Chapter 1

Mathematical and Statistical Properties of Least Squares Solutions

> De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus general, de plus exact, ni d'une application plus facile que celui qui consiste à rendre *minimum* la somme de carrés des erreurs.[1]
>
> *Adrien Marie Legendre, Nouvelles méthodes pour la détermination des orbites des comètes. Appendice. Paris, 1805.*

1.1. Introduction

The linear least squares problem is a computational problem of primary importance, which originally arose from the need to fit a linear mathematical model to given observations. In order to reduce the influence of errors in the observations one would then like to use a greater number of measurements than the number of unknown parameters in the model. The resulting problem is to "solve" an **overdetermined** linear system of equations. In matrix terms, given a vector $b \in \mathbf{R}^m$ and a matrix $A \in \mathbf{R}^{m\times n}$, $m > n$, we want to find a vector $x \in \mathbf{R}^n$ such that Ax is the "best" approximation to b.

EXAMPLE 1.1.1. Consider a model described by a scalar function $y(t) = f(x,t)$, where $x \in \mathbf{R}^n$ is a parameter vector to be determined from measurements (y_i, t_i), $i = 1,\dots,m$, $m > n$. In particular, let $f(x,t)$ be *linear* in x:

$$f(x,t) = \sum_{j=1}^{n} x_j \phi_j(t).$$

Then the equations $y_i = \sum_{j=1}^{n} x_j\phi_j(t_i)$, $i = 1,\dots,m$ form an overdetermined linear system $Ax = b$, where $a_{ij} = \phi_j(t_i)$ and $b_i = y_i$. ▮

There are many possible ways of defining the "best" solution. A choice which can often be motivated for statistical reasons (see below) and which also leads to a simple computational problem is to let x be a solution to the minimization problem

$$\min_x \|Ax - b\|_2, \quad A \in \mathbf{R}^{m\times n}, \quad b \in \mathbf{R}^m, \tag{1.1.1}$$

[1] Of all the principles that can be proposed, I think there is none more general, more exact, and more easy of application, than that which consists of rendering the sum of the squares of the errors a minimum.

where $\|\cdot\|_2$ denotes the Euclidean vector norm. We call this a **linear least squares problem** and x a linear least squares solution of the system $Ax = b$. We refer to $r = b - Ax$ as the residual vector. A least squares solution minimizes $\|r\|_2^2 = \sum_{i=1}^m r_i^2$, i.e., the sum of the squared residuals. If rank $(A) < n$, then the solution x to (1.1) is not unique. However, among all least squares solutions there is a unique solution which minimizes $\|x\|_2$; see Theorem 1.2.10.

1.1.1. Historical remarks.

Laplace in 1799 used the principle of minimizing the sum of the absolute errors $\sum_{i=1}^m |r_i|$, with the added condition that the sum of the errors be equal to zero; see Goldstine [363, 1977]. He showed that the solution x must then satisfy exactly n out of the m equations. Gauss argued that since, by the principles of probability, greater or smaller errors are equally possible in all equations, it is evident that a solution which satisfies precisely n equations must be regarded as less consistent with the laws of probability. He was then led to the principle of least squares. The algebraic procedure of the method of least squares was first published by Legendre [523, 1805]. It was justified as a statistical procedure by Gauss [320, 1809], where he (much to the annoyance of Legendre) claimed to have discovered the method of least squares in 1795.[2]

Most historians agree that Gauss was right in his claim. Gauss used the least squares principle for analyzing survey data and in astronomical calculations. A famous example is when Gauss successfully predicted the orbit of the asteroid Ceres in 1801. The method of least squares quickly became the standard procedure for analysis of astronomical and geodetic data. There are several good accounts of the history of the invention of least squares and the dispute between Gauss and Legendre; see Placket [660, 1972], Stigler [757, 1977], [758, 1981], and Goldstine [363, 1977].

Gauss gave the method a sound theoretical basis in "Theoria Combinationis" [322, 1821], [323, 1823]. These two memoirs of Gauss, which contain his definitive treatment of the area, have recently been collected for the first time in an English translation by Stewart [325, 1995]. Gauss proves here the optimality of the least squares estimate without any assumptions that the random variables follow a particular distribution. This contribution of Gauss was somehow neglected until being rediscovered by Markoff [566, 1912]; see Theorem 1.1.1.

