SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 914–931

SOLVING THE INDEFINITE LEAST SQUARES PROBLEM BY HYPERBOLIC QR FACTORIZATION*

ADAM BOJANCZYK†, NICHOLAS J. HIGHAM‡, AND HARIKRISHNA PATEL‡

Abstract. The indefinite least squares (ILS) problem involves minimizing a certain type of indefinite quadratic form. We develop perturbation theory for the problem and identify a condition number. We describe and analyze a method for solving the ILS problem based on hyperbolic QR factorization. This method has a lower operation count than one recently proposed by Chandrasekaran, Gu, and Sayed that employs both QR and Cholesky factorizations. We give a rounding error analysis of the new method and use the perturbation theory to show that under a reasonable assumption the method is forward stable. Our analysis is quite general and sheds some light on the stability properties of hyperbolic transformations. In our numerical experiments the new method is just as accurate as the method of Chandrasekaran, Gu, and Sayed.

Key words. indefinite least squares problem, downdating, hyperbolic rotation, hyperbolic QR factorization, rounding error analysis, forward stability, perturbation theory, condition number

AMS subject classifications. 65F20, 65G05

PII. S0895479802401497

1. Introduction.

The indefinite least squares problem (ILS) takes the form

$$\text{ILS}: \qquad \min_x (b-Ax)^T J(b-Ax), \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$, $m \geq n$, and $b \in \mathbb{R}^m$ are given and J is the signature matrix

$$J = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}, \qquad p+q = m. \tag{1.2}$$

For $p = 0$ or $q = 0$ we have the standard least squares (LS) problem and the quadratic form is definite, while for $pq > 0$ the problem is to minimize a genuinely indefinite quadratic form. Chandrasekaran, Gu, and Sayed [3] discuss the application of the ILS problem to the solution of total least squares problems [18] and to the area of optimization known as H^∞ smoothing [8], [14].

*Received by the editors January 24, 2002; accepted for publication (in revised form) by L. Eldén June 25, 2002; published electronically February 12, 2003. This work was supported by Engineering and Physical Sciences Research Council Visiting Fellowship GR/R22414. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defence Advanced Research Project Agency (DARPA), Rome Laboratory, or the U.S. Government. This work was performed by an employee of the U.S. Government or under U.S. government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

http://www.siam.org/journals/simax/24-4/40149.html

†School of Electrical and Computer Engineering, Cornell University, 335 Theory Center and Engineering, Ithaca, NY 14853-3801 (adamb@ee.cornell.edu, http://www.ee.cornell.edu/~adamb/). This author was sponsored by the Defence Advanced Research Project Agency (DARPA) and Rome Laboratory, Air Force Material Command, USAF, under agreement F30602-97-1-0292.

‡Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/; hpatel@ma.man.ac.uk, http://www.ma.man.ac.uk/~hpatel/). The second author's work was supported by Engineering and Physical Sciences Research Council grant GR/R22612. The third author's work was supported by an Engineering and Physical Sciences Research Council Ph.D. Studentship.

The normal equations for (1.1), which are first order conditions for optimality, are

$$A^TJ(b - Ax) = 0. \tag{1.3}$$

Since the Hessian matrix of the quadratic to be minimized in (1.1) is $2A^TJA$, it follows that the ILS problem has a unique solution if and only if

$$A^TJA \text{ is positive definite.} \tag{1.4}$$

We will assume throughout this paper that (1.4) holds. Note that (1.4) implies $p \geq n$ and that $A(1\colon p, 1\colon n)$ (and hence A) has full rank. For a genuinely indefinite LS problem we therefore need $m > n$.

We note in passing that (1.3) gives $x = M^{-1}A^TJb$, where $M = A^TJA$, and the matrix $X = M^{-1}A^TJ$ is a pseudoinverse of A but not the Moore–Penrose pseudoinverse ($XA = I$, but AX is not symmetric).

One way of solving the ILS problem is to form the normal equations and solve them with the aid of a Cholesky factorization. Since this method has poor numerical stability properties for the standard LS problem it is clearly not a good choice for the ILS problem, except perhaps when A^TJA is well conditioned.

Chandrasekaran, Gu, and Sayed [3] propose a method for solving the ILS problem based on a QR factorization of A,

$$A = QR = \begin{matrix} \\ p \\ q \end{matrix} \begin{matrix} \overset{n}{} \\ \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \end{matrix} R, \qquad R \in \mathbb{R}^{n\times n}.$$

This factorization yields

$$A^TJA = R^T(Q_1^TQ_1 - Q_2^TQ_2)R,$$

which, in view of (1.4), implies that R is nonsingular and $Q_1^TQ_1 - Q_2^TQ_2$ is positive definite. Hence the normal equations (1.3) can be rewritten as

$$(Q_1^TQ_1 - Q_2^TQ_2)Rx = Q^TJb. \tag{1.5}$$

Using the Cholesky factorization

$$Q_1^TQ_1 - Q_2^TQ_2 = U^TU,$$

(1.5) becomes

$$U^TURx = Q^TJb.$$

This system can be solved for x by one forward and two backward substitutions. We will refer to this method as the "QR-Cholesky" method. It is shown in [3] that this method produces a computed solution $\widehat{x}$ that solves the problem

$$\min_x \big(b + \Delta b - (A + \Delta A)x)^TJ(b + \Delta b - (A + \Delta A)x\big),$$

where

$$\|\Delta A\|_F \leq c_{m,n}u\|A\|_F, \qquad \|\Delta b\|_2 \leq c_{m,n}u\|b\|_2,$$

with $c_{m,n}$ a constant depending on the problem dimensions and u the unit roundoff; in other words, the QR-Cholesky method is backward stable.

In this work we investigate the solution of the ILS problem via hyperbolic QR factorization. This approach has a lower operation count than the QR-Cholesky method but, in view of the use of hyperbolic transformations, its stability is questionable. We give rounding error analysis and perturbation analysis that combine to show that the method is forward stable under a reasonable assumption and hence of practical interest.

We begin, in the next section, with the perturbation analysis. The hyperbolic QR factorization method is described in section 3, its error analysis is given in section 4, and numerical experiments are presented in section 5. It is an important fact that obtaining useful error bounds for the application of a product of hyperbolic transformations to a vector is much more difficult than when the transformations are orthogonal. In section 4.1 we show how such products can be analyzed under a natural assumption on the form of the hyperbolic transformations.

2. Perturbation theory. In this section we derive normwise and componentwise perturbation bounds for the solution x and a residual r of the ILS problem. Our approach is based on that used by Cox and Higham [4] to obtain perturbation bounds for the equality constrained LS problem. We let $\widetilde{x}$ be the solution of the perturbed ILS problem

$$\min_x \bigl(b + \Delta b - (A + \Delta A)x\bigr)^T J \bigl(b + \Delta b - (A + \Delta A)x\bigr) \tag{2.1}$$

and define

$$\widetilde{r} = b + \Delta b - (A + \Delta A)\widetilde{x}, \qquad r = b - Ax$$

to be the residuals of the perturbed and unperturbed problems, respectively. We assume that $A + \Delta A$ satisfies the uniqueness condition (1.4), which will always be the case for ΔA sufficiently small in norm. The perturbations to the data will be measured by the smallest ϵ for which

$$\|\Delta A\|_F \le \epsilon \|\mathbf{A}\|_F, \quad \|\Delta b\|_2 \le \epsilon \|\mathbf{b}\|_2, \tag{2.2}$$

where $\mathbf{A}$ and $\mathbf{b}$ are a matrix and vector of tolerances.

