2: Singular values

| | | | | |
|---|---|---|---|---|
| 1.0e+00 | 1.4e−01 | 1.6e−03 | 4.6e−06 | 1.8e−07 |

| $\hat{\alpha}$ | $\tau_x$ | $\tau_1$ | $\tau_2$ | $\tau_q$ | $\omega$ | $\hat{\alpha}_{\mathrm{new}}$ |
|---|---|---|---|---|---|---|
| 2.2e−16 | 6.8e+00 | 8.0e−16 | 5.3e−17 | 1.2e−16 | 4.6e−16 | 4.8e−15 |
| 4.8e−15 | 1.1e+04 | 3.0e−12 | 6.8e−17 | 3.4e−13 | 4.7e−13 | 7.6e−12 |
| 7.6e−12 | 6.9e+02 | 3.6e−09 | 1.9e−13 | 1.2e−11 | 1.5e−11 | 5.0e−11 |
| 5.0e−11 | 2.1e+06 | 1.3e−05 | 1.4e−12 | 4.4e−11 | 1.3e−10 | 1.3e−09 |

Example 3: Singular values

| | | | | |
|---|---|---|---|---|
| 1.0e+00 | 4.6e−04 | 2.3e−07 | 1.2e−11 | 7.3e−16 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2.2e−16 | 2.9e+03 | 1.8e−13 | 0.0e+00 | 1.4e−13 | 1.7e−13 | 2.0e−12 |
| 2.0e−12 | 4.3e+05 | 1.7e−07 | 2.7e−13 | 9.1e−11 | 1.3e−10 | 1.1e−09 |
| 1.1e−09 | 4.3e+10 | 9.0e+00 | 2.2e−09 | 2.6e−06 | 3.6e−06 | 2.6e−05 |
| 2.6e−05 | 3.8e+14 | 2.2e+05 | 1.1e+00 | 2.7e+03 | 1.0e+00 | 7.2e−05 |

$$\hat{\alpha} = \|R^{-1}\|\epsilon_{\mathrm{M}}, \quad \tau_x = \tau(x), \quad \tau_1 = \tau(u_1), \quad \tau_2 = \tau(b_2), \quad \tau_q = \tau(q), \quad \omega = \|I - Q^{\mathrm{T}}Q\|$$

Figure 3.1: Three examples

The third example shows that QGS method can fail. The singular values are steeply graded from 1 to about $10^{-15}$. The first three rows behave in the manner we have come to expect. But in the fourth row the two orthogonalizations are not sufficient to make $\tau_1$ of the same order of magnitude of as $\hat{\alpha}$, and orthogonality in $Q$ is lost. We could overcome this problem by including additional reorthogonalizations. But as we shall see, there are good reasons for not doing so.

To summarize, we must explain three things about the algorithm.

1. Why do the orthogonalization steps enhance orthogonality by a factor of $\hat{\alpha}$?

2. Why is $\hat{\alpha}$ a lower bound on the attainable orthogonality?
3. Why is orthogonality lost in passing from $b_2$ to $q$ and why is it proportional to the increase in $\hat{\alpha}$?

## 4. The exact QGS step

The idea for our analysis is to show first that, absent rounding errors, repeated QGS steps produce increasingly orthogonal vectors $q$. Then, in the next section, we show that rounding errors perturb that $q$ by quantities of order $\|R^{-1}\|\epsilon_{\mathrm{M}}$. There we also analyze the formation of $q$ from $b_2$.

Without rounding error, the QGS step without reorthogonalization can be written in the form

$$y = (I - QQ^{\mathrm{T}})x \equiv Px.$$

Here we skip the last normalization step. As usual, decompose

$$x = x_X + x_\perp \quad \text{and} \quad y = y_X + y_\perp.$$

The problem then is to find expressions for $y_X$ and $y_\perp$.

The vector $y_\perp$ is easy to find. Let $U_\perp$ be any orthonormal basis for $\mathcal{R}(X)^\perp$. Then

$$\begin{aligned} y_\perp &= U_\perp U_\perp^{\mathrm{T}} y \\ &= U_\perp U_\perp^{\mathrm{T}}(I - QQ^{\mathrm{T}})x \\ &= U_\perp U_\perp^{\mathrm{T}} x \qquad\qquad \text{(since } U_\perp^{\mathrm{T}} X = 0) \\ &= x_\perp. \end{aligned}$$

In other words, multiplication by $P$ does not change the component of $x$ along $\mathcal{R}(X)^\perp$.

