Section 2.2 Condition numbers

* Overview of Error Analysis:
  - Error analysis is important subject of numerical analysis
  - Given a problem f and an algorithm $\tilde{f}$ with an input x, the absolute error is $\|\tilde{f}(x) - f(x)\|$ and relative error is $\|\tilde{f}(x) - f(x)\| / \|f(x)\|$.
  - We would like the solution to be ~~accurate~~ good, i.e., with small errors.

* Condition number $\Rightarrow$ a measure of sensitivity of a problem.
  - Consider a system $Ax = b$. (A is nonsingular, b is nonzero)

$\Rightarrow$ has unique solution $x \neq 0$.

How if the system is perturbed? That is,

$$A\hat{x} = b + \underbrace{\delta b}_{\text{small vector / noise}}$$

Thm: Let A be nonsingular, and consider $Ax = b$ and the perturbed linear system $A\hat{x} = b + \delta b$. Then

$$\frac{\|x - \hat{x}\|_2}{\|x\|_2} \leq \|A\|_2 \|A^{-1}\|_2 \frac{\|\delta b\|_2}{\|b\|_2}.$$

Pf: Subtracting $Ax = b$ from $A\hat{x} = b + \delta b$,

$$A(\hat{x} - x) = \delta b.$$

$$\Rightarrow \hat{x} - x = A^{-1}\delta b.$$

(*) $$\|\hat{x} - x\|_2 = \|A^{-1}\delta b\|_2 \leq \|A^{-1}\| \, \|\delta b\|.$$

Since $b = Ax$,

(**) $$\|b\|_2 = \|Ax\|_2 \leq \|A\| \, \|x\|_2 \Rightarrow \frac{1}{\|x\|_2} \leq \frac{\|A\|}{\|b\|}$$

(*) and (**), $$\frac{\|\hat{x} - x\|_2}{\|x\|_2} \leq \|A\| \, \|A^{-1}\| \frac{\|\delta b\|_2}{\|b\|_2}$$

Def: $\kappa(A) = \|A\| \, \|A^{-1}\|$ is called the condition number of A

$$\Rightarrow \quad \frac{\|\delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\delta b\|}{\|b\|} \quad \text{, where } \delta x = \hat{x} - x.$$

$\Rightarrow$ $\kappa(A)$ is a factor that determines how the magnitude of δx depends on the magnitude of a perturbation of b.

- If $\kappa(A)$ is small, the small values of $\frac{\|\delta b\|}{\|b\|}$ ~~implies~~ imply small values of $\frac{\|\delta x\|}{\|x\|}$.
  $\Rightarrow$ the solution to $Ax = b$ is not sensitive to small changes in b. (A is well-conditioned).
- If $\kappa(A)$ is large, small value of $\frac{\|\delta b\|}{\|b\|}$ does not guarantee that $\frac{\|\delta x\|}{\|x\|}$ will be small.
  (A is ill-conditioned).

Prop: $\kappa(A) \geq 1$.

Pf: $$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \, \|A^{-1}\| = \kappa(A).$$

$\Rightarrow$ the best (smallest) possible condition number is 1.

* Geometric interpretation of the condition number

By definition,

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

$$\Rightarrow \|A^{-1}\| = \max_{x \neq 0} \frac{\|A^{-1}x\|_2}{\|x\|_2} = \max_{y \neq 0} \frac{\|y\|_2}{\|Ay\|_2} \quad \text{, where } y = A^{-1}x.$$

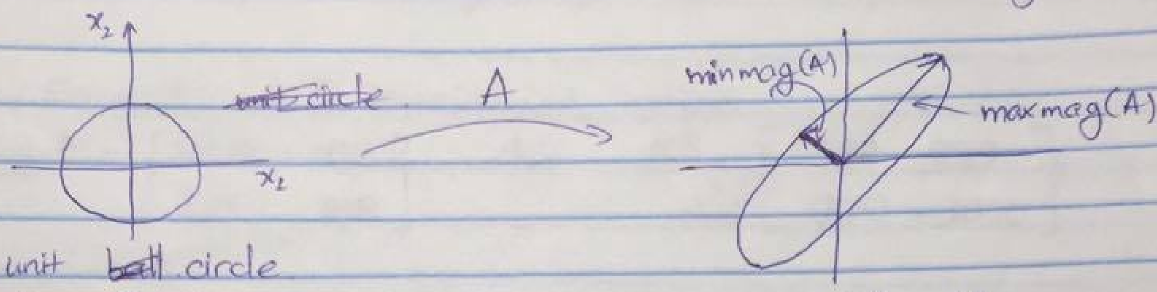
$$= \frac{1}{\min\limits_{y \neq 0} \frac{\|Ay\|_2}{\|y\|_2}}$$

$$= \frac{1}{\min\limits_{\|y\|_2 = 1} \|Ay\|_2}$$

maximum magnification

$$\kappa(A) = \frac{\max\limits_{\|x\|_2=1} \|Ax\|_2}{\min\limits_{\|x\|_2=1} \|Ax\|_2} = \frac{\text{maxmag}(A)}{\text{minmag}(A)}$$

minimum magnification



unit ~~ball~~ circle

$\{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$.

$\{Ax : \|x\|_2 = 1\}$.

A matrix A is ill-conditioned if $\kappa(A)$ is large.

68

~~Example of~~

Def: $\kappa_p(A) = \|A\|_p \, \|A^{-1}\|_p$ for $1 \le p \le \infty$.

. Example of ill-conditioned matrix.

E.g. $A = \begin{bmatrix} 1000 & 999 \\ 999 & 998 \end{bmatrix}$, then $A^{-1} \underset{\text{check!}}{=} \begin{bmatrix} -998 & 999 \\ 999 & -1000 \end{bmatrix}$

$$\Rightarrow \kappa_\infty(A) = \|A\|_\infty \, \|A^{-1}\|_\infty = (1999).(1999) = 1999^2$$

$$\kappa_1(A) = \|A\|_1 \, \|A^{-1}\|_1 = 1999^2.$$

$$= 3.996 \times 10^6.$$

* $\kappa_2(A) = ?$ (Exercise).