The normal equations (1.3) can be rewritten as the augmented system (with $r = b - Ax$)

$$\begin{bmatrix} I & A \\ A^T J & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

It is convenient to define $s = Jr$ and rewrite the system with a symmetric coefficient matrix:

$$\begin{bmatrix} J & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} s \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}. \tag{2.3}$$

The perturbed augmented system corresponding to (2.3) is

$$\begin{bmatrix} J & A + \Delta A \\ (A + \Delta A)^T & 0 \end{bmatrix} \begin{bmatrix} \widetilde{s} \\ \widetilde{x} \end{bmatrix} = \begin{bmatrix} b + \Delta b \\ 0 \end{bmatrix}. \tag{2.4}$$

Writing

$$\widetilde{s} = s + \Delta s, \quad \widetilde{x} = x + \Delta x$$

and subtracting (2.3) from (2.4), we obtain

$$\begin{bmatrix} J & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta s \\ \Delta x \end{bmatrix} = \begin{bmatrix} \Delta b - \Delta A\widetilde{x} \\ -\Delta A^T \widetilde{s} \end{bmatrix}. \tag{2.5}$$

It is straightforward to verify that the inverse of the matrix on the left-hand side of (2.5) is

$$\begin{bmatrix} J - JAM^{-1}A^TJ & JAM^{-1} \\ M^{-1}A^TJ & -M^{-1} \end{bmatrix}, \qquad \text{where } M = A^TJA.$$

Premultiplying by the inverse and expanding the right-hand side, we obtain

$$\Delta s = (J - JAM^{-1}A^TJ)(\Delta b - \Delta A\widetilde{x}) - (JAM^{-1})\Delta A^T\widetilde{s}, \tag{2.6a}$$

$$\Delta x = M^{-1}A^TJ(\Delta b - \Delta A\widetilde{x}) + M^{-1}\Delta A^T\widetilde{s}. \tag{2.6b}$$

If we put $J = I$, then we recover perturbation expressions for the standard LS problem.

Since the perturbations Δs and Δx are of order ϵ, we can substitute $s = Jr$ and x for their perturbed counterparts to obtain first order expressions. Then, taking norms, we deduce

$$\begin{aligned} \|\Delta r\|_2 &\le \epsilon\Big[\|I - JAM^{-1}A^T\|_2(\|\mathbf{b}\|_2 + \|\mathbf{A}\|_F\|x\|_2) + \|AM^{-1}\|_2\|\mathbf{A}\|_F\|r\|_2\Big] \\ &\quad + O(\epsilon^2), \\ \|\Delta x\|_2 &\le \epsilon\Big[\|M^{-1}A^T\|_2(\|\mathbf{b}\|_2 + \|\mathbf{A}\|_F\|x\|_2) + \|M^{-1}\|_2\|\mathbf{A}\|_F\|r\|_2\Big] + O(\epsilon^2). \end{aligned} \tag{2.7}$$

Hence, provided $x \neq 0$,

$$\begin{aligned} \frac{\|\Delta x\|_2}{\|x\|_2} \le \epsilon\Bigg[&\|M^{-1}A^T\|_2\|\mathbf{A}\|_F\left(\frac{\|\mathbf{b}\|_2}{\|\mathbf{A}\|_F\|x\|_2} + 1\right) \\ &+ \|M^{-1}\|_2\|A\|_F^2\frac{\|\mathbf{A}\|_F}{\|A\|_F}\frac{\|r\|_2}{\|A\|_F\|x\|_2}\Bigg] + O(\epsilon^2). \end{aligned} \tag{2.8}$$

This bound shows that the sensitivity of the ILS problem is bounded in terms of $\|M^{-1}A^T\|_2\|\mathbf{A}\|_F$ when the residual is zero or small and $\|M^{-1}\|_2\|A\|_F^2$ otherwise; note that for $\mathbf{A} = A$ the former quantity is no larger than the latter and is potentially much smaller.

Now we examine whether (2.8) is attainable for some ΔA and Δb. The three terms in brackets in (2.7) are

$$E_1 = \|M^{-1}A^T\|_2\|\mathbf{b}\|_2, \quad E_2 = \|M^{-1}A^T\|_2\|\mathbf{A}\|_F\|x\|_2, \quad E_3 = \|M^{-1}\|_2\|\mathbf{A}\|_F\|r\|_2,$$

and they result from the perturbations Δb, ΔA, and ΔA^T, respectively. It follows that the bound (2.7) can fail to be achieved for some Δb and ΔA only if $E_1 < E_2 \approx E_3$ and there is substantial cancellation in the expression $-M^{-1}A^TJ\Delta Ax + M^{-1}\Delta A^TJr$ for all ΔA. We can show in various special cases that these circumstances cannot

arise (for example, when r is small, or when $|r^T Jr| \approx \|r\|_2^2$), but we have been unable to establish attainability of the bound (2.8) in general.

A natural definition of the condition number of the ILS problem is

$$\kappa_{\mathrm{ILS}}(A,b) = \lim_{\epsilon\to 0}\sup\left\{\frac{\|x-\widetilde{x}\|_2}{\|x\|_2} : (2.2)\text{–}(2.4)\text{ hold}\right\}. \tag{2.9}$$

Without a guarantee of sharpness, the bound (2.8) does not provide an estimate of $\kappa_{\mathrm{ILS}}(A,b)$ to within a readily identifiable constant factor. Therefore we take a different approach in which we combine the two ΔA terms in (2.6b) before taking norms. To do this, we use the vec operator, which stacks the columns of a matrix into one long column vector, together with the Kronecker product $A\otimes B = (a_{ij}B)$, which for $A\in\mathbb{R}^{m\times n}$ and $B\in\mathbb{R}^{p\times q}$ is the block matrix $(a_{ij}B)\in\mathbb{R}^{mp\times nq}$ (see [9], [11, Chap. 4]). Applying the vec operator to (2.6b) and using the relation $\mathrm{vec}(AXB) = (B^T\otimes A)\mathrm{vec}(X)$, we obtain

$$\Delta x = M^{-1}A^T J\Delta b - (x^T\otimes M^{-1}A^T J)\mathrm{vec}(\Delta A) + (r^T J\otimes M^{-1})\mathrm{vec}(\Delta A^T) + O(\epsilon^2).$$

Using the relation $\mathrm{vec}(\Delta A^T) = \Pi\mathrm{vec}(\Delta A)$, where Π is the vec-permutation matrix [9], gives

$$\Delta x = M^{-1}A^T J\Delta b - \left[(x^T\otimes M^{-1}A^T J) - (r^T J\otimes M^{-1})\Pi\right]\mathrm{vec}(\Delta A) + O(\epsilon^2).$$

Now we take 2-norms. Using (2.2) and the fact that $\|\mathrm{vec}(\Delta A)\|_2 = \|\Delta A\|_F$, we deduce that

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \psi\epsilon + O(\epsilon^2), \tag{2.10}$$

where

$$\psi = \left(\|M^{-1}A^T\|_2\|\mathbf{b}\|_2 + \|(x^T\otimes M^{-1}A^T J) - (r^T J\otimes M^{-1})\Pi\|_2\|\mathbf{A}\|_F\right)/\|x\|_2,$$

and we have

$$\kappa_{\mathrm{ILS}}(A,b) \le \psi \le 2\kappa_{\mathrm{ILS}}(A,b).$$

In extensive numerical comparisons between the first order terms of the bounds (2.8) and (2.10), including with direct search optimization, we have found these terms always to be within a small factor of each other. We believe that (2.8) is nearly attainable and, because this bound is much easier to work with than (2.10), we will use it when we investigate the stability of hyperbolic QR factorization for solving the ILS problem.