To determine $y_X$ we must construct a specific orthonormal basis for $\mathcal{R}(X)$. As above let $W = Q^{\mathrm{T}}Q - I$ and assume that $\omega = \|W\| < 1$. Then $I + W$ is positive definite and has a positive definite square root $(I+W)^{\frac{1}{2}}$. It follows that

$$U_X = Q(I+W)^{-\frac{1}{2}}$$

is orthogonal. Since the column space of $Q$ and $X$ are the same, $U_X$ forms an orthonormal basis for the column space of $X$.

Now

$$\begin{aligned} y_X &= U_X U_X^{\mathrm{T}}(I - QQ^{\mathrm{T}})x \\ &= U_X U_X^{\mathrm{T}}[I - U_X(I+W)^{\frac{1}{2}}(I+W)^{\frac{1}{2}}U_X^{\mathrm{T}}]x \quad \text{(since } (I+W)^{\frac{1}{2}} \text{ is symmetric)} \\ &= U_X[I - (I+W)]U_X^{\mathrm{T}}x \\ &= U_X W U_X^{\mathrm{T}} x \\ &= U_X W U_X^{\mathrm{T}} x_X \qquad\qquad \text{(since } U_X x_\perp = 0). \end{aligned}$$

It follows that

$$\|y_X\| \leq \omega \|x_X\|.$$

In other words, multiplication by $I - QQ^{\mathrm{T}}$ reduces the component along $\mathcal{R}(X)$ by a factor of at least $\omega$.

The consequence of all this is that the repeated application of the QGS step produces a sequence of vectors that converges to $y_\perp$ — or with normalization to $y_\perp/\|y_\perp\|$. The rate of convergence is that of the approach of $\omega^k$ to zero with increasing $k$. It remains to determine to what extent rounding error limits this convergence.

## 5. The effects of rounding error

We now aassume that the QGS step (1.3) without reorthogonalization is computed in floating-point arithmetic with rounding unit $\epsilon_{\mathrm{M}}$. We assume that the arithmetic is standard in the sense that

$$\mathrm{fl}(a \circ b) = (a \circ b)(1 + \epsilon), \qquad |\epsilon| \leq |\epsilon_{\mathrm{M}}|, \quad \circ = +, -, \times, \div,$$

where $\mathrm{fl}(a \circ b)$ denotes the computed value of $a \circ b$.

Using standard techniques of rounding error analysis, we have the following relations among the computed values:

$$\begin{array}{ll} a = (X + E)^{\mathrm{T}} x, & \|E\| \leq \langle\gamma\rangle \|X\| \epsilon_{\mathrm{M}}, \\ r = (R + F)^{-\mathrm{T}} a, & \|F\| \leq \langle\gamma\rangle \|R\| \epsilon_{\mathrm{M}}, \\ b = (R + G)^{-1} r, & \|G\| \leq \langle\gamma\rangle \|R\| \epsilon_{\mathrm{M}}, \\ \tilde{u} = x - (X + H)b + h, & \|H\| \leq \langle\gamma\rangle \|X\| \epsilon_{\mathrm{M}}, \\ & \|h\| \leq \langle\gamma\rangle \|x\| \epsilon_{\mathrm{M}}. \end{array}$$

Here we have assumed that $\|\tilde{u}\| \leq \|x\|$, which is reasonable in this context. Since by our scaling we have insured that the norms of $X$, $R$, and $x$ are near one, we we may absorb their norms into $\langle\gamma\rangle$ and use the simpler bounds

$$\|E\|, \|F\|, \|G\|, \|H\|, \|h\| \leq \langle\gamma\rangle \epsilon_{\mathrm{M}}.$$