1.1.2. Statistical preliminaries.

Let y be a random variable having the distribution function $F(y)$, where $F(y)$ is nondecreasing, right continuous, and

$$0 \le F(y) \le 1, \quad F(-\infty) = 0, \quad F(\infty) = 1.$$

The expected value and the variance of y is then defined as

$$\mathcal{E}(y) = \mu = \int_{-\infty}^{\infty} y dF(y), \quad \mathcal{E}((y-\mu)^2) = \sigma^2 = \int_{-\infty}^{\infty} (y-\mu)^2 dF(y).$$

[2] "Our principle, which we have made use of since 1795, has lately been published by Legendre...," C. F. Gauss, *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, Hamburg [320, 1809].

Let $y = (y_1, \ldots, y_n)^T$ be a vector of random variables and let $\mu = (\mu_1, \ldots, \mu_n)$, where $\mu_i = \mathcal{E}(y_i)$. Then we write $\mu = \mathcal{E}(y)$. If y_i and y_j have the joint distribution $F(y_i, y_j)$ the **covariance** σ_{ij} between y_i and y_j is defined by

$$\operatorname{cov}(y_i, y_j) = \mathcal{E}[(y_i - \mu_i)(y_j - \mu_j)] = \sigma_{ij} = \int_{-\infty}^{\infty} (y_i - \mu_i)(y_j - \mu_j) dF(y_i, y_j).$$

Note that $\sigma_{ij} = \mathcal{E}(y_i y_j) - \mu_i \mu_j$. The variance-covariance matrix $V \in \mathbf{R}^{n \times n}$ of y is defined by

$$\mathcal{V}(y) = V = \mathcal{E}[(y - \mu)(y - \mu)^T] = \mathcal{E}(yy^T) - \mu\mu^T.$$

We now prove some properties which will be useful in the remainder of the book.

LEMMA 1.1.1. *Let $z = Fy$, where $F \in \mathbf{R}^{r \times n}$ is a given matrix and y a random vector with $\mathcal{E}(y) = \mu$ and covariance matrix V. Then*

$$\mathcal{E}(z) = F\mu, \qquad \mathcal{V}(z) = FVF^T.$$

Proof. The first property follows directly from the definition of expected value. The second is proved as

$$\mathcal{V}(Fy) = \mathcal{E}[F(y - \mu)(y - \mu)^T F^T] = F\mathcal{E}[(y - \mu)(y - \mu)^T]F^T = FVF^T. \quad \blacksquare$$

In the special case when $F = f^T$ is a row vector, then $z = f^T y$ is a linear functional of y and $\mathcal{V}(z) = \mu\|b\|_2^2$. The following lemma is given without proof.

LEMMA 1.1.2. *Let $A \in \mathbf{R}^{n \times n}$ be a symmetric matrix and consider the quadratic form $y^T Ay$, where y is a random vector with expected value μ and covariance matrix V. Then*

$$\mathcal{E}(y^T Ay) = \mu^T A\mu + \operatorname{trace}(AV),$$

where trace(AV) *denotes the sum of diagonal elements of AV.*

1.1.3. Linear models and the Gauss–Markoff theorem.

In linear statistical models one assumes that the vector $b \in \mathbf{R}^m$ of observations is related to the unknown parameter vector $x \in \mathbf{R}^n$ by a linear relation

$$Ax = b + \epsilon, \tag{1.1.2}$$

where $A \in \mathbf{R}^{m \times n}$ is a known matrix and ϵ is a vector of random errors. In the **standard linear model** we have

$$\mathcal{E}(\epsilon) = 0, \qquad \mathcal{V}(\epsilon) = \sigma^2 I, \tag{1.1.3}$$

i.e., the random variables ϵ_i are uncorrelated and all have zero means and the same variance. We also assume that rank$(A) = n$.

We make the following definitions.

DEFINITION 1.1.1. *A function $g(y)$ of the random vector y is an unbiased estimate of a parameter θ if $\mathcal{E}(g(y)) = \theta$. When such a function exists, then θ is called an estimable parameter.*

DEFINITION 1.1.2. *Let c be a constant vector. Then the linear function $g = c^T y$ is called a minimum variance unbiased estimate of θ if $\mathcal{E}(g) = \theta$, and $\mathcal{V}(g)$ is minimized over all linear estimators.*

The following theorem by Gauss placed the method of least squares on a sound theoretical basis without any assumption that the random errors follow a normal distribution.