To end this section, we note that we can also use (2.6) to obtain componentwise perturbation bounds for the ILS problem. For the solution, we obtain

$$|\Delta x| \le \epsilon|M^{-1}A^T|(\mathbf{b} + \mathbf{A}|x|) + |M^{-1}|\mathbf{A}^T|r| + O(\epsilon^2),$$

where inequalities and the absolute value are interpreted componentwise and ϵ has been redefined as the smallest value for which $|\Delta A| \le \epsilon\mathbf{A}$, $|\Delta b| \le \epsilon\mathbf{b}$, where $\mathbf{A}$ and $\mathbf{b}$ are now assumed to have nonnegative entries.

3. Hyperbolic QR factorization method. We define a matrix $Q \in \mathbb{R}^{m\times m}$ to be J-orthogonal if

$$Q^T J Q = J,$$

or, equivalently, $QJQ^T = J$, where J is defined in (1.2). Suppose we can find a J-orthogonal matrix Q such that

$$Q^T A = Q^T \begin{array}{c} {\scriptstyle p} \\ {\scriptstyle q} \end{array} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{array}{c} {\scriptstyle n} \\ {\scriptstyle m-n} \end{array} \begin{array}{c} {\scriptstyle n} \\ \end{array} \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{3.1}$$

where $R \in \mathbb{R}^{n\times n}$ is upper triangular. We refer to this factorization as a *hyperbolic QR factorization*. Then

$$Q^T(b - Ax) = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} - \begin{bmatrix} R \\ 0 \end{bmatrix} x = \begin{bmatrix} d_1 - Rx \\ d_2 \end{bmatrix}, \qquad \begin{array}{c} {\scriptstyle n} \\ {\scriptstyle m-n} \end{array} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = Q^T b,$$

and so

$$\begin{aligned} (b - Ax)^T J (b - Ax) &= (b - Ax)^T Q J Q^T (b - Ax) \\ &= \begin{bmatrix} d_1 - Rx \\ d_2 \end{bmatrix}^T J \begin{bmatrix} d_1 - Rx \\ d_2 \end{bmatrix} \\ &= \|d_1 - Rx\|_2^2 + d_2^T J(n+1\colon m, n+1\colon m) d_2, \end{aligned} \tag{3.2}$$

recalling that (1.4) implies $p \geq n$ in (1.2). Hence the ILS solution is obtained by solving $Rx = d_1$. This method is an analogue of Golub's method for the LS problem [7].

The matrix Q can be constructed as a product of hyperbolic rotations and orthogonal matrices. A 2×2 hyperbolic rotation has the form

$$H = \begin{bmatrix} c & -s \\ -s & c \end{bmatrix}, \qquad c^2 - s^2 = 1,$$

and it is so named because $|c| = \cosh\theta$ and $s = \sinh\theta$ for some θ. It is easy to check that H is J-orthogonal for $J = \operatorname{diag}(1, -1)$. We will choose H to effect the zeroing operation

$$\begin{bmatrix} c & -s \\ -s & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix},$$

which requires that $cx_2 = sx_1$. The latter equation has a real solution only when $|x_1| > |x_2|$, in which case

$$c = \frac{x_1}{\sqrt{x_1^2 - x_2^2}}, \qquad s = \frac{x_2}{\sqrt{x_1^2 - x_2^2}}. \tag{3.3}$$

In practice a rescaling of these formulas is desirable to reduce the risk of overflow.

```
function [c, s] = Hrotate(x1, x2)
% Compute c and s defining hyperbolic rotation H such that
% Hx has zero second element.
if |x1| > |x2|
   t = x2/x1, c = 1/sqrt(1 - t^2), s = ct
else
   No real rotation exists—abort.
end
```

Unlike for orthogonal rotations, how hyperbolic rotations are applied to a vector is crucial to the stability of the computation [1], [15]. Consider the computation of $y = Hx$:

$$\begin{aligned} y_1 &= cx_1 - sx_2, \\ y_2 &= -sx_1 + cx_2. \end{aligned} \tag{3.4}$$

The first equation gives

$$x_1 = \frac{y_1}{c} + \frac{s}{c}x_2, \tag{3.5}$$

which allows the second to be rewritten as

$$\begin{aligned} y_2 &= -\frac{s}{c}y_1 + \left(-\frac{s^2}{c} + c\right)x_2 \\ &= -\frac{s}{c}y_1 + \frac{x_2}{c}. \end{aligned} \tag{3.6}$$

We will apply hyperbolic rotations using (3.4) and (3.6). As noted by Park and Eldén [13], this way of forming the product $y = Hx$ corresponds to use of the rescaled LU factorization

$$H = \begin{bmatrix} c & -s \\ -s & c \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -s/c & 1/c \end{bmatrix} \begin{bmatrix} c & -s \\ 0 & 1 \end{bmatrix}.$$

That this way of forming y is advantageous for stability was proved in [1] in the context of downdating a Cholesky factorization. We express the formation as follows:

function $B = \text{Happly}(c, s, B)$
% Apply hyperbolic rotation defined by c and s to $2 \times n$ matrix B.
for $j = 1\colon n$
    $B(1,j) = cB(1,j) - sB(2,j)$
    $B(2,j) = -(s/c)B(1,j) + B(2,j)/c$
end

For later use we note that (3.5) and (3.6) can be expressed together in the form

$$\begin{bmatrix} x_1 \\ y_2 \end{bmatrix} = G \begin{bmatrix} y_1 \\ x_2 \end{bmatrix}, \tag{3.7}$$

where

$$G = \begin{bmatrix} 1/c & s/c \\ -s/c & 1/c \end{bmatrix} \equiv \begin{bmatrix} \widetilde{c} & \widetilde{s} \\ -\widetilde{s} & \widetilde{c} \end{bmatrix}, \qquad \widetilde{c}^{\,2} + \widetilde{s}^{\,2} = 1.$$

The matrix G is a Givens rotation. Hence function Happly can be interpreted as forming the first row of the product HB by a hyperbolic rotation and the second row by a Givens rotation.

Our algorithm for computing the triangular factor R in (3.1) begins by computing the QR factorization

$$A_1 = Q_1 R_1, \qquad Q_1 \in \mathbb{R}^{p\times p}, \quad R_1 \in \mathbb{R}^{p\times n},$$

where Q_1 is orthogonal. Defining $\widetilde{Q} = \text{diag}(Q_1^T, I_q)$ we have

$$A^{(1)} = \widetilde{Q}A = \begin{bmatrix} R_1 \\ A_2 \end{bmatrix},$$

and $\widetilde{Q}$ is trivially J-orthogonal. We now zero A_2 with the aid of hyperbolic rotations. This can be done entirely with hyperbolic rotations or with a mix of hyperbolic and orthogonal rotations. Since hyperbolic rotations do not preserve the norms of vectors to which they are applied, we will use the minimum number, n, of them.

From a 2×2 hyperbolic rotation we build an $m \times m$ rotation in the (i, j) plane, $H_{i,j}$, defined to be the identity matrix modified according to $h_{ii} = h_{jj} = c$ and $h_{ij} = h_{ji} = -s$. Note that, provided the indices satisfy $i \leq p$ and $j > p$, H_{ij} is J-orthogonal. The parameters c and s are chosen to zero the jth element of the vector to which H_{ij} is applied.