E.g. A famous example is Hilbert matrix, defined by

$$h_{ij} = \frac{1}{i+j-1}, \quad 1 \le i, j \le n.$$

The matrix is ill-conditioned for even quite small n.

For $n \geq 4$, we have

$$H_4 = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{bmatrix}$$

then $\kappa_2(H_4) \approx 1.6 \times 10^4$ (check! using Matlab).

$n = 8$, $\kappa_2(H_8) \approx 1.5 \times 10^{10}$.

(69)

Thm: Given a nonsingular A and a perturbation δA, consider $Ax = b$ and $(A+\delta A)(x+\delta x) = b$.
Suppose $\frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)}$ (and thus $A+\delta A$ is nonsingular)
Then

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)\frac{\|\delta A\|}{\|A\|}}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}.$$

Section 2.3 Perturbing the coefficient matrix.

Given $Ax = b$ with A nonsingular, consider a perturbed linear system $(\underbrace{A + \delta A}_{\text{small matrix}})(\underbrace{x + \delta x}_{\hat{x}}) = b$.

- We first consider a result that guarantees that this perturbed linear system has a unique sol.

Thm: If A is nonsingular and $\|A^{-1}\delta A\| < 1$, then $A + \delta A$ is nonsingular.

Thm: Let A be nonsingular, if $\|A^{-1}\| \, \|\delta A\| < 1$, then $A + \delta A$ is nonsingular.

Remark: $\|A^{-1}\| \, \|\delta A\| < 1 \iff \frac{\|\delta A\|}{\|A\|} < \frac{1}{\kappa(A)}$.

Thm: If A is nonsingular, and let $b \neq 0$. Then

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \kappa(A) \frac{\|\delta A\|}{\|A\|}.$$

Pf:

$$(A + \delta A)(\hat{x}) = b.$$

$$A\hat{x} + \delta A\hat{x} = b.$$

$$\underbrace{Ax}_{b} + A\delta x + \delta A\hat{x} = b.$$

$$\Rightarrow \quad A\delta x + \delta A\hat{x} = 0.$$

$$A\delta x = -\delta A\hat{x}.$$

$$\Rightarrow \quad \delta x = -A^{-1}\delta A\hat{x}.$$

$$\|\delta x\|_2 \leq \|A^{-1}\| \, \|\delta A\| \, \|\hat{x}\|_2.$$

$$\Rightarrow \quad \frac{\|\delta x\|_2}{\|\hat{x}\|_2} \leq \|A^{-1}\| \, \|\delta A\| = \kappa(A) \frac{\|\delta A\|}{\|A\|}.$$

7L

Theorem 2.3.3 (page 134)

Given a nonsingular matrix A and a perturbation δA, consider the linear systems $Ax = b$ and $(A + \delta A)(x + \delta x) = b$, where $b \neq 0$. Then

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \kappa(A)\frac{\|\delta A\|}{\|A\|}.$$

Proof. Subtracting $Ax = b$ from $(A + \delta A)(x + \delta x) = b$ gives

$$\delta A\,(x + \delta x) = -A\,\delta x.$$

Therefore

$$\begin{aligned} \delta x &= -A^{-1}\delta A\,(x + \delta x) \\ \Rightarrow\quad & \|\delta x\| \leq \|A^{-1}\|\,\|\delta A\|\,\|x + \delta x\| \\ \Rightarrow\quad & \frac{\|\delta x\|}{\|x + \delta x\|} \leq \kappa(A)\frac{\|\delta A\|}{\|A\|} \end{aligned}$$

NOTE. In the textbook, $x + \delta x$ is denoted by $\hat{x}$. Also note that this result does not require that $A + \delta A$ is nonsingular or that δA is small. Since δx is in both the numerator and denominator of $\frac{\|\delta x\|}{\|x + \delta x\|}$, it is possible to bound this even if δx is not uniquely determined.

APPLICATION OF THEOREM 2.3.6

Consider solving $H_n x = b$, where H_n is the $n \times n$ Hilbert matrix, defined by

$$h_{ij} = \frac{1}{i + j - 1}.$$

Suppose that b is known exactly, but that H_n must be rounded to 7 significant decimal digits when stored in the computer, so that the actual system solved is

$$(H_n + \delta H_n)(x + \delta x) = b,$$

which is some perturbation of the exact linear system $H_n x = b$ with $\frac{\|\delta H_n\|}{\|H_n\|} \approx 10^{-7}$.

72

Here are some condition numbers (with respect to the 2-norm):

n	$\kappa(H_n)$
3	5.2×10^{2}
5	4.8×10^{5}
7	4.8×10^{8}
9	4.9×10^{11}

(Note that $\|H_n^{-1}\| \approx \kappa(H_n)$.) So for $n=3$ or 5, $\|H_n^{-1}\|\|\delta H_n\| << 1$ and (ignoring the denominator term in the bound in Theorem 2.3.6)

$$\frac{\|\delta x\|}{\|x\|} \le \kappa(H_n)\frac{\|\delta H_n\|}{\|H_n\|}.$$

But for $n \ge 7$, $\|H_n^{-1}\|\|\delta H_n\| > 1$ and thus the bound in Theorem 2.3.6 isn't even applicable. In this case, the result of Theorem 2.3.3 above applies, but note that since $\|H_n^{-1}\|\|\delta H_n\| = \kappa(H_n)\frac{\|\delta H_n\|}{\|H_n\|} > 1$, this result simply says that

$$\frac{\|\delta x\|}{\|x+\delta x\|} \le \{\text{something} > 1\}$$

which gives no information about how small δx might be.

****

SIMULTANEOUS perturbation of both A and b -- see Theorem 2.3.8 and Theorem 2.3.9 on page 135. These are extensions of the results of Theorem 2.3.3 and Theorem 2.3.6, respectively.