THEOREM 1.1.1. The Gauss–Markoff theorem. *Consider the linear model* (1.1.2), *where $A \in \mathbf{R}^{m\times n}$ is a known matrix of rank n, $\hat{b} = b + \epsilon$, where ϵ is a random vector with mean and variance given by* (1.1.3). *Then the best linear unbiased estimator of any linear function $c^T x$ is $c^T \hat{x}$, where $\hat{x}$ is the least squares estimator, obtained by minimizing the sum of squares $\|Ax - \hat{b}\|_2^2$. Furthermore, $\mathcal{E}(s^2) = \sigma^2$, where s^2 is the quadratic form*

$$s^2 = \frac{1}{m-n}(b - A\hat{x})^T(b - A\hat{x}) = \frac{1}{m-n}\|b - A\hat{x}\|_2^2. \tag{1.1.4}$$

Proof. For a modern proof see Zelen [850, 1962, pp. 560–561]. ∎

COROLLARY 1.1.1. *The variance-covariance matrix of the least squares estimate $\hat{x}$ is*

$$\mathcal{V}(\hat{x}) = \sigma^2(A^T A)^{-1}. \tag{1.1.5}$$

Proof. Since $\hat{x} = (A^T A)^{-1} A^T \hat{b}$ it follows from Lemma 1.1.1 that

$$\mathcal{V}(\hat{x}) = (A^T A)^{-1} A^T \mathcal{V}(\hat{b}) A (A^T A)^{-1} = \sigma^2 (A^T A)^{-1}. \quad ∎$$

The residual vector $\hat{r} = b - A\hat{x}$ satisfies $A^T\hat{r} = 0$, and hence there are n linear relations among the m components of $\hat{r}$. It can be shown that the residuals $\hat{r}$, and therefore also the quadratic form s^2, are uncorrelated with $\hat{x}$, i.e.,

$$\operatorname{cov}(\hat{r}, \hat{x}) = 0, \qquad \operatorname{cov}(s^2, \hat{x}) = 0.$$

In the **general univariate linear model** the covariance matrix is $\mathcal{V}(\epsilon) = \sigma^2 W$, where $W \in \mathbf{R}^{m\times m}$ is a positive semidefinite symmetric matrix. If A has rank n and W is positive definite then the best unbiased linear estimate for x was shown by Aiken [4, 1934] to be the solution of

$$\min_x (Ax - b)^T W^{-1} (Ax - b). \tag{1.1.6}$$

General linear models are considered in Section 4.3.1, and special methods for the corresponding generalized least squares problems are treated in Sections 4.3 and 4.4. It is important to note that for singular W the best unbiased linear estimate of x can *not* always be obtained by replacing W^{-1} in (1.1.6) by the Moore–Penrose pseudoinverse $W^\dagger$!

In some applications it might be more adequate to consider the more general minimization problem

$$\min_x \|Ax - b\|_p, \tag{1.1.7}$$

where the Hölder vector p-norms $\|\cdot\|_p$ are defined by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \qquad 1 \le p < \infty. \tag{1.1.8}$$

The Euclidian norm corresponds to $p = 2$, and in the limiting case

$$\|x\|_\infty = \max_{1\le i\le n} |x_i|. \tag{1.1.9}$$

EXAMPLE 1.1.2. To illustrate the effect of using a Hölder norm with $p \neq 2$, we consider the problem of estimating the scalar γ from m observations $y \in \mathbf{R}^m$. This is equivalent to minimizing $\|A\gamma - y\|_p$, with $A = (1, 1, \ldots, 1)^T$. It is easily verified that if $y_1 \ge y_2 \ge \cdots \ge y_m$, then the solution for some different values p are

$$\begin{aligned} \gamma_1 &= y_{(m+1)/2}, \quad (m \text{ odd}), \\ \gamma_2 &= (y_1 + y_2 + \cdots + y_m)/m, \\ \gamma_\infty &= (y_1 + y_m)/2. \end{aligned}$$

These estimates correspond to the **median**, **mean**, and **midrange**, respectively. Note that the estimate γ_1 is insensitive to extreme values of y_i. This property carries over to more general problems, and a small number of isolated large errors will usually not change the l_1 solution. For a treatment of problem (1.1.7) when $p \neq 2$ see Section 4.5.

1.1.4. Characterization of least squares solutions. We begin by characterizing the set of all solutions to the least squares problem (1.1.1).