Consider the first column of $A^{(1)}$. We first zero the elements in positions $(p+2, 1)$, $(p+3, 1), \ldots, (m, 1)$ using a Householder transformation, P_1, acting on rows $p+1\colon m$. Then we eliminate the $(p+1, 1)$ element, which is the sole remaining subdiagonal element in column 1, by a hyperbolic rotation $H_{1,p+1}$. It is clear that these operations do not disturb the existing zeros in positions $(2\colon p, 1)$ of $A^{(1)}$. At this point we have formed

$$A^{(2)} := H_{1,p+1} P_1 A^{(1)} =: Q^{(1)} A, \tag{3.8}$$

where $A^{(2)}(2\colon m, 1) = 0$. The matrix $Q^{(1)}$ is a product of J-orthogonal matrices and so is J-orthogonal. Elements below the diagonal in the remaining columns are eliminated in an analogous way, with the hyperbolic rotation used for the jth column being in the $(j, p+1)$ plane. The complete algorithm for solving the ILS problem is summarized as follows.

ALGORITHM 1. *This algorithm solves the ILS problem* (1.1) *using Householder QR factorization and hyperbolic rotations.*

```
Compute the Householder QR factorization A(1:p,:) = Q_1 R_1
    (Q_1 ∈ R^{p×p}, R_1 ∈ R^{p×n}), overwriting A(1:p,:) with R and
    b(1:n) with Q(1:n,:)^T b(1:n).
for j = 1:min(m-1, n)
    Construct a Householder transformation H_j such that
        H_j A(p+1:m, j) = σ_j e_1.
    A(p+1:m, j:n) = H_j A(p+1:m, j:n)
    b(p+1:m) = H_j b(p+1:m)
    % Eliminate sole remaining subdiagonal element in column j by a
    % hyperbolic rotation.
    [c, s] = Hrotate(A(j, j), A(p+1, j))
    A([j p+1], j:n) = Happly(c, s, A([j p+1], j:n))
    b([j p+1]) = Happly(c, s, b([j p+1]))
end
R = A(1:n,:)
Solve Rx = b(1:n) by substitution.
```

The operation count of Algorithm 1 is the same as that for solution of the standard LS problem by Householder QR factorization (essentially because the hyperbolic rotations contribute only to the lower order terms in the operation count). Table 3.1 compares the cost of Algorithm 1 with the cost of forming and solving the normal equations (1.3) and the cost of the QR-Cholesky method. Algorithm 1 requires fewer operations than the QR-Cholesky method by a factor 2.5–3.

It remains to show that the desired hyperbolic rotations exist. Suppose the algorithm has succeeded in eliminating the first $k-1$ columns of A_2, yielding $A_2^{(k)}$, and

TABLE 3.1
Operation counts for methods for solving the ILS problem.

	Normal equations	Hyperbolic QR	QR-Cholesky
	$n^2(m+n/3)$	$2n^2(m-n/3)$	$n^2(5m-n)$
$m \approx n$	$4n^3/3$	$4n^3/3$	$4n^3$
$m \gg n$	mn^2	$2mn^2$	$5mn^2$

define $C \in \mathbb{R}^{n\times q}$ and $R_1^{(k)} \in \mathbb{R}^{n\times n}$ by

$$R_1^{(k)^T} C \equiv A_1^{(k)^T} \begin{bmatrix} C \\ 0 \end{bmatrix} = A_2^{(k)^T}. \tag{3.9}$$

Since

$$\begin{bmatrix} A_1^{(k)} \\ A_2^{(k)} \end{bmatrix} = Q^{(k-1)} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

for a J-orthogonal matrix $Q^{(k-1)}$ and (1.4) holds, the matrix

$$A_1^{(k)^T} A_1^{(k)} - A_2^{(k)^T} A_2^{(k)} = R_1^{(k)^T} (I - CC^T) R_1^{(k)}$$

is positive definite. Since $R_1^{(k)}$ is upper triangular and nonsingular, it follows that $I - CC^T$ is positive definite and hence that $|c_{ij}| < 1$ for all i and j. Now $A_1^{(k)}$ is upper triangular and $A_2^{(k)}(1\colon q, 1\colon k-1) = 0$, so, using (3.9),

$$1 > |c_{ki}| = \left| \frac{a_{p+i,k}^{(k)}}{a_{k,k}^{(k)}} \right|, \qquad i = 1\colon q,$$

which ensures the existence of the hyperbolic rotation required on the $(k+1)$st stage.

4. Rounding error analysis. We now give a rounding error analysis of Algorithm 1. First, we note that from (3.1) we have $R^TR = A_1^TA_1 - A_2^TA_2$. Hence if A_1 is upper trapezoidal, then the hyperbolic QR factor R is the result of (block) downdating a Cholesky factorization. Various algorithms, both hyperbolic and nonhyperbolic, are known for downdating Cholesky factorizations, and error analysis is available; see, for example, [1], [2], [5], [6], [15], [17]. While we could invoke some of the earlier results in the part of the analysis that does not involve the right-hand side, b, we have chosen to give an independent development, aiming to make clear how the various errors combine and provide building blocks that should be of use in future analyses. In particular, we emphasize the high-level features of the analysis and thereby provide new insight into what is required of a sequence of hyperbolic transformations in order for satisfactory error bounds to be obtainable.

We use the standard model of floating point arithmetic [10, sect. 2.2]:

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\delta)^{\pm 1}, \qquad |\delta| \le u, \quad \operatorname{op} = +, -, *, /,$$

where u is the unit roundoff. Our bounds are expressed in terms of the constants

$$\gamma_k = \frac{ku}{1-ku}, \qquad \widetilde{\gamma}_k = \frac{cku}{1-cku}, \tag{4.1}$$

where c denotes a small integer constant whose exact value is unimportant. We also employ the relative error counter, $\langle k \rangle$:

$$\langle k \rangle = \prod_{i=1}^{k} (1+\delta_i)^{\rho_i}, \qquad \rho_i = \pm 1, \quad |\delta_i| \le u. \tag{4.2}$$

We use the fact that $|\langle k \rangle - 1| \le \gamma_k = ku/(1-ku)$ [10, Lem. 3.1].

Given an error bound for a single orthogonal transformation it is relatively easy to obtain a useful error bound for a product of several orthogonal transformations, as first shown by Wilkinson in the 1960s. The situation is quite different for a product of hyperbolic transformations, $y = H_p \dots H_2 H_1 x$, say. It is possible to mimic the analysis for orthogonal transformations and write, for example, $\widehat{y} = (H_p + \Delta H_p) \dots (H_2 + \Delta H_2)(H_1 + \Delta H_1)x$, with each ΔH_j bounded relative to H_j. However, this expression does not lead to a satisfactory forward or backward error bound, because the H_j are unbounded in norm. A better approach is to exploit the following equivalence between orthogonal and hyperbolic transformations.

Let

$$A = \begin{matrix} p \\ q \end{matrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \begin{matrix} n \\ \end{matrix} = \begin{matrix} p \\ q \end{matrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{matrix} p & q \\ \end{matrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \begin{matrix} p \\ q \end{matrix} \begin{matrix} n \\ \end{matrix} = QB, \tag{4.3}$$

where Q is J-orthogonal, for J in (1.2). Then $Q_{11}^T Q_{11} = I + Q_{21}^T Q_{21}$, and hence Q_{11} is nonsingular. It is not hard to show that

$$\begin{bmatrix} B_1 \\ A_2 \end{bmatrix} = \operatorname{exc}(Q) \begin{bmatrix} A_1 \\ B_2 \end{bmatrix}, \tag{4.4}$$

where the matrix

$$\operatorname{exc}(Q) = \begin{bmatrix} Q_{11}^{-1} & -Q_{11}^{-1} Q_{12} \\ Q_{21} Q_{11}^{-1} & Q_{22} - Q_{21} Q_{11}^{-1} Q_{12} \end{bmatrix}$$

is orthogonal. Moreover, if P is an orthogonal matrix partitioned in the same way as Q and its $(1,1)$ block is nonsingular, then $\operatorname{exc}(P)$ is J-orthogonal. In fact, the exchange operator is involutary: $\operatorname{exc}(\operatorname{exc}(P)) = P$. Note that (3.7) is a special case of (4.4). For proofs of these properties, see [12, Lem. 1], [17, sect. 2].