Theorem 2.3.8

Given a nonsingular matrix A and perturbations δA and δb, consider the linear systems $Ax=b$ and $(A+\delta A)(x+\delta x)=b+\delta b$, where $x+\delta x \ne 0$ and $b+\delta b \ne 0$. Then

$$\frac{\|\delta x\|}{\|x+\delta x\|} \le \kappa(A)\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b+\delta b\|}+\frac{\|\delta A\|}{\|A\|}\frac{\|\delta b\|}{\|b+\delta b\|}\right).$$

73

SECTION 2.4 A POSTERIORI ERROR ANALYSIS USING THE RESIDUAL

TYPES OF ERROR BOUNDS

<u>a priori</u> -- can be evaluated without solving for the solution of the problem.

<u>a posteriori</u> -- the bound uses information about the computed solution or information obtained during the computation.

All bounds given previously are *a priori* bounds: they involve A, δA, b, δb or A^{-1} but not a computed solution $\hat{x}$ to $Ax = b$.

A simple EXAMPLE of an *a posteriori* bound: given $Ax = b$, let $\hat{x}$ denote a computed solution (obtained by any means). Define the <u>residual vector</u>

$$\hat{r} = b - A\hat{x}.$$

NOTES.

(i) $\hat{r} = 0$ if and only if $\hat{x} = x$, where x is the exact solution of $Ax = b$.

(ii) If $\hat{r}$ is small, then $\hat{x}$ is the solution of a linear system that is close to $Ax = b$ because if we define $\delta b = -\hat{r}$, then $\hat{x}$ is the exact solution of $A\hat{x} = b + \delta b$.

(iii) However, it is unfortunately the case that even if $\hat{r}$ is small, $\hat{x}$ is not necessarily close to the exact solution x. The condition number of A must also be taken into account. Restating Theorem 2.2.4 for the case that $\delta b = -\hat{r}$, we obtain the following result.

<u>Theorem 2.4.1</u> (page 137)

Let A be nonsingular, $b \neq 0$, and let $\hat{x}$ be any vector (for example, any computed approximation to x). Let $\hat{r} = b - A\hat{x}$. Then

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \kappa(A) \frac{\|\hat{r}\|}{\|b\|}.$$

<u>Proof.</u>

This follows from Theorem 2.2.4 with $\delta b = -\hat{r}$ since $A\hat{x} = b - \hat{r}$.

This is an *a posteriori* bound since it depends on the computed solution $\hat{x}$.

74

INTERPRETATION OF THIS RESULT

If A is well conditioned (that is, $\kappa(A)$ is small) and if $\frac{\|\hat{r}\|}{\|b\|} = \frac{\|b - A\hat{x}\|}{\|b\|}$ is small, then $\hat{x} \approx x$. But if A is ill-conditioned, then a small residual does not necessarily imply that $\hat{x} \approx x$.

SECTION 3.1 THE DISCRETE LEAST-SQUARES (ℓ_2) PROBLEM

THE PROBLEM

Given a set of discrete data $\{(t_i, y_i), 1 \le i \le n\}$ and a set of basis functions $\{\varphi_1(t), \ldots, \varphi_m(t)\}$, find the best least-squares approximation of the form

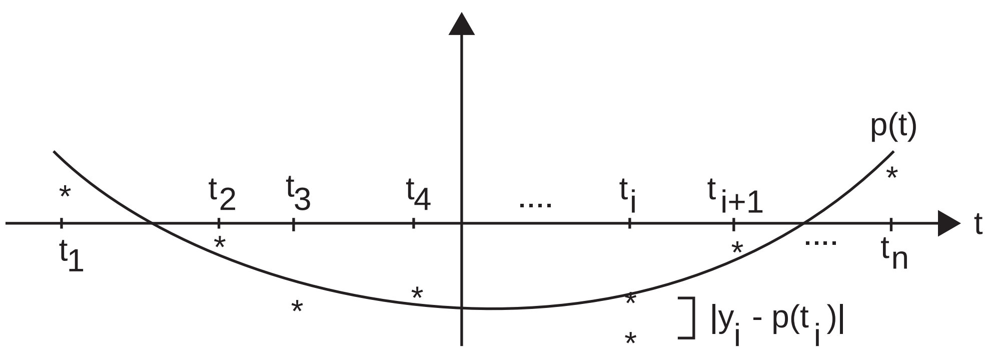
$$p(t) = x_1\varphi_1(t) + x_2\varphi_2(t) + \cdots + x_m\varphi_m(t)$$

to the given data. That is, determine values $x_1, x_2, \ldots, x_m$ so as to

$$\min_{\{x_1, \ldots, x_m\}} \left(\sum_{i=1}^{n} (y_i - p(t_i))^2 \right)^{1/2}$$

or equivalently

$$\min_{\{x_1, \ldots, x_m\}} \left(\sum_{i=1}^{n} (y_i - p(t_i))^2 \right).$$



ONE APPROACH: set the partial derivatives of

$$S(x_1, \ldots, x_m) = \sum_{i=1}^{n} \left(y_i - x_1\varphi_1(t_i) - x_2\varphi_2(t_i) - \cdots - x_m\varphi_m(t_i)\right)^2$$

with respect to each of $x_1, x_2, \ldots. x_m$ to 0. This gives a system of *m* linear equations in *m* unknowns, which are called the <u>normal equations</u>.

EXAMPLE

Consider the case $m = 3$. Then

$$S(x_1, x_2, x_3) = \sum_{i=1}^{n} \left(y_i - x_1 \varphi_1(t_i) - x_2 \varphi_2(t_i) - x_3 \varphi_3(t_i) \right)^2$$