THEOREM 1.1.2. *Denote the set of all solutions to* (1.1.1) *by*

$$\mathcal{S} = \{x \in \mathbf{R}^n \mid \|Ax - b\|_2 = \min\}. \tag{1.1.10}$$

Then $x \in \mathcal{S}$ if and only if the following orthogonality condition holds:

$$A^T(b - Ax) = 0. \tag{1.1.11}$$

Proof. Assume that $\hat{x}$ satisfies $A^T\hat{r} = 0$, where $\hat{r} = b - A\hat{x}$. Then for any $x \in \mathbf{R}^n$ we have $r = b - Ax = \hat{r} + A(\hat{x} - x) \equiv \hat{r} + Ae$. Squaring this we obtain

$$r^T r = (\hat{r} + Ae)^T(\hat{r} + Ae) = \hat{r}^T\hat{r} + \|Ae\|_2^2,$$

which is minimized when $x = \hat{x}$.

On the other hand suppose $A^T\hat{r} = z \neq 0$, and take $x = \hat{x} + \epsilon z$. Then $r = \hat{r} - \epsilon Az$, and

$$r^T r = \hat{r}^T\hat{r} - 2\epsilon z^T z + \epsilon^2 (Az)^T Az < \hat{r}^T\hat{r}$$

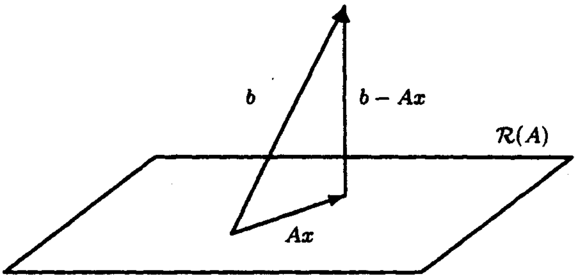

FIG. 1.1.1. *Geometric interpretation of least squares property.*

for sufficiently small ϵ. Hence $\hat{x}$ is not a least squares solution. ∎

The **range** (or column space) of a matrix $A \in \mathbf{R}^{m \times n}$ is defined to be

$$\mathcal{R}(A) = \{ z = Ax \mid x \in \mathbf{R}^n \}. \tag{1.1.12}$$

The set of solutions to $A^T y = 0$ is a subspace called the **nullspace** of A^T and denoted by

$$\mathcal{N}(A^T) = \{ y \in \mathbf{R}^m \mid A^T y = 0 \}, \tag{1.1.13}$$

and is the orthogonal complement in $\mathbf{R}^m$ to the space $\mathcal{R}(A)$. These are two of the four fundamental subspaces of the matrix A; see Section 1.2. Theorem 1.1.2 asserts that the residual vector $r = b - Ax$ of a least squares solution lies in $\mathcal{N}(A^T)$. Hence any least squares solution x uniquely decomposes the right-hand side b into two orthogonal components

$$b = Ax + r, \quad Ax \in \mathcal{R}(A), \quad r \in \mathcal{N}(A^T). \tag{1.1.14}$$

This geometric interpretation is illustrated for $n = 2$ in Figure 1.1.1.

From (1.1.11) it follows that a least squares solution satisfies the **normal equations**

$$A^T A x = A^T b. \tag{1.1.15}$$

The matrix $A^T A \in \mathbf{R}^{n \times n}$ is symmetric and nonnegative definite. The normal equations are always consistent since

$$A^T b \in \mathcal{R}(A^T) = \mathcal{R}(A^T A).$$

Furthermore we have the following theorem.

THEOREM 1.1.3. *The matrix $A^T A$ is positive definite if and only if the columns of A are linearly independent, i.e.,* rank $(A) = n$.

Proof. If the columns of A are linearly independent, then $x \neq 0 \Rightarrow Ax \neq 0$ and therefore $x \neq 0 \Rightarrow x^T A^T A x = \|Ax\|_2^2 > 0$. Hence $A^T A$ is positive definite. On the other hand, if the columns are linearly dependent then for some $x_0 \neq 0$ we have $Ax_0 = 0$ and so $x_0^T A^T A x_0 = 0$, and $A^T A$ is not positive definite. ∎

From Theorem 1.1.3 it follows that if rank$(A) = n$, then the unique least squares solution x and the corresponding residual $r = b - Ax$ are given by

$$x = (A^TA)^{-1}A^Tb, \qquad r = b - A(A^TA)^{-1}A^Tb. \tag{1.1.16}$$

If $S \subset \mathbf{R}^m$ is a subspace, then $P_S \in \mathbf{R}^{m\times m}$ is the **orthogonal projector** onto S if $\mathcal{R}(P_S) = S$, and

$$P_S^2 = P_S, \quad P_S^T = P_S. \tag{1.1.17}$$

Moreover,

$$(I - P_S)^2 = (I - P_s), \quad (I - P_S)P_S = 0,$$

and $(I - P_S)$ is the projector for the subspace complementary to that of S.