The advantage of (4.4) is that because the transformation matrix is orthogonal error terms can be moved around in the equation without changing their norm. The disadvantage is that it is hard to analyze more than one transformation. For example, let $C = PA$, where P is J-orthogonal. Then $C = PQB$ and corresponding to (4.4) we have

$$\begin{bmatrix} A_1 \\ C_2 \end{bmatrix} = \operatorname{exc}(PQ) \begin{bmatrix} C_1 \\ A_2 \end{bmatrix}. \tag{4.5}$$

Despite the elegance of this relation, $\operatorname{exc}(PQ)$ is a complicated function of P and Q. In practice the equations $A = QB$ and $C = PA$ must be modified to include rounding error terms, and these terms appear to preclude a suitably perturbed version of (4.5) with satisfactory bounds on the perturbations.

The gist of this analysis is that it is unclear how to obtain useful error bounds for the product of two or more *arbitrary* hyperbolic transformations. Fortunately, the transformations in Algorithm 1 are far from arbitrary, and in the next two sections we show that by exploiting their structure we can make useful progress.

4.1. Combining two hyperbolic transformations. We now analyze a product of two hyperbolic transformations that satisfy one key assumption: that the two transformations are "nonoverlapping" in components $1\!:\!p$. Nonoverlapping means that for $i = 1\!:\!p$ at least one of the two transformations agrees with the identity matrix in row i and column i. Without loss of generality, we consider a transformation $H_1(2,3)$ agreeing with the identity matrix in rows and columns $1\!:\!t$ and a transformation $H_2(1,3)$ agreeing with the identity matrix in rows and columns $t+1\!:\!p$, where $1 \le t < p$. Let

$$H_1(2,3)\begin{bmatrix} R \\ S \\ X \end{bmatrix}\begin{matrix} t \\ p-t \\ q \end{matrix} =: \begin{bmatrix} R \\ S_1 \\ X_1 \end{bmatrix}\begin{matrix} t \\ p-t \\ q \end{matrix}, \qquad H_2(1,3)\begin{bmatrix} R \\ S_1 \\ X_1 \end{bmatrix} =: \begin{bmatrix} R_1 \\ S_1 \\ X_2 \end{bmatrix},$$

or, overall,

$$H_2(1,3)H_1(2,3)\begin{bmatrix} R \\ S \\ X \end{bmatrix} = \begin{bmatrix} R_1 \\ S_1 \\ X_2 \end{bmatrix}.$$

We know from (4.3) and (4.4) that these two operations can be rewritten in terms of orthogonal transformations G_i as follows, where we now express the relations in terms of the affected components only:

$$G_1\begin{bmatrix} S_1 \\ X \end{bmatrix} = \begin{bmatrix} S \\ X_1 \end{bmatrix}, \tag{4.6a}$$

$$G_2\begin{bmatrix} R_1 \\ X_1 \end{bmatrix} = \begin{bmatrix} R \\ X_2 \end{bmatrix}. \tag{4.6b}$$

These two relations can be rewritten as

$$\begin{bmatrix} R_1 \\ S_1 \\ X \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & G_1^T \end{bmatrix}\begin{bmatrix} R_1 \\ S \\ X_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & G_1^T \end{bmatrix}\widetilde{G}_2^T\begin{bmatrix} R \\ S \\ X_2 \end{bmatrix} \equiv G\begin{bmatrix} R \\ S \\ X_2 \end{bmatrix}, \tag{4.7}$$

where $\widetilde{G}_2([1\!:\!t, p{+}1\!:\!m], [1\!:\!t, p{+}1\!:\!m]) = G_2$ and elsewhere $\widetilde{G}_2$ agrees with the identity matrix, and G is orthogonal. This relation shows that $\operatorname{exc}(H_2(1,3)H_1(2,3)) = G$ is of a relatively simple form given the no-overlap assumption.

Now we incorporate errors into the analysis. Consider the perturbed versions of (4.6),

$$G_1\begin{bmatrix} S_1 + E_1 \\ X + E_2 \end{bmatrix} = \begin{bmatrix} S \\ X_1 \end{bmatrix}, \tag{4.8a}$$

$$G_2\begin{bmatrix} R_1 + F_1 \\ X_1 + F_2 \end{bmatrix} = \begin{bmatrix} R \\ X_2 \end{bmatrix}, \tag{4.8b}$$

where

$$\max_{i=1,2}\|E_i\|_2 \le \mu\max(\|S_1\|_2, \|X\|_2), \quad \max_{i=1,2}\|F_i\|_2 \le \mu\max(\|R_1\|_2, \|X_1\|_2). \tag{4.9}$$

We will show below that perturbations of this form model rounding errors in Algorithm 1.

We now obtain an analogue of (4.7) for the perturbed quantities. We have

$$\begin{aligned}\begin{bmatrix} R_1 + F_1 \\ S_1 + E_1 \\ X + E_2 \end{bmatrix} &= \begin{bmatrix} I & 0 \\ 0 & G_1^T \end{bmatrix} \begin{bmatrix} R_1 + F_1 \\ S \\ X_1 \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & G_1^T \end{bmatrix} \left(\widetilde{G}_2^T \begin{bmatrix} R \\ S \\ X_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ F_2 \end{bmatrix} \right).\end{aligned}$$

This may be rewritten as

$$\begin{bmatrix} R_1 \\ S_1 \\ X \end{bmatrix} + \Delta = G \begin{bmatrix} R \\ S \\ X_2 \end{bmatrix}, \tag{4.10}$$

where, using $\|X_1\|_2 \le 2\max(\|S_1\|_2, \|X\|_2) + O(\mu)$,

$$\Delta = \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{bmatrix}, \qquad \max_i \|\Delta_i\|_2 \le 3\mu \max(\|R_1\|_2, \|X\|_2, \|S_1\|_2) + O(\mu^2).$$

The key fact is that the error bound for the two transformations combined is commensurate with that for the individual transformations. Because G is orthogonal the relation (4.10) can, if desired, be rewritten so that the 3×1 block matrix on the right is perturbed instead of the one on the left, as in the assumptions (4.8a) and (4.8b).

4.2. One rotation. Now we analyze the application of a hyperbolic rotation. We make the simplifying assumption that c and s in (3.3) are computed exactly.