Let us denote the composite mapping from $x$ to $\tilde{u}$ by $\tilde{P}$. Then up to second order terms in $\epsilon_{\mathrm{M}}$, we have

$$\begin{aligned} \tilde{P} \cong\ & XR^{-1}R^{-\mathrm{T}}X^{\mathrm{T}} \\ & + HR^{-1}R^{-\mathrm{T}}X^{\mathrm{T}} \\ & - XR^{-1}GR^{-1}R^{-\mathrm{T}}X^{\mathrm{T}} \\ & - XR^{-1}R^{-\mathrm{T}}FR^{-\mathrm{T}}X^{\mathrm{T}} \\ & + XR^{-1}R^{-\mathrm{T}}E^{\mathrm{T}}. \end{aligned}$$

Since $Q = XR^{-1}$, we may write

$$\tilde{P} \cong P + HR^{-1}Q^{\mathrm{T}} - QGR^{-1}Q^{\mathrm{T}} - QFR^{-\mathrm{T}}Q + QR^{-\mathrm{T}}E^{\mathrm{T}}.$$

It follows that with $u = (I - P)x$ and $\tilde{u} = (I - \tilde{P})x + h$ we have

$$\tilde{u} = u + e, \tag{5.1}$$

where

$$\|e\| \le \langle\gamma\rangle\|R^{-1}\|\|x\|\epsilon_{\mathrm{M}} = \alpha\|x\|, \tag{5.2}$$

where

$$\alpha = \langle\gamma\rangle\|R^{-1}\|\epsilon_{\mathrm{M}}.$$

(Note that $\alpha$ differs from $\hat{\alpha}$ in Section 3 only by the generic factor $\langle\gamma\rangle$.) Thus the effect of rounding error is to replace $u = (I - QQ^{\mathrm{T}})x$ by $u + e$, $\|e\|$ is bounded by $\alpha\|x\|$.

We have seen in the last section that the iterated quasi-projection $(I-P)^k x$ converges to a vector orthogonal to $\mathcal{R}(X)$. We shall now show that the presence of the error $e$ causes the convergence of the vectors

$$x^k = (I - \tilde{P})^k x$$

to stagnate.

We first note that the presence of $e$ in (5.1) limits the size of of the orthogonality measure $\tau(x)$. Specifically, from (5.2) and (1.8) we have that

$$\sqrt{1+\tau(x)^2}\,\alpha \lesssim \frac{\|e\|}{\|x_\perp\|}.$$

If the quantity on the left is greater one, then $\|e\|$ must be larger than $\|x_\perp\|$, and the addition of $e$ in (5.1) may obliterate $x_\perp$, after which there is no way to compute it. Since $\alpha$ will be small, the condition $\sqrt{1+\tau(x)^2}\,\alpha < 1$ is tantamount to the condition

$$\tau(x)\alpha < 1,$$

and this is the form we will use in what follows. It is worth noting that this condition is violated in the fourth row of Example 3 above.

Turning now to the behavior of the $x^k$, write

$$x = x^0 = x_X^0 + x_\perp^0.$$

By the results of the preceding section,

$$x^1 \equiv (I - \tilde{P})x^0 = (I - P)x^1 + e^0 \cong \hat{x}_X^0 + x_\perp^0 + e^0 \equiv x_X^1 + x_\perp^1,$$

where

$$\|\hat{x}_X^1\| \le \omega \|x_X^0\|.$$

Now suppose we have constants $\eta_0$, $\zeta_0$, and $\theta_0$ satisfying

$$\|x_X^0\| \le \eta_0 \quad \text{and} \quad \theta_0 \le \|x_\perp^0\| \le \zeta_0.$$

Then

$$\|x_X^1\| \le \omega\eta_0 + \alpha\|x^0\| \le \omega\eta_0 + \alpha(\|x_X^0\| + \|x_\perp^0\|) \le (\alpha+\omega)\eta_0 + \alpha\zeta_0.$$

Similarly,

$$\|x_\perp^1\| \le \alpha\eta_i + (1+\alpha)\zeta_i$$

and

$$\|x_\perp^1\| \ge -\alpha\eta_0 - \alpha\zeta_0 + \theta_0$$

Thus if we set

$$C = \begin{pmatrix} \alpha+\omega & \alpha & 0 \\ \alpha & 1+\alpha & 0 \\ -\alpha & -\alpha & 1 \end{pmatrix}$$

and define

$$\begin{pmatrix} \eta_k \\ \zeta_k \\ \theta_k \end{pmatrix} = C^k \begin{pmatrix} \eta_0 \\ \zeta_0 \\ \theta_0 \end{pmatrix},$$

then we have

$$\|x_X^k\| \le \eta_k, \quad \text{and} \quad \theta_k \le \|x_\perp^k\| \le \zeta_k.$$