Setting the partial derivatives to 0 gives

$$\frac{\partial S}{\partial x_1} = 2 \sum_{i=1}^{n} \left(y_i - x_1 \varphi_1(t_i) - x_2 \varphi_2(t_i) - x_3 \varphi_3(t_i) \right) \left(-\varphi_1(t_i) \right) = 0$$
$$\frac{\partial S}{\partial x_2} = 2 \sum_{i=1}^{n} \left(y_i - x_1 \varphi_1(t_i) - x_2 \varphi_2(t_i) - x_3 \varphi_3(t_i) \right) \left(-\varphi_2(t_i) \right) = 0$$
$$\frac{\partial S}{\partial x_3} = 2 \sum_{i=1}^{n} \left(y_i - x_1 \varphi_1(t_i) - x_2 \varphi_2(t_i) - x_3 \varphi_3(t_i) \right) \left(-\varphi_3(t_i) \right) = 0$$

which can be rewritten as

$$\begin{bmatrix} \sum_{i=1}^{n} (\varphi_1(t_i))^2 & \sum_{i=1}^{n} \varphi_1(t_i)\varphi_2(t_i) & \sum_{i=1}^{n} \varphi_1(t_i)\varphi_3(t_i) \\ \sum_{i=1}^{n} \varphi_1(t_i)\varphi_2(t_i) & \sum_{i=1}^{n} (\varphi_2(t_i))^2 & \sum_{i=1}^{n} \varphi_2(t_i)\varphi_3(t_i) \\ \sum_{i=1}^{n} \varphi_1(t_i)\varphi_3(t_i) & \sum_{i=1}^{n} \varphi_2(t_i)\varphi_3(t_i) & \sum_{i=1}^{n} (\varphi_3(t_i))^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \varphi_1(t_i) \\ \sum_{i=1}^{n} y_i \varphi_2(t_i) \\ \sum_{i=1}^{n} y_i \varphi_3(t_i) \end{bmatrix}.$$

Such a system can be solved by Gaussian elimination -- or in fact by the Cholesky algorithm, since it can be shown that the coefficient matrix of the normal equations is positive definite.

A numerically better approach -- ORTHOGONALIZATION METHODS

The ℓ_2 problem

$$\min_{\{x_1, \ldots, x_m\}} \left(\sum_{i=1}^{n} (y_i - p(t_i))^2 \right)$$

can be restated (in terms of vectors and matrices) as

$$\min_x \| y - Ax \|_2^2$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad , \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

and

$$A = \begin{bmatrix} \varphi_1(t_1) & \varphi_2(t_1) & \cdots & \varphi_m(t_1) \\ \varphi_1(t_2) & \varphi_2(t_2) & \cdots & \varphi_m(t_2) \\ \vdots & \vdots & & \vdots \\ \varphi_1(t_n) & \varphi_2(t_n) & \cdots & \varphi_m(t_n) \end{bmatrix}.$$

Usually $n >> m$.

NOTE: with this notation, the <u>normal equations</u> are

$$A^T Ax = A^T y \, .$$

GENERAL ℓ_2 PROBLEM (from a vector/matrix point-of-view)

Given

$$A \in \Re^{n \times m} \quad \text{with} \quad n \geq m$$

$$y \in \Re^n$$

determine $x \in \Re^m$ such that

$$\|y - Ax\|_2^2$$

is minimized. Such a solution vector x is also called the best least-squares solution to the over-determined (if $n > m$) linear system $Ax = y$.

SECTION 3.2 ORTHOGONAL MATRICES, ROTATORS AND REFLECTORS

A real matrix Q of order n is orthogonal if $QQ^T = I$ (or, equivalently, if $Q^TQ = I$ or if $Q^T = Q^{-1}$).

The row vectors and the column vectors of an orthogonal matrix Q form an orthonormal set: that is,

$$q_i^T q_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

EXAMPLE

$$Q = \begin{bmatrix} 2/3 & -2/3 & 1/3 \\ 2/3 & 1/3 & -2/3 \\ 1/3 & 2/3 & 2/3 \end{bmatrix}$$

PROPERTIES OF ORTHOGONAL MATRICES

1. $\|Qx\|_2 = \|x\|_2$ for all vectors x.

Proof.

$$\|Qx\|_2 = \sqrt{(Qx)^T(Qx)} = \sqrt{x^TQ^TQx} = \sqrt{x^Tx} = \|x\|_2$$

2. Orthogonal matrices preserve angles. The angle θ between 2 nonzero vectors x and y is such that

$$\cos\theta = \frac{y^Tx}{\|x\|_2\|y\|_2}.$$

The angle between Qx and Qy is the same since

$$\frac{(Qy)^T(Qx)}{\|Qx\|_2\|Qy\|_2} = \frac{y^TQ^TQx}{\|x\|_2\|y\|_2} = \frac{y^Tx}{\|x\|_2\|y\|_2}.$$

3. The product of orthogonal matrices is orthogonal: if all Q_i are orthogonal, then

$$(Q_1 Q_2 \cdots Q_k)(Q_1 Q_2 \cdots Q_k)^T = Q_1 Q_2 \cdots Q_k Q_k^T \cdots Q_2^T Q_1^T = I \, .$$

ROTATORS (or PLANE ROTATION MATRICES)

Read pages 188-190 of the text: discusses these matrices in $\Re^2$ from a geometrical point-of-view.

General form (in $\Re^n$):

$$Q_{ji} = \begin{array}{cc} \begin{bmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & c & & & & -s & & & \\ & & & & 1 & & & & & & \\ & & & & & \ddots & & & & & \\ & & & & & & 1 & & & & \\ & & & s & & & & c & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{bmatrix} & \begin{matrix} \\ \\ \\ \leftarrow i \\ \\ \\ \\ \leftarrow j \\ \\ \\ \\ \end{matrix} \\ \begin{matrix} \qquad\qquad \uparrow & \qquad\qquad\quad \uparrow \\ \qquad\qquad i & \qquad\qquad\quad j \end{matrix} & \end{array}$$

where $c^2 + s^2 = 1$.

Without loss of generality, one can consider $c = \cos\theta$ and $s = \sin\theta$ for some angle θ.

NOTE: $Q_{ji} Q_{ji}^T = I$, which implies that the matrix Q_{ji} is orthogonal.

NOTE the effect on a vector x of multiplication by Q_{ji}^T:

$$Q_{ji}^T x = \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ c\,x_i + s\,x_j \\ x_{i+1} \\ \vdots \\ x_{j-1} \\ -s\,x_i + c\,x_j \\ x_{j+1} \\ \vdots \\ x_n \end{bmatrix} \begin{array}{l} \\ \\ \\ \leftarrow i-th \text{ entry} \\ \\ \\ \\ \leftarrow j-th \text{ entry} \\ \\ \\ \\ \end{array}$$

A frequent goal of a numerical algorithm is to introduce zeros into an array. Multiplication by Q_{ji}^T can accomplish this if c and s are appropriately chosen.