The computed quantities from (3.4) and (3.6) satisfy

$$\widehat{y}_1\langle 1\rangle = cx_1\langle 1\rangle - sx_2\langle 1\rangle,$$

that is,

$$x_1 = \frac{\widehat{y}_1}{c}\langle 2\rangle + \frac{s}{c}x_2\langle 2\rangle,$$

and

$$\widehat{y}_2 = -\frac{s}{c}\widehat{y}_1\langle 3\rangle + \frac{x_2}{c}\langle 2\rangle.$$

Hence the analogue of (3.7) for the computed quantities is

$$\begin{bmatrix} x_1 \\ \widehat{y}_2 \end{bmatrix} = \begin{bmatrix} g_{11}\langle 2\rangle & g_{12}\langle 2\rangle \\ g_{21}\langle 3\rangle & g_{22}\langle 2\rangle \end{bmatrix} \begin{bmatrix} \widehat{y}_1 \\ x_2 \end{bmatrix} = (G + \Delta G) \begin{bmatrix} \widehat{y}_1 \\ x_2 \end{bmatrix}, \qquad |\Delta G| \le \gamma_3 |G|,$$

where G is orthogonal. This result can be rewritten as

$$\begin{bmatrix} x_1 \\ \widehat{y}_2 \end{bmatrix} = G \begin{bmatrix} \widehat{y}_1 + e_1 \\ x_2 + e_2 \end{bmatrix},$$

where

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = G^T \Delta G \begin{bmatrix} \widehat{y}_1 \\ x_2 \end{bmatrix},$$

so that

$$\max(|e_1|, |e_2|) \le \gamma_3(1 + 2|\widetilde{c}\,||\widetilde{s}|)\max(|\widehat{y}_1|, |x_2|) \le \gamma_6 \max(|\widehat{y}_1|, |x_2|).$$

This is a mixed backward–forward error result, since one element of each of the input and output vectors is perturbed. Importantly, this result is of the form (4.8).

4.3. Hyperbolic QR factorization. As before, we partition

$$A = \begin{matrix} p \\ q \end{matrix} \overset{n}{\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}}.$$

The first stage of Algorithm 1 computes the Householder QR factorization $A_1 = Q_1 \widetilde{R}_1$, where $\widetilde{R}_1 \in \mathbb{R}^{n\times n}$ is upper trapezoidal. We know from standard error analysis that the computed $\widetilde{R}_1$ is the exact factor of $A_1 + \Delta_1$, with $\|\Delta_1\|_F \le \widetilde{\gamma}_{pn}\|A_1\|_F$ [10, Thm. 19.4]. To simplify the notation, we will assume for the moment that A_1 is already in upper trapezoidal form and will introduce the error Δ_1 at the end.

Consider the jth column of A,

$$A(:,j) = \begin{matrix} p \\ q \end{matrix} \overset{n}{\begin{bmatrix} a_j^{(1)} \\ a_j^{(2)} \end{bmatrix}}.$$

It undergoes n Householder transformations in the last q components, intertwined with n hyperbolic rotations in the planes $(1, p+1), \dots, (n, p+1)$; the jth pair of these transformations introduce the required zeros in this column. The final $n-j$ pairs of transformations leave the column unchanged.

Consider a Householder transformation and the subsequent hyperbolic rotation. The Householder transformation agrees with the identity in rows and columns $1\colon p$ and its application is described by a standard backward stability result [10, Lem. 19.2]. It satisfies (4.8a) and (4.9) with $E_1 = 0$ and $\mu = \widetilde{\gamma}_q$. The hyperbolic rotation satisfies the bound of section 4.2 and is nonoverlapping with the Householder transformation. Therefore the analysis of section 4.1 can be applied to these two transformations. Importantly, all the subsequent pairs of Householder and hyperbolic rotations are mutually nonoverlapping and so the result of section 4.1 can be applied inductively.

The overall finding relating the jth columns of A and the final upper trapezoidal factor R_1 is that

$$\begin{matrix} p \\ q \end{matrix} \begin{bmatrix} \widehat{r}_j \\ a_j^{(2)} \end{bmatrix} + h_j = G \begin{bmatrix} a_j^{(1)} \\ 0 \end{bmatrix} \begin{matrix} p \\ q \end{matrix}, \qquad \|h_j\|_2 \le \widetilde{\gamma}_{qj} \max(\|\widehat{r}_j\|_2, \|a_j^{(2)}\|_2)$$

for some exactly orthogonal G that is *independent of* j. Importantly, $h_j(n+1\colon p) = 0$, because after the initial Householder QR factorization rows $n+1\colon p$ of A rest untouched. Putting these equations together for $j = 1\colon n$ and incorporating the error from the initial QR factorization of A_1 gives

$$\begin{bmatrix} \widehat{R}_1 + \Delta_3 \\ A_2 + \Delta_2 \end{bmatrix} = G \begin{bmatrix} A_1 + \Delta_1 \\ 0 \end{bmatrix}, \tag{4.11}$$

where $\Delta_3(n+1\colon p, :) = 0$ and

$$\|\Delta_1\|_F \le \widetilde{\gamma}_{pn}\|A_1\|_F, \tag{4.12a}$$

$$\|\Delta_i\|_F \le \widetilde{\gamma}_{qn} \max(\|\widehat{R}_1\|_F, \|A_2\|_F) \le \widetilde{\gamma}_{qn}\|A_1\|_F, \quad i = 2\colon 3. \tag{4.12b}$$

Certainly, $\max_i \|\Delta_i\|_F \le \widetilde{\gamma}_{mn}\|A\|_F$.

Note that in view of the equivalence (4.3) and (4.4), as long as G has a nonsingular $(1,1)$ block this result is equivalent to

$$\begin{bmatrix} A_1 + \Delta_1 \\ A_2 + \Delta_2 \end{bmatrix} = Q \begin{bmatrix} \widehat{R}_1 + \Delta_3 \\ 0 \end{bmatrix} \tag{4.13}$$

for a J-orthogonal Q. Both (4.11) and (4.13) are mixed backward–forward error results, because both the original data A and the trapezoidal factor R_1 are perturbed. We can obtain a genuine backward error result with the aid of the following lemma (for a proof, see [16, pp. 302–304]).

LEMMA 4.1. *Let $m = p + q$ and $n \geq p$. Given a full rank matrix $A \in \mathbb{R}^{p\times n}$ and $E \in \mathbb{R}^{q\times n}$ there exists an orthogonal $Q \in \mathbb{R}^{m\times m}$ such that*

$$Q \begin{bmatrix} A \\ E \end{bmatrix} = \begin{bmatrix} A + F \\ 0 \end{bmatrix}, \tag{4.14}$$

where, for small $\|E\|_2$,

$$\|F\|_2 \leq \frac{\|E\|_2^2}{2\sigma_{\min}(A)} + O(\|E\|_2^4). \qquad \square$$

Rewriting (4.11) as

$$\begin{bmatrix} \widehat{R}_1 \\ A_2 + \Delta_2 \end{bmatrix} = G \begin{bmatrix} A_1 + \Delta_1 + \widetilde{\Delta}_1 \\ \widetilde{\Delta}_2 \end{bmatrix}, \qquad \widetilde{\Delta} = -G^T \begin{bmatrix} \Delta_3 \\ 0 \end{bmatrix}$$

and applying Lemma 4.1 to the right-hand side leads to the conclusion that

$$\begin{bmatrix} \widehat{R}_1 \\ A_2 + \Delta_2 \end{bmatrix} = \widetilde{G} \begin{bmatrix} A_1 + \overline{\Delta}_1 \\ 0 \end{bmatrix},$$

where $\widetilde{G}$ is orthogonal and

$$\begin{aligned} \|\Delta_2\|_F &\leq \widetilde{\gamma}_{mn} \|A\|_F, \\ \|\overline{\Delta}_1\|_F &\leq \frac{\tilde{\gamma}_{mn}^2 \|A_1\|_F^2}{2\sigma_{\min}(A_1 + \Delta_1 + \widetilde{\Delta}_1)} + O(u^4) \\ &\leq \frac{\sqrt{n}}{2} \big(\kappa_2(A_1)\widetilde{\gamma}_{mn}\big)\widetilde{\gamma}_{mn} \|A\|_F + O(u^3). \end{aligned}$$

We conclude that backward stability of the factorization is guaranteed if $\kappa_2(A_1)u$ is of order 1. Thus the factorization is only conditionally backward stable, although the condition is quite weak. To relate the condition of A_1 to the sensitivity of the ILS problem, we note that

$$\kappa_2(A_1) \leq \big(\|M^{-1}\|_2 \|A\|_2^2\big)^{1/2},$$

where $M = A^T J A = A_1^T A_1 - A_2^T A_2$, from which it follows that if the perturbation bound (2.8) is small and the residual is not small, then A_1 must be well conditioned.