We shall be particularly interested in the evolution of the ratio $\tau_k = \eta_k/\theta_k$, which is a bound on $\tau(x_\perp^k)$. We can trace this if we know the eigensystem of $C$. Specifically, let

$$Z^{-1}CZ = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3) \equiv \Lambda.$$

be an eigendecompostion of $C$. Let $s_k = (\eta_k\ \zeta_k\ \theta_k)^{\mathrm{T}}$. If we set $b = Z^{-1}s_0$ so that $s_0 = Zb$, then

$$s_k = C^k s_0 = C^k Z b = Z\Lambda^k b = b_1\lambda_1^k z_1 + b_2\lambda_2^k z_2 + b_3\lambda_3^k z_3.$$

Thus we obtain an explicit formula for the $s_k$ from which their behavior as $k$ increases can be read off.

Under the assumption that $\alpha$ and $\omega$ are small, we can use perturbation theory to determine approximations to the eigenvalues and eigenvectors of $C$. Specifically (see the appendix for details), up to terms of order $\alpha^2$

$$\Lambda \cong \operatorname{diag}(\alpha+\omega, 1+\alpha, 1) \quad \text{and} \quad Z \cong \begin{pmatrix} 1 & \alpha & 0 \\ -\alpha & 1 & 0 \\ \alpha & -1 & 1 \end{pmatrix}. \tag{5.3}$$

Moreover, up to terms of order $\alpha^2$

$$Z^{-1} \cong \begin{pmatrix} 1 & -\alpha & 0 \\ \alpha & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

It follows that if we take $\zeta_0 = \theta_0$, then

$$b \cong \begin{pmatrix} \eta_0 - \alpha\theta_0 \\ \alpha\eta_0 + \theta_0 \\ 2\theta_0 \end{pmatrix}$$

Moreover the ratio $\tau_k = \eta_k/\theta_k$ is

$$\begin{aligned} \tau_k &\cong \frac{b_1(\alpha+\omega)^k + b_2\alpha(1+\alpha)^k}{b_1\alpha(\alpha+\omega)^k - b_2(1+\alpha)^k + b_3} \\ &= \frac{(\tau_0-\alpha)(\alpha+\omega)^k + (1+\alpha\tau_0)\alpha(1+\alpha)^k}{(\tau_0-\alpha)\alpha(\alpha+\omega)^k - (1+\alpha\tau_0)(1+\alpha)^k + 2}. \end{aligned} \tag{5.4}$$

Let us look at the second expression for $\tau_k$ more carefully under the assumptions that $\alpha = \omega$ and that $\alpha$ is small. Note that this assumption is essentially an induction hypothesis. For it says that the loss of orthogonality in $Q$ is of the same order of magnitude as the condition number of $R$ times $\epsilon_{\mathrm{M}}$—just the proposition we are trying to establish.

Both the numerator and denominator in the second expression in (5.4) have terms that are converging to zero as $(2\alpha)^k$—that is rapidly. They both also have terms that remain essentially constant, since $(1+\alpha)^k$ is near one. The ratios of these latter terms,

$$\frac{1+\alpha\tau_0}{1-\alpha\tau_0}\alpha \tag{5.5}$$

approximate the value of $\tau$ at which the values of $\tau_k$ stagnates. For

$$\alpha\tau_0 < 0.1 \tag{5.6}$$

this value is essentially $\alpha$. For greater values, the limiting value of $\tau_k$ becomes greater, blowing up as $\alpha\tau_0$ approaches one.