Let P_1 and P_2 be orthogonal projectors onto S. Then using (1.1.17) we have for all $z \in \mathbf{R}^m$

$$\|(P_1 - P_2)z\|_2^2 = z^TP_1(I - P_2)z + z^TP_2(I - P_1)z = 0.$$

It follows that $P_1 = P_2$, and hence the orthogonal projector is unique.

From the geometric interpretation (see Figure 1.1.1) Ax is the orthogonal projection of b onto $\mathcal{R}(A)$. We have $r = (I - P_{\mathcal{R}(A)})b$, and in the full rank case

$$P_{\mathcal{R}(A)} = A(A^TA)^{-1}A^T. \tag{1.1.18}$$

If rank$(A) < n$ then A has a nontrivial nullspace and the least squares solution is not unique. If $\hat{x}$ is a particular least squares solution then the set of all least squares solutions is

$$\mathcal{S} = \{x = \hat{x} + z \mid z \in \mathcal{N}(A)\}.$$

If $\hat{x} \perp \mathcal{N}(A)$ then $\|x\|_2^2 = \|\hat{x}\|_2^2 + \|z\|_2^2$, and therefore $\hat{x}$ is the unique least squares solution of minimum norm.

The problem of computing the minimum norm solution $y \in \mathbf{R}^m$ to an underdetermined system of linear equations

$$\min \|y\|_2, \qquad A^Ty = c, \tag{1.1.19}$$

where $A \in \mathbf{R}^{m\times n}$, occurs as a subproblem in optimization algorithms. If rank$(A) = n$, then the system $A^Ty = c$ is consistent and the unique solution of (2.5) is given by the **normal equations of the second kind**

$$A^TAz = c, \qquad y = Az, \tag{1.1.20}$$

that is, $y = A(A^TA)^{-1}c$.

The classical method for solving the normal equations is based on the following matrix factorization.

THEOREM 1.1.4. Cholesky Decomposition. *Let the matrix $C \in R^{n\times n}$ be symmetric and positive definite. Then there is a unique upper triangular matrix R with positive diagonal elements such that*

$$C = R^T R. \tag{1.1.21}$$

R is called the Cholesky factor of C and (1.1.21) *is called the Cholesky factorization.*

Proof. The proof is by induction on the order n of C. The result is trivial for $n = 1$. Assume that (1.1.21) holds for all positive definite matrices of order n. Consider the positive definite matrix $\tilde{C}$ of order $n+1$, and seek a factorization

$$\tilde{C} = \begin{pmatrix} C & c \\ c^T & \gamma \end{pmatrix} = \begin{pmatrix} R^T & 0 \\ r^T & \rho \end{pmatrix} \begin{pmatrix} R & r \\ 0 & \rho \end{pmatrix}. \tag{1.1.22}$$

C is a principal minor of $\tilde{C}$ and hence positive definite. By the induction hypothesis the factorization $C = R^T R$ exists and thus (1.1.22) holds provided r and $\rho > 0$ satisfy

$$R^T r = m, \quad \rho^2 = \gamma - r^T r. \tag{1.1.23}$$

Since R^T has positive diagonal elements and is lower triangular, $r = R^{-T} m$ is uniquely determined. Now, from the positive definiteness of $\tilde{C}$ it follows that

$$\begin{aligned} 0 < (r^T R^{-T} \quad -1) \begin{pmatrix} C & c \\ c^T & \gamma \end{pmatrix} \begin{pmatrix} R^{-1} r \\ -1 \end{pmatrix} &= r^T R^{-T} C R^{-1} r - 2 r^T R^{-T} m + \gamma \\ &= r^T r - 2 r^T r + \gamma = \gamma - r^T r. \end{aligned}$$

Hence also $\rho = (\gamma - r^T r)^{1/2}$ is uniquely determined. ∎

Another characterization of the least squares solution is given in the following theorem.