4.4. Solving the ILS problem. In solving the ILS problem we also transform the right-hand side $b = [b_1^T\ b_2^T]^T$ to $d = [d_1^T\ d_2^T]^T$. The above analysis gives

$$\begin{matrix} p \\ q \end{matrix} \begin{bmatrix} \widehat{d}_1 + \delta_3 \\ b_2 + \delta_2 \end{bmatrix} = G \begin{bmatrix} b_1 + \delta_1 \\ \widehat{d}_2 \end{bmatrix} \begin{matrix} p \\ q \end{matrix}, \tag{4.15}$$

where $\delta_3(n+1\colon p) = 0$ and

$$\|\delta_1\|_2 \le \widetilde{\gamma}_p \|b_1\|_2, \qquad \|\delta_i\|_2 \le \widetilde{\gamma}_q \max(\|\widehat{d}_1(1\colon n)\|_2, \|b_2\|_2), \quad i = 2\colon 3.$$

The ensuing analysis is simpler if $\widehat{d}_1$ is not perturbed, so we rewrite this relation as

$$\begin{bmatrix} \widehat{d}_1 \\ b_2 + \overline{\delta}_2 \end{bmatrix} = G \begin{bmatrix} b_1 + \overline{\delta}_1 \\ \widehat{d}_2 + \overline{\delta}_3 \end{bmatrix}, \tag{4.16}$$

where $\overline{\delta}_2 = \delta_2$ and

$$\max_{i=1\colon 3} \|\overline{\delta}_i\|_2 \le \widetilde{\gamma}_m \max(\|\widehat{d}_1(1\colon n)\|_2, \|b\|_2). \tag{4.17}$$

In Algorithm 1 the final step is to solve the triangular system $Rx = d_1$, where $R = R_1(1\colon n, :)$. The computed solution $\widehat{x}$ satisfies $(\widehat{R}+\Delta R)\widehat{x} = \widehat{d}_1(1\colon n)$, $|\Delta R| \le \gamma_n |\widehat{R}|$ [10, Thm. 8.5]; that is, the rounding errors in the substitution correspond to a further small perturbation of $\widehat{R}$.

We now consider the forward error of the computed solution $\widehat{x}$. First, let z_1 be the solution of the perturbed ILS problem with data

$$A + \Delta A := \begin{bmatrix} A_1 + \Delta_1 \\ A_2 + \Delta_2 \end{bmatrix}, \qquad b + \Delta b := \begin{bmatrix} b_1 + \overline{\delta}_1 \\ b_2 + \overline{\delta}_2 \end{bmatrix},$$

for which we know from (4.11) that the exact upper triangular R-factor is $\widehat{R} + \widetilde{\Delta}_3$, where $\widetilde{\Delta}_3 = \Delta_3(1\colon n, :)$. Then, in view of (4.16),

$$(\widehat{R} + \widetilde{\Delta}_3) z_1 = \widehat{d}_1(1\colon n).$$

Write

$$x - \widehat{x} = (x - z_1) + (z_1 - \widehat{x}).$$

Using the bounds on ΔA and Δb in (4.12) and (4.17), we have, from (2.8),

$$\begin{aligned} \frac{\|x - z_1\|_2}{\|x\|_2} \le \widetilde{\gamma}_{mn} \Bigg[& \|M^{-1}A^T\|_2 \|A\|_F \left(\frac{\max(1,\theta)\|b\|_2}{\|A\|_F \|x\|_2} + 1 \right) \\ & + \|M^{-1}\|_2 \|A\|_F^2 \frac{\|r\|_2}{\|A\|_F \|x\|_2} \Bigg] + O(u^2), \end{aligned} \tag{4.18}$$

where

$$\theta = \frac{\|d_1(1\colon n)\|_2}{\|b\|_2}. \tag{4.19}$$

The quantity θ measures the growth in the leading n components of the right-hand side as a result of the transformations that reduce A to triangular form. We now show

that even though θ can be large, it is innocuous. Suppose that $\theta \gg 1$. Note first that, since $\|d_1\|_2^2 + \|b_2\|_2^2 = \|b_1\|_2^2 + \|d_2\|_2^2$, we have $\|d_2\|_2 \approx \|d_1\|_2 \gg \|b_1\|_2$. Note also that $b_1(n+1\colon p)$ is not subjected to hyperbolic rotations and hence $\|d_1(n+1\colon p)\|_2 \le \|b\|_2$. Hence, from (3.2),

$$\|r\|_2^2 \ge |(b-Ax)^T J(b-Ax)| = |\,\|d(n+1\colon p)\|_2^2 - \|d_2\|_2^2\,| \approx \|d_2\|_2^2 \approx \|d_1\|_2^2.$$

Therefore $\theta \lesssim \|r\|_2/\|b\|_2$ and it follows that the first term in (4.18) is no larger than the second, showing that a large θ does not worsen the bound. Therefore (4.18) is essentially the same as (2.8) with $\epsilon = \widetilde{\gamma}_{mn}$.

From standard perturbation theory for square linear systems, the term $\|z_1 - \widehat{x}\|_2/\|x\|_2$ is bounded by

$$\begin{aligned}\phi &= \kappa_2(R)\left(\gamma_n + \widetilde{\gamma}_{qn}\frac{\max(\|R\|_F, \|A_2\|_F)}{\|R\|_2}\right)\\ &= \|R^{-1}\|_2\big(\gamma_n\|R\|_2 + \widetilde{\gamma}_{qn}\max(\|R\|_F, \|A_2\|_F)\big)\\ &\le \widetilde{\gamma}_{qn}\|R^{-1}\|_2\|A\|_F. \qquad (4.20)\end{aligned}$$

Now from the exact arithmetic analogue of (4.11) we have

$$\begin{bmatrix} R \\ A_2 \end{bmatrix} = G \begin{bmatrix} A_1 \\ 0 \end{bmatrix},$$

where G is orthogonal. Postmultiplying by $R^{-1}R^{-T}$ and transposing gives

$$[\,R^{-1} \quad R^{-1}R^{-T}A_2^T\,] = [\,R^{-1}R^{-T}A_1^T \quad 0\,]\,G^T.$$

Recalling that $M = A^TJA = R^TR$, it follows that

$$\begin{aligned}\|R^{-1}\|_2 &\le \|\,[\,R^{-1}R^{-T}A_1^T \quad 0\,]\,\|_2\\ &\le \|R^{-1}R^{-T}\,[\,A_1^T \quad A_2^T\,]\,\|_2\\ &= \|M^{-1}A^T\|_2.\end{aligned}$$

Hence

$$\phi \le \widetilde{\gamma}_{qn}\|M^{-1}A^T\|_2\|A\|_F,$$

which is smaller than the first term in (4.18). Our overall conclusion is that $\|x - \widehat{x}\|_2/\|x\|_2$ has an upper bound no larger than (2.8) with $\epsilon = \widetilde{\gamma}_{mn}$.