The rate of the approach to stagnation is controlled by the the term $(\tau_0 - \alpha)(2\alpha)^k$. Specifically, stagnation occurs when $\tau_0(2\alpha)^k \le \alpha$. For $\tau_0 = 1$, this happens essentially when $k = 1$. However, as $\tau_0$ grows, we will generally need $k = 2$ to achieve stagnation. This suggests that that if (5.6) is satisfied then only one reorthogonalization in the QGS algorithm is needed to to attain the best possible orthogonalization.

| | | | | | | |
|---|---|---|---|---|---|---|
| $\tau_0$ | 1.000e+00 | 1.000e+01 | 1.000e+02 | 1.000e+03 | 1.000e+04 | 1.000e+05 |
| $\tau_1$ | 3.000e−06 | 2.100e−05 | 2.010e−04 | 2.003e−03 | 2.020e−02 | 2.222e−01 |
| $\tau_2$ | 1.000e−06 | 1.000e−06 | 1.001e−06 | 1.006e−06 | 1.061e−06 | 1.667e−06 |
| $\tau_3$ | 1.000e−06 | 1.000e−06 | 1.000e−06 | 1.002e−06 | 1.020e−06 | 1.222e−06 |
| $\tau_{\min}$ | 1.000e−06 | 1.000e−06 | 1.000e−06 | 1.002e−06 | 1.020e−06 | 1.222e−06 |

Figure 5.1: The progress of $\tau$ for $\alpha = \omega = 10^{-6}$

Figure 5.1 shows the values of $\tau_k$ for $\alpha = \omega = 10^{-6}$ and for $\tau_0 = 10^i$ $(i = 0, \ldots, 5)$. The row labeled $\tau_{\min}$ contains the value (5.5) at which the orthogonalization stagnates. It is seen that even when $\alpha\tau = 0.1$, one reorthogonalization is sufficient to put $\tau$ in the ball park, and a second makes it equal to $\tau_{\min}$ to four digits. Since (5.6) is satisfied, the rise of $\tau_{\min}$ as $\tau_0$ increases is gradual and insignificant.

With two steps of orthogonalization, we have computed the vector $u_2$ at step 8 in (1.3), which satisfies $\tau(u_2) \cong \alpha$. However, as we have mentioned, we do not use $q_2 = u_2/\rho$ as our $q$. Instead we define it as the last column of (1.4). Mathematically, this amounts to setting

$$u = x - XR^{-1}r = XR^{-1}(r_1 + r_2). \quad \text{and} \quad q = u/\rho. \tag{5.7}$$

Thus we must evaluate the degree of orthonormality of the matrix $(Q\ q)$. Let

$$(Q\ q)^{\mathrm{T}}(Q\ q) = \begin{pmatrix} W & w \\ w^{\mathrm{T}} & \nu \end{pmatrix}.$$

Since we know $\|W\| = \omega$ we must bound $\|w\|$ and $|1 - \nu|$.

From our rounding error analysis we know that

$$u_2 = x - X[(R + E_1)^{-1}r_1 + (R + E_2)r_2],$$

where $\|E_i\| \leq \langle\gamma\rangle\epsilon_{\mathrm{M}}$. It follows from (5.7) that

$$g = u - u_2 \cong -XR^{-1}E_1R^{-1}r_1 - XR^{-1}E_2R^{-1}r_2 = Q(E_1R^{-1}r_1 + E_2R^{-1}r_2).$$

Remembering that $\|Q\|$, and $\|r_1\|$ are near one and that $\|r_2\| \leq \langle\gamma\rangle\|r_1\|$, we have

$$\|g\| \leq \langle\gamma\rangle\|R^{-1}\|\|r\|\epsilon_{\mathrm{M}}.$$

To bound $\|w\|$, observe that $\tau(u_2) = \|u_X^{(2)}/u_\perp^{(2)}\| \cong \alpha$, which we assume is small. This implies that

$$\|u_X^{(2)}\| \cong \alpha\|u_\perp^{(2)}\| \cong \alpha\|u_2\| = \alpha\rho.$$

On the other hand

$$Q^{\mathrm{T}}u = Q^{\mathrm{T}}(u_2 + g) = Q^{\mathrm{T}}(u_X^{(2)} + u_\perp^{(2)} + g) = Q^{\mathrm{T}}(u_X^{(2)} + g),$$