EXAMPLE

Given any vector x, to make the j^{th} entry of $Q_{ji}^T x$ equal to 0 requires that

$$-s\,x_i + c\,x_j = 0$$

$$\Rightarrow \quad c = \frac{x_i}{\sqrt{x_i^2 + x_j^2}} \text{ and } s = \frac{x_j}{\sqrt{x_i^2 + x_j^2}}, \quad \text{since } c^2 + s^2 = 1.$$

NOTE: the effect on A of forming the product $Q_{ji}^T A$ is similar:

-- rows $1, \ldots, i-1, i+1, \ldots. j-1, j+1, \ldots, n$ of A are unchanged

-- rows i and j of $Q_{ji}^T A$ are linear combinations of rows i and j of A.

GEOMETRICAL INTERPRETATION of the multiplication by a rotator: see the middle of page 193 of the text.

THEOREM 3.2.20 (page 193)

Let $A \in \Re^{n\times n}$. Then there exists an orthogonal matrix Q such that $Q^T A = R$, where R is upper triangular. (Equivalently, $A = QR$.)

<u>Sketch of a proof</u>.

Let Q_{21} be such that

$$Q_{21}^T \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ a_{31} \\ \vdots \\ a_{n1} \end{bmatrix}.$$

Similarly, let Q_{31}^T create a 0 in the (3, 1) position of $Q_{21}^T A$, so that

$$Q_{31}^T Q_{21}^T \begin{bmatrix} * \\ 0 \\ a_{31} \\ a_{41} \\ \vdots \\ a_{n1} \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ 0 \\ a_{41} \\ \vdots \\ a_{n1} \end{bmatrix}.$$

Thus one can choose rotators $Q_{21}, Q_{31}, Q_{41}, \ldots, Q_{n1}$ so that

$$Q_{n1}^T Q_{n-1,1}^T \cdots Q_{21}^T A = \left[\begin{array}{c|c} * & \\ 0 & \\ 0 & \\ 0 & * \\ \vdots & \\ 0 & \end{array}\right].$$

Now choose Q_{32}^T to zero out the (3, 2) entry; then Q_{42}^T to zero out the (4, 2) entry, and so on. Thus rotators $Q_{32}, Q_{42}, \ldots, Q_{n2}$ can be chosen so that

$$(Q_{n2}^T Q_{n-1,2}^T \cdots Q_{32}^T)(Q_{n1}^T \cdots Q_{21}^T) A = \left[\begin{array}{cc|c} * & * & \\ 0 & * & \\ 0 & 0 & \\ 0 & 0 & * \\ \vdots & \vdots & \\ 0 & 0 & \end{array}\right].$$

Clearly the process can continue, so that

$$(Q_{n,n-1}^T)(Q_{n,n-2}^T Q_{n-1,n-2}^T)\cdots(Q_{n1}^T Q_{n-1,1}^T \cdots Q_{21}^T)A = R$$

is upper triangular. Thus, letting

$$Q^T = Q_{n,n-1}^T Q_{n,n-2}^T Q_{n-1,n-2}^T \cdots Q_{21}^T \quad \Rightarrow \quad Q = Q_{21}Q_{31}\cdots Q_{n,n-2}Q_{n,n-1},$$

we have that

$$Q^T A = R \text{ or } A = QR.$$

REFLECTORS (or HOUSEHOLDER MATRICES)

Definition

A matrix of the form

$$Q = I - 2uu^T, \quad \text{where } u^T u = 1,$$

is a Householder matrix.

EXAMPLE

Case $n = 3$

$$Q = \begin{bmatrix} 1-2u_1^2 & -2u_1u_2 & -2u_1u_3 \\ -2u_1u_2 & 1-2u_2^2 & -2u_2u_3 \\ -2u_1u_3 & -2u_2u_3 & 1-2u_3^2 \end{bmatrix}, \quad \text{where } u_1^2 + u_2^2 + u_3^2 = 1.$$

Another (equivalent) form for a Householder matrix:

$$Q = I - \frac{2vv^T}{v^T v} \quad \text{for any vector } v \neq 0$$

$$= I - 2\left(\frac{v}{\|v\|_2}\right)\left(\frac{v^T}{\|v\|_2}\right)$$

This is equivalent to the previous definition since $u = \dfrac{v}{\|v\|_2}$ is a unit vector.

PROPERTIES OF HOUSEHOLDER MATRICES

1. symmetric: $Q = Q^T$

2. orthogonal

$$\begin{aligned} Q^T Q &= (I - 2uu^T)(I - 2uu^T) \\ &= I - 4uu^T + 4uu^T uu^T \\ &= I \quad \text{since } u^T u = 1 \end{aligned}$$

Thus $Q = Q^T = Q^{-1}$.

THEOREM 3.2.30 (page 196)

Let $x, y \in \Re^n$, $x \neq y$, and $\|x\|_2 = \|y\|_2$. Define $u = \dfrac{x - y}{\|x - y\|_2}$. Then

$$(I - 2uu^T)x = y.$$

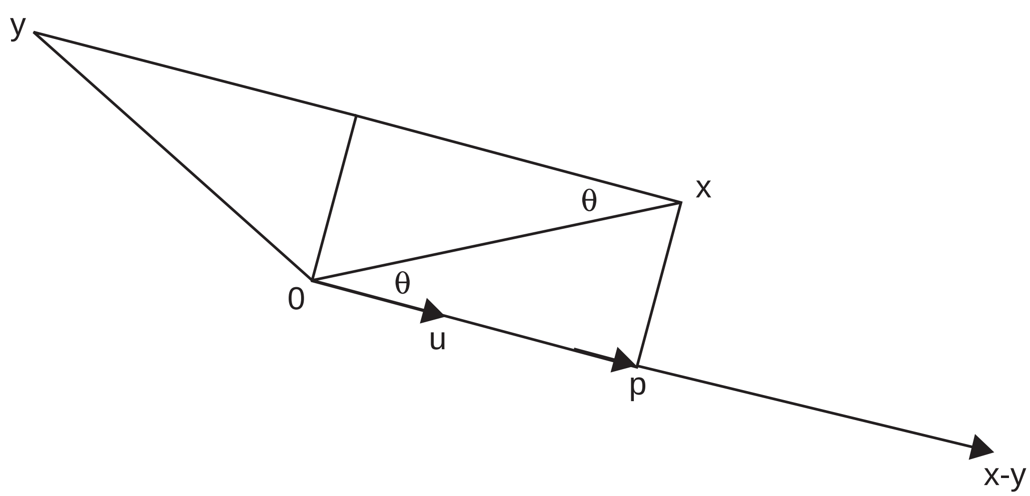
Proof.