Recall that a method for solving the ILS problem is forward stable if it produces a computed solution with forward error similar to that for a backward stable method. If we make the reasonable assumption that the perturbation bound (2.8) is approximately attainable, then our rounding error analysis has shown that the hyperbolic QR factorization method for solving the ILS problem is forward stable.

It is unclear whether the hyperbolic QR factorization method is mixed backward–forward stable, or even backward stable. It is an open problem to determine a computable formula for the backward error of an arbitrary approximate solution to the ILS problem, and without such a formula it is difficult to test numerically for backward instability.

TABLE 5.1

Errors $\|x - \widehat{x}\|_2/\|x\|_2$ *for the three methods. In every case* $\theta \le 0.78$, $\|r\|_2/(\|A\|_2\|x\|_2) \approx u$, $\|Q\|_2 = 1.73$, $\|Q^T A - [R^T\ 0]^T\|_2/\|A\|_2 \approx u$, *and* $\|Q^T JQ - J\|_2 \le 10u$.

κ	Hyperbolic QR	QR-Cholesky	Normal equations	ψu
10^2	4.9e-15	4.0e-15	2.0e-13	2.1e-14
10^6	3.0e-11	1.7e-11	2.6e-5	1.9e-10
10^{10}	9.7e-8	1.5e-7	2.4e0	1.3e-6
10^{12}	4.8e-4	9.8e-3	6.4e0	1.4e-2

TABLE 5.2

Errors $\|x - \widehat{x}\|_2/\|x\|_2$ *for the three methods. In every case* $\|r\|_2/(\|A\|_2\|x\|_2) \approx 10^{-1}$ *and* $\|Q^T JQ - J\|_2 \approx$ *the error for hyperbolic QR.*

μ	Hyperbolic QR	QR-Cholesky	Normal equations	ψu	θ	$\|Q\|_2$	$\frac{\|Q^T A - [R^T\ 0]^T\|_2}{\|A\|_2}$
10	4.4e-13	1.9e-13	1.6e-12	2.0e-12	3.1e1	4.3e1	1.7e-14
10^2	1.6e-12	3.1e-12	1.5e-12	1.1e-11	1.1e2	1.5e2	2.8e-13
10^3	1.0e-10	5.4e-11	9.7e-11	7.8e-10	8.8e2	1.2e3	5.6e-12
10^4	2.2e-5	6.7e-5	5.0e-5	3.0e-4	5.6e5	8.0e5	3.2e-9
10^5	1.3e-1	5.4e-2	3.3e-2	3.1e-1	1.8e7	2.6e7	2.7e-7

5. Numerical experiments. We have carried out MATLAB experiments to compare the forward errors $\|x - \widehat{x}\|_2/\|x\|_2$ from Algorithm 1, the normal equations method (which forms and solves (1.3)), and the QR-Cholesky method. We approximated the exact solution by forming and solving the normal equations in 100-digit arithmetic using MATLAB's Symbolic Math Toolbox. We report results with $m = 16$, $n = 8$, and $p = 10$.

We formed the first class of test problems as

$$A = \begin{matrix} p \\ q \end{matrix} \begin{bmatrix} Q_1 DU \\ \frac{1}{2} Q_2 DU \end{bmatrix}, \tag{5.1}$$

where U, Q_1, and Q_2 are random orthogonal matrices and D is diagonal with diagonal elements distributed exponentially from κ^{-1} to 1. We have $A^T JA = (3/4)U^T D^2 U$, so A satisfies (1.4). The solution x is chosen from the random $N(0,1)$ distribution and $b := Ax$. Table 5.1 shows some results. In the table ψu is the first order term in (2.10) with $\epsilon = u$, $\mathbf{A} = A$, and $\mathbf{b} = b$; thus ψu is a first order bound for the forward error for a backward stable method. Recall that θ is defined in (4.19). For the statistics shown in the caption we explicitly formed Q by accumulating all the orthogonal and hyperbolic transformations.

In the second set of tests we generated A as in (5.1) and then premultiplied it by a random J-orthogonal matrix that is the product of 5 random hyperbolic rotations of norm approximately μ; this gives a Q factor of norm depending on μ in the hyperbolic QR factorization. Then we defined b as the right singular vector corresponding to the largest singular value of Q^T, which tends to make θ in (4.19) large. Results are shown in Table 5.2.

In all the tests the relative difference between ψu and the first order term from (2.8) was at most 0.1.

Three main conclusions can be drawn from the results shown. First, as expected, the normal equations method is not forward stable. Second, Algorithm 1 behaves in

a forward stable way in these tests and is just as accurate as the backward stable QR-Cholesky method, even when θ and $\|Q\|_2$ are large. The latter behavior adequately summarizes more extensive experiments that we have carried out. Third, the last column of Table 5.2 is consistent with the fact that we have proved our algorithm for computing the hyperbolic QR factorization to be mixed backward–forward stable and only conditionally backward stable.

Acknowledgments. Tony Cox contributed to the analysis in section 2. The first author would like to thank the Department of Mathematics, The University of Manchester, UK, where this work was carried out. He would also like to thank his hosts Professor Nick Higham and Dr. Françoise Tisseur for their great hospitality.

REFERENCES

[1] A. W. Bojanczyk, R. P. Brent, P. Van Dooren, and F. R. de Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 210–221.

[2] A. W. Bojanczyk and A. O. Steinhardt, *Stability analysis of a Householder-based algorithm for downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1255–1265.

[3] S. Chandrasekaran, M. Gu, and A. H. Sayed, *A stable and efficient algorithm for the indefinite linear least-squares problem*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 354–362.

[4] A. J. Cox and N. J. Higham, *Accuracy and stability of the null space method for solving the equality constrained least squares problem*, BIT, 39 (1999), pp. 34–50.

[5] L. Eldén and H. Park, *Perturbation analysis for block downdating of a Cholesky decomposition*, Numer. Math., 68 (1994), pp. 457–467.

[6] L. Eldén and H. Park, *Perturbation and error analyses for block downdating of a Cholesky decomposition*, BIT, 36 (1996), pp. 247–263.

[7] G. H. Golub, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.

[8] B. Hassibi, A. H. Sayed, and T. Kailath, *Linear estimation in Krein spaces—Part* I: *Theory*, IEEE Trans. Automat. Control, 41 (1996), pp. 18–33.

[9] H. V. Henderson and S. R. Searle, *The vec-permutation matrix, the vec operator and Kronecker products: A review*, Linear Multilinear Algebra, 9 (1981), pp. 271–288.

[10] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, PA, 2002.

[11] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[12] C.-T. Pan and R. J. Plemmons, *Least squares modifications with inverse factorizations: Parallel implications*, J. Comput. Appl. Math., 27 (1989), pp. 109–127.

[13] H. Park and L. Eldén, *Stability analysis and fast algorithms for triangularization of Toeplitz matrices*, Numer. Math., 76 (1997), pp. 383–402.

[14] A. H. Sayed, B. Hassibi, and T. Kailath, *Inertia properties of indefinite quadratic forms*, IEEE Signal Process. Lett., 3 (1996), pp. 57–59.

[15] G. W. Stewart, *On the stability of sequential updates and downdates*, IEEE Trans. Signal Process., 43 (1995), pp. 2642–2648.

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

[17] M. Stewart and G. W. Stewart, *On hyperbolic triangularization: Stability and pivoting*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 847–860.

[18] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, PA, 1991.