Hence

$$\|Q^{\mathrm{T}}u\| \leq \langle\gamma\rangle(\alpha\rho + \|R^{-1}\|\|r\|\epsilon_{\mathrm{M}}).$$

and

$$\|w\| = \|Q^{\mathrm{T}}u\|/\rho \leq \langle\gamma\rangle(\alpha + \rho^{-1}\|R^{-1}\|\|r\|\epsilon_{\mathrm{M}}).$$

Since $\alpha = \langle\gamma\rangle\|R^{-1}\|\epsilon_{\mathrm{M}}$,

$$\|w\| \leq \langle\gamma\rangle(\alpha + \rho^{-1}\|R^{-1}\|\|r\|\epsilon_{\mathrm{M}}) = \langle\gamma\rangle\alpha(1 + \rho^{-1}\|r\|). \tag{5.8}$$

To bound $1 - \nu$, we first observe that since $\rho$ is the computed value of $\|u_2\|$

$$\rho^{-1}\|u_2\| = 1 + \langle\gamma\rangle\epsilon_{\mathrm{M}}.$$

$$\begin{aligned}
\nu &= \rho^{-2}(u_2 + g)^{\mathrm{T}}(u_2 + g) \\
&= \rho^{-2}(u_2^{\mathrm{T}}u_2 + 2u^{\mathrm{T}}g + g^{\mathrm{T}}g) \\
&= 1 + \langle\gamma\rangle\epsilon_{\mathrm{M}} + 2\rho^{-2}u_2^{\mathrm{T}}g + \rho^{-2}g^{\mathrm{T}}g \\
&\cong 1 + \langle\gamma\rangle\epsilon_{\mathrm{M}} + 2\rho^{-2}u_2^{\mathrm{T}}g
\end{aligned}$$

Hence

$$|1 - \nu| \leq \langle\gamma\rangle(1 + \rho^{-1}\|R^{-1}\|\|r\|)\epsilon_{\mathrm{M}}. \tag{5.9}$$

To interpret bounds (5.8) and (5.9) let

$$R_{\mathrm{new}} = \begin{pmatrix} R & r \\ 0 & \rho \end{pmatrix}.$$

It then follows from (1.4), that the bounds are themselves bounded by $\langle\gamma\rangle\|R_{\mathrm{new}}^{-1}\|$. Since $\|R_{\mathrm{new}}\| \cong 1$, they are also bounded by $\langle\gamma\rangle\kappa(R_{\mathrm{new}})$. In other words the deterioration in orthogonality in the current $Q$ is bounded by a multiple of the condition number of the current $R$, which is what we set out to establish.

## 6. Discussion

We are now in a position to answer the three questions raised at the end of Section 3.1. The first two questions — why does each orthogonalization decrease $\tau$ by a factor $\alpha$ and why does the decrease stop at $\alpha$ — are essentially answered by our analysis of the second expression in (5.4). We assume that $\alpha = \omega$, but, as we have pointed out, this is essentially an induction hypothesis.

The answer third question — why is there a loss of orthogonality in passing from $u_2$ to $q$ and why is it proportional to the increase in $\alpha$ — is more complicated. In fact, there there can be little loss of orthogonality. If, for example, $x$ is orthogonal to $\mathcal{R}(X)$, $r$ will be small and $\rho$ will be near one, so that the bounds (5.8) and (5.9) will not be much larger than $\alpha$. On the other hand if there is a loss of orthogonality the last column of $R_{\text{new}}^{-1}$ must be bigger than $R^{-1}$, which will cause a proportional increase in $\alpha$.

In the the third example of Section 3.1 we saw that unless $\alpha\tau_x < 1$, the QGS method with only a single reorthogonalization can fail. In our analysis of the process we saw that a slightly stronger condition — $\alpha\tau(x) < 0.1$ should be imposed. Although additional orthogonalizations can revive the process, they are of little avail. For when the condition fails most of the information about $x_\perp$ is obliterated in the first orthogonalization. However, this requirement is not very important in the principal application of the QGS method, which is to produce well-conditioned low-rank approximations to a matrix. One would use column pivoting to bring in linearly independent (read small $\tau$) vectors and stop process before $\alpha/\epsilon_{\mathrm{M}}$ became large.