$$\begin{aligned} \|x - y\|_2^2 &= (x - y)^T (x - y) \\ &= x^T x - x^T y - y^T x + y^T y \\ &= 2(x^T x - y^T x) \quad \text{since } x^T y = y^T x \text{ and } y^T y = x^T x. \end{aligned}$$

Therefore

$$\begin{aligned} (I - 2uu^T)x &= x - 2\,\frac{x - y}{\|x - y\|_2}\frac{(x - y)^T}{\|x - y\|_2}x \\ &= x - \frac{2(x - y)}{2(x^T x - y^T x)}(x^T x - y^T x) \\ &= y. \end{aligned}$$

GEOMETRICAL INTERPRETATION (page 197)



Let

$$u = \frac{x-y}{\|x-y\|_2}.$$

The vector p is defined by $y = x - 2p$. Now find an expression for p.

We have

$$\cos\theta = \frac{\|p\|_2}{\|x\|_2} = \frac{u^T x}{\|u\|_2 \|x\|_2}$$

$$\Rightarrow \quad \|p\|_2 = u^T x$$

$$\Rightarrow \quad p = u(u^T x)$$

That is, p is a vector of length $u^T x$ in the direction of u.

Therefore

$$\begin{aligned} y &= x - 2u(u^T x) \\ &= (I - 2uu^T)x. \end{aligned}$$

It can be shown that the vector u is uniquely determined up to a $\pm$ sign:

$$u = \pm\frac{x-y}{\|x-y\|_2}.$$

The above theorem can be used to <u>introduce 0's</u> into an array.

COROLLARY

Let $x \neq 0$, x not a scalar multiple of $e_1 = (1, 0, 0, \ldots, 0)^T$.

Let $\sigma = \pm\|x\|_2$, $\quad v = x + \sigma e_1$, and $u = \dfrac{v}{\|v\|_2}$.

Then

$$\begin{aligned}(I - 2uu^T)x &= -\sigma e_1 \\ &= (-\sigma, 0, 0, \ldots, 0)^T.\end{aligned}$$

Proof.

This is just the case of $y = -\sigma e_1$ in Theorem 3.2.30.

<u>Choice of the $\pm$ sign</u>: note that

$$uu^T = \frac{vv^T}{\|v\|_2^2}$$

and

$$\begin{aligned}\|v\|_2^2 &= (x + \sigma e_1)^T (x + \sigma e_1) \\ &= x^T x + \sigma x^T e_1 + \sigma e_1^T x + \sigma^2 e_1^T e_1 \\ &= 2(\sigma^2 + \sigma x_1) \\ &= 2\sigma(\sigma + x_1).\end{aligned}$$

Thus, as one divides by $\|v\|_2^2$ in forming uu^T, in order to avoid cancellation choose $\sigma = sign(x_1)\|x\|_2$.

Another <u>computational aspect</u> of the calculations in the above Corollary: see pages 198-199. The vector x should be normalized (scaled) so as to avoid unnecessary overflows or underflows when computing $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$. If x is replaced by $\dfrac{x}{\|x\|_\infty}$, then no overflows can occur when computing $\|x\|_2$ and any underflows that do occur can be safely set to 0.

ALGORITHM for computing a Householder matrix Q such that $Qx = -\sigma e_1$: see page 199.

EFFICIENT COMPUTATION of the product of a Householder matrix and a vector (or another matrix):

suppose that $a = (a_1, a_2, \ldots, a_n)^T$. Then

$$(I - 2uu^T)a = a - (2u^T a)u,$$

which is just a difference of 2 vectors, since $2u^T a$ is a scalar. Thus, multiplication of a vector by a Householder matrix $Q = I - 2uu^T$

-- requires that only u be stored (and not the $n \times n$ matrix Q)
-- does not require a matrix/vector multiplication, but only the computation of an inner product and a vector subtraction

PAGES 201-202 of the text: a PROOF that there exists a factorization $A = QR$ using Householder matrices. (Note: we saw this result previously in terms of rotators.)

Given any matrix A, pre-multiplication by $n-1$ Householder matrices can reduce A to upper triangular form:

(1) determine a Householder matrix Q_1 so that

$$Q_1 A = \left[\begin{array}{c|c} * & \\ 0 & \\ 0 & * \\ \vdots & \\ 0 & \end{array}\right]$$

(2) determine a Householder matrix $Q_2 = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \hat{Q}_2 \end{array}\right]$ so that

$$Q_2(Q_1 A) = \left[\begin{array}{cc|c} * & * & \\ 0 & * & \\ 0 & 0 & * \\ \vdots & \vdots & \\ 0 & 0 & \end{array}\right]$$

(3) determine a Householder matrix $Q_3 = \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & \hat{Q}_3 \end{array}\right]$ so that

$$Q_3(Q_2Q_1A) = \left[\begin{array}{ccc|c} * & * & * & \\ 0 & * & * & \\ 0 & 0 & * & \\ 0 & 0 & 0 & * \\ \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \end{array}\right]$$

and so on. After $n-1$ such steps, A will be reduced to upper triangular form.

ALGORITHM for computing the QR factorization using reflectors: see page 203.

NOTE: this algorithm could be used to solve a linear system $Ax = b$, where A is $n \times n$ and nonsingular. See the top of page 204.

$$Ax = b \quad \Rightarrow \quad QRx = b \quad \Rightarrow \quad Rx = Q^T b.$$

Since R is upper triangular and nonsingular, multiply $Q^T \times b$ and then solve for x by back substitution.
Cost: $4n^3/3$ flops, whereas Gaussian elimination is only $2n^3/3$ flops.
Note: if rotators are used to compute the QR factorization, the cost is $8n^3/3$ flops.

The following is a uniqueness result for QR factorization.

THEOREM 3.2.46 (page 204)
Let A be an $n \times n$ nonsingular matrix. Then there exist unique $n \times n$ matrices Q (orthogonal) and R (upper triangular with all of its main diagonal entries positive) such that $A = QR$.

Sketch of the proof: note that if $A = \hat{Q}\hat{R}$ is any QR factorization, then there exists a diagonal matrix D with $d_{ii} = \pm 1$ so that $D\hat{R}$ has all of its diagonal entries positive. Thus $A = (\hat{Q}D^{-1})(D\hat{R})$ is the desired QR factorization. Uniqueness of this factorization is obtained from uniqueness of the Cholesky factorization -- see the proof in the text.

(Note: since $R = D\hat{R}$ is upper triangular with positive diagonal entries and $A^T A = (QR)^T (QR) = R^T R$, it follows that R^T is the Cholesky factor L of $A^T A$.)

STABILITY of computations with rotators and reflectors: see pages 205-206.

Results due to Wilkinson: let $\hat{Q}$ denote the computed approximation to any rotator or reflector Q. Then

$$f\ell(\hat{Q}A) = Z(A+E)$$

where Z is some exactly orthogonal matrix (that is close to Q) and $\dfrac{\|E\|_2}{\|A\|_2}$ is small.

That is, the product of the computed approximation to Q and A is exactly equal to the product of some orthogonal matrix and a small perturbation of A.

This extends to products of several rotators or reflectors and a matrix A: for example,

$$\begin{aligned} f\ell(\hat{Q}_1\hat{Q}_2 A) &= Z_2(Z_1(A+E_1)+E_2) \\ &= Z_2 Z_1(A+E), \quad \text{where } E = E_1 + Z_1^T E_2, \end{aligned}$$

Z_1 and Z_2 are exactly orthogonal, and

$$\begin{aligned} \|E\|_2 &\le \|E_1\|_2 + \left\|Z_1^T E_2\right\|_2 \\ &= \|E_1\|_2 + \|E_2\|_2 \quad \text{since } \left\|Z_1^T E_2\right\|_2 = \|E_2\|_2 \text{ because} \\ &\qquad \|QB\|_2 = \max_{\|x\|_2=1} \|QBx\|_2 = \max_{\|x\|_2=1} \|Bx\|_2 = \|B\|_2 \\ &\qquad \text{for any matrix } B \text{ and any orthogonal } Q \\ \Rightarrow &\ \frac{\|E\|_2}{\|A\|_2} \text{ is small.} \end{aligned}$$

This kind of analysis shows that any algorithm involving <u>repeated multiplication by orthogonal matrices is stable</u> -- the computed product of any number of orthogonal matrices and a matrix A is equal to the exact product of some exactly orthogonal matrix and a small perturbation of A. The stability essentially follows from the above fact that

$$\|QB\|_2 = \|B\|_2$$

for any orthogonal matrix Q and any matrix B.

ANOTHER FORM OF THE STABILITY of orthogonal matrices -- see page 116 of Numerical Linear Algebra by Trefethen and Bau:

$$\hat{Q}\hat{R} = A + E, \quad \text{where } \frac{\|E\|}{\|A\|} \text{ is small,}$$

$\hat{R}$ is the computed upper triangular matrix, and $\hat{Q}$ is an exactly orthogonal matrix (that is close to the computed approximation to Q). The computed approximation to Q is not used in this result since it is not exactly orthogonal.

COMPLEX ANALOG OF ORTHOGONAL MATRICES: see pages 206-207.

An $n \times n$ complex matrix U is called <u>unitary</u> if $UU^* = I$ (or, equivalently, if $U^*U = I$ or $U^* = U^{-1}$).

EXAMPLES

$$U = \frac{1}{\sqrt{|a|^2 + |b|^2}} \begin{bmatrix} a & -\bar{b} \\ b & \bar{a} \end{bmatrix},$$

where a and b are complex numbers, is a complex-valued rotator.

If u is a complex vector with n entries and $\|u\|_2 = 1$, then $I - 2uu^*$ is a complex-valued reflector. Note that

$$\|u\|_2 = \sqrt{\sum_{i=1}^{n} u_i \bar{u}_i} = \sqrt{\sum_{i=1}^{n} |u_i|^2} .$$

* <u>Compute QR-Factorization by Householder reflectors.</u>

<u>Overview of the algorithm (on a 3x3 matrix)</u>

$$A = \begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \end{bmatrix} \xrightarrow[Q_1]{} \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & x & x \end{bmatrix} \longrightarrow \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \end{bmatrix} = R$$

Define the HH reflector Q_1 in terms of this vector

define HH reflector Q_2

$$Q_2 Q_1 A = R \Rightarrow A = Q_1^T Q_2^T R.$$

* <u>Reflectors (or Householder matrices)</u>

<u>Def:</u> A matrix of the form

$$Q = I - 2uu^T, \text{ where } u^T u = 1,$$

is a Householder matrix.

Another (equivalent) form for a Householder matrix:

$$Q = I - \frac{2 v v^T}{v^T v} \quad \text{for any vector } v \neq 0.$$

$$= I - 2\left(\frac{v}{\|v\|_2}\right)\left(\frac{v^T}{\|v\|_2}\right).$$

This is equivalent to the previous definition since $u = \frac{v}{\|v\|_2}$ is a unit vector.