## 7. Appendix: The eigensystem of C

In this appendix we will determine approximations to the eigenvalues and eigenvectors of the matrix

$$C = \begin{pmatrix} \alpha+\omega & \alpha & 0 \\ \alpha & 1+\alpha & 0 \\ -\alpha & -\alpha & 1 \end{pmatrix}$$

under the assumption that $\alpha$ and $\omega$ are small.

One of the eigenvalues is exactly one, and its corresponding to the eigenvector is $(0\ \ 0\ \ 1)^{\mathrm{T}}$. The other two eigenvalues are the eigenvalues of the leading principal submatrix

$$\begin{pmatrix} \alpha+\omega & \alpha \\ \alpha & 1+\alpha \end{pmatrix}.$$

Using the quadratic formula, one can easily verify that these eigenvalues are $\alpha+\omega+O(\alpha^2)$ and $1+\alpha+O(\alpha^2)$.

Because the eigenvalue near zero is well separated from the ones near one and because $C$ is diagonal up to terms of order $\alpha$, we can approximate the eigenvector corresponding to the eigenvalue near zero by $(1\ \ \beta\ \ \langle\gamma\rangle)^{\mathrm{T}}$, where $\beta$ and $\langle\gamma\rangle$ are $O(\alpha)$ [6, Section 1.3.2]. Write

$$\begin{pmatrix} \alpha+\omega & \alpha & 0 \\ \alpha & 1+\alpha & 0 \\ -\alpha & -\alpha & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \beta \\ \langle\gamma\rangle \end{pmatrix} \cong (\alpha+\omega) \begin{pmatrix} 1 \\ \beta \\ \langle\gamma\rangle \end{pmatrix}.$$

From the second row of the relation, we get

$$\alpha + (1+\alpha)\beta = (\alpha+\omega)\beta,$$

and ignoring second order terms, we get

$$\beta = -\alpha.$$

Similarly, from the third row we get

$$\langle\gamma\rangle = 0.$$

Thus our second eigenvalue and its eigenvector are $\omega+\alpha$ and $(1\ \ -\alpha\ \ 0)^{\mathrm{T}}$.

The remaining eigenvector cannot be approximated so simply, since its eigenvalue is near the eigenvalue 1. Instead, we use the observation that if $(\lambda, x)$ is a right eigenpair of a matrix and $(\mu, y)$ is a left eigenpair with $\lambda \neq \mu$, then $y^{\mathrm{H}}x = 0$. Thus we will approximate the left eigenvectors corresponding to 1 and $\alpha+\omega$. The eigenvector we seek is then the unique (up to a constant multiple) vector that is orthogonal to both.

It can be easily verified that the left eigenvector corresponding to 1 is $(0\ 1\ 1)^{\mathrm{T}}$. An approximation of the left eigenvector corresponding to $\alpha+\omega$ may be approximated as above. The result is $(1\ \ -\alpha\ \ 0)^{\mathrm{T}}$. The vector orthogonal to both these vectors is $(\alpha\ \ 1\ \ -1)^{\mathrm{T}}$.

Collecting the above results, we obtain (5.3).

## 8. Acknowledgement

Part of this work was performed as faculty appointee at the Mathematical and Computational Sciences Division of the National Institute for Standards and Technology.

## References

[1] Å. Björck. Stability analysis of the method of seminormal equations. *Linear Algebra and Its Applications*, 88/89:31–48, 1987.

[2] Å. Björck. *Numerical Methods for Least Squares Problems.* SIAM, Philadelphia, 1996.

[3] W. Hoffman. Iterative algorithms for Gram–Schmidt orthogonalization. *Computing*, 41:335–348, 1989.

[4] G. W. Stewart. *Matrix Algorithms I: Basic Decompositions.* SIAM, Philadelphia, 1998.

[5] G. W. Stewart. Four algorithms for the the efficient computation of truncated pivoted qr approximations to a sparse matrix. *Numerische Mathematik*, 83:313–323, 1999.

[6] G. W. Stewart. *Matrix Algorithms II: Eigensystems.* SIAM, Philadelphia, 2001.

[7] A. van der Sluis. Condition numbers and equilibration of matrices. *Numerische Mathematik*, 14:14–23, 1969.