* Properties of Householder matrices:

1. Symmetric: $Q = Q^T$.
2. orthogonal.

$$Q^TQ = (I - 2uu^T)(I - 2uu^T)$$
$$= I - 2uu^T - 2uu^T + 4uu^Tuu^T$$
$$= I \quad \text{since } u^Tu = 1.$$

Thm: Let $x, y \in \mathbb{R}^n$ and $x \neq y$, $\|x\|_2 = \|y\|_2$. Define $u = \frac{x-y}{\|x-y\|_2}$. Then

$$(I - 2uu^T)x = y.$$

Cor. Let $x \neq 0$, x not a scalar multiple of $e_1 = (1, 0, \dots, 0)^T$.
Let $\sigma = \pm\|x\|_2$, $v = x + \sigma e_1$, and $u = \frac{v}{\|v\|_2}$.

Then $(I - 2uu^T)x = -\sigma e_1 = (-\sigma, 0, \dots, 0)^T$.

Remark: Choose $\sigma = \text{sign}(x_1)\|x\|_2$.

E.g. Find the QR factorization for

$$A = \begin{bmatrix} 3 & 1 \\ 4 & 2 \end{bmatrix}$$

Determine the HH reflector Q such that for $x = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$,

$$Qx = -\underbrace{\text{sign}(x_1)\|x\|}_{\sigma} e_1 = \begin{bmatrix} 5 \\ 0 \end{bmatrix}.$$

1) Find u: $\quad u = \frac{x + \sigma e_1}{\|x + \sigma e_1\|} = \frac{\begin{bmatrix} 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 5 \\ 0 \end{bmatrix}}{\|v\|_2} = \frac{\begin{bmatrix} 8 \\ 4 \end{bmatrix}}{\sqrt{64 + 16}} = \frac{1}{\sqrt{5}}\begin{bmatrix} 2 \\ 1 \end{bmatrix}.$

2) Find Q:

$$
\begin{aligned}
Q &= I - 2uu^T \\
&= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{2}{5}\begin{bmatrix} 2 \\ 1 \end{bmatrix}\begin{bmatrix} 2 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{2}{5}\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \\
&= \begin{bmatrix} -\frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & \frac{3}{5} \end{bmatrix}
\end{aligned}
$$

Check:

$$
\overset{Q}{\begin{bmatrix} -\frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & \frac{3}{5} \end{bmatrix}}
\overset{A}{\begin{bmatrix} 3 & 1 \\ 4 & 2 \end{bmatrix}}
= \overset{R}{\begin{bmatrix} -5 & -\frac{11}{5} \\ 0 & \frac{2}{5} \end{bmatrix}}
$$

Solution of the least squares problem by QR factorization.

(LSP) $$\min_{x \in \mathbb{R}^n} \|b - Ax\|_2$$

$A = \quad \in \mathbb{R}^{m\times n}$ $\quad x = \quad \in \mathbb{R}^n$ and $b = \quad \in \mathbb{R}^m$

$m > n$.

We have

$$
\begin{aligned}
\|Ax - b\|_2 &= \|QRx - b\|_2 \\
&= \|Q^TQRx - Q^Tb\|_2 \\
&= \|Rx - Q^Tb\|_2
\end{aligned}
$$

$$\Rightarrow \min_{x \in \mathbb{R}^n} \|Ax - b\|_2 = \min_{x \in \mathbb{R}^n} \|Rx - Q^Tb\|_2$$

Partition $R \in \mathbb{R}^{m\times n}$ and $Q^T b$ as follows.

$$R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \boxed{\diagdown} \\ 0 \end{bmatrix}. \quad \text{where } R_1 \in \mathbb{R}^{n\times n} \text{ is upper triangular.}$$

$$Q^T b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad \text{where } b_1 \in \mathbb{R}^n \text{ and } b_2 \in \mathbb{R}^{m-n}.$$

Then

$$\|Ax - b\|_2 = \Bigg\| \underbrace{\begin{bmatrix} R_1 \\ 0 \end{bmatrix}}_{R} x - \underbrace{\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}}_{b} \Bigg\|_2$$

$$= \Bigg\| \begin{bmatrix} R_1 x - b_1 \\ -b_2 \end{bmatrix} \Bigg\|_2.$$

$$= \sqrt{\|R_1 x - b_1\|_2^2 + \underbrace{\|b_2\|_2^2}_{\text{positive constant, independent of } x.}}$$

$$\Rightarrow \min_{x \in \mathbb{R}^n} \|Ax - b\|_2 = \|A\hat{x} - b\|_2 = \|b_2\|_2,$$

where $\hat{x}$ is a solution of $R_1 x = b_1$.

<u>E.g.</u> Consider the LSP

$$\min_{x \in \mathbb{R}^2} \Bigg\| \underbrace{\begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 0 & 1 \end{bmatrix}}_{A} x - \underbrace{\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}}_{b} \Bigg\|_2 .$$

Given the QR factorization.

$$\underbrace{\begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 0 & 1 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \sqrt{2} & 2\sqrt{2} \\ 0 & 1 \\ 0 & 0 \end{bmatrix}}_{R}$$

R_1

1

$$\Rightarrow \|Ax - b\|_2 = \|Rx - Q^Tb\|.$$

$$Q^Tb = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} \boxed{\begin{matrix} -1/\sqrt{2} \\ 1 \end{matrix}} \\ \boxed{-3/\sqrt{2}} \end{bmatrix} \begin{matrix} \rightarrow b_1 \\ \\ b_2 \end{matrix}$$

$$\Rightarrow \min_x \|Ax - b\|_2 = \|A\hat{x} - b\|_2 = \|b_2\|_2 = \frac{3}{\sqrt{2}},$$

where $\hat{x}$ is the solution of $\begin{bmatrix} \sqrt{2} & 2\sqrt{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} \\ 1 \end{bmatrix}.$

$$\Rightarrow \hat{x} = \begin{bmatrix} -5/2 \\ 1 \end{bmatrix}.$$

* Procedure (To solve LSP).

1) Compute a full QR factorization.

$$A = QR$$

2) Partition R and Q^Tb.

$$R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \rightarrow n \times n \qquad Q^Tb = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

3) Solve the upper triangular system.

$$R_1 \hat{x} = b_1.$$

4) $\min_x \|Ax - b\|_2 = \|b_2\|_2 = \|A\hat{x} - b\|_2.$