THEOREM 1.1.5. *Assume that $A \in \mathbf{R}^{m\times n}$ has rank n. Then the symmetric linear system*

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix} \tag{1.1.24}$$

is nonsingular and gives the condition for the solution of both the primal and dual least squares problem

$$\min_x \left\{ \|Ax - b\|_2^2 + 2c^T x \right\}, \tag{1.1.25}$$

$$\min_y \|y - b\|_2^2, \qquad A^T y = c. \tag{1.1.26}$$

Proof. The system (1.1.24), often called the **augmented system**, can be obtained by differentiating (1.1.25) to give $A^T(b - Ax) = c$, and setting y to be the residual $y = b - Ax$. The system can also be obtained by differentiating the Lagrangian

$$L(x, y) = \|y - b\|_2^2 + 2x^T(A^T y - c)$$

of (1.1.26), and equating to zero. Here x is the vector of Lagrange multipliers. ∎

Setting $c = 0$ in (1.1.25) gives the linear least squares problem (1.1.2). Setting $b = 0$ in (1.1.26) gives the problem of minimum 2-norm solution of an underdetermined linear system $A^Ty = c$; see (1.1.19).

1.2. The Singular Value Decomposition

1.2.1. The singular value decomposition. The singular value decomposition (SVD) of a matrix $A \in \mathbf{R}^{m\times n}$ is a matrix decomposition of great theoretical and practical importance for the treatment of least squares problems. It provides a diagonal form of A under an orthogonal equivalence transformation. The history of this matrix decomposition goes back more than a century; see the very interesting survey of the early history of the SVD by Stewart [750, 1993]. However, only recently has the SVD been as much used as it should. Now it is a main tool in numerous application areas such as signal and image processing, control theory, pattern recognition, time-series analysis, etc.

Because applications exist also for complex matrices we state the theorem below for matrices with complex elements. (The matrix A^H will denote the matrix formed by conjugating each element and taking the transpose.)

Theorem 1.2.1. Singular Value Decomposition. *Let $A \in \mathbf{C}^{m\times n}$ be a matrix of rank r. Then there exist unitary matrices $U \in \mathbf{C}^{m\times m}$ and $V \in \mathbf{C}^{n\times n}$ such that*

$$A = U\Sigma V^H, \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{1.2.1}$$

where $\Sigma \in \mathbf{R}^{m\times n}$, $\Sigma_1 = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

The σ_i are called the singular values of A, and if we write

$$U = (u_1, \ldots, u_m), \quad V = (v_1, \ldots, v_n), \tag{1.2.2}$$

the u_i and v_i are, respectively, the left and right singular vectors associated with σ_i, $i = 1, \ldots, r$.

Proof. (See Golub and Van Loan [389, 1989].) Let $v_1 \in \mathbf{C}^n$ be a vector such that

$$\|v_1\|_2 = 1, \quad \|Av_1\|_2 = \|A\|_2 = \sigma,$$

where σ is real and positive. The existence of such a vector follows from the definition of a matrix subordinate norm $\|A\|$. If $\sigma = 0$, then $A = 0$, and we can take $\Sigma = 0$ and U and V arbitrary unitary matrices. Therefore assume that $\sigma > 0$, and take $u_1 = (1/\sigma)Av_1 \in \mathbf{C}^m$, $\|v_1\|_2 = 1$. Let the matrices

$$V = (v_1, V_1) \in \mathbf{C}^{n\times n}, \quad U = (u_1, U_1) \in \mathbf{C}^{m\times m}$$

be unitary. (Recall that it is always possible to extend a unitary set of vectors to a unitary basis for the whole space.) Since $U_1^H Av_1 = \sigma U_1^H u_1 = 0$ it follows that $U^H AV$ has the following structure:

$$A_1 \equiv U^H AV = \begin{pmatrix} \sigma & w^H \\ 0 & B \end{pmatrix},$$

where $w^H = u_1^H AV_1$, and $B = U_1^H AV_1 \in \mathbf{C}^{(m-1)\times(n-1)}$. From the two inequalities

$$\|A_1\|_2(\sigma^2 + w^H w)^{1/2} \geq \left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \sigma^2 + w^H w \\ Bw \end{pmatrix} \right\|_2 \geq \sigma^2 + w^H w,$$

it follows that $\|A_1\|_2 \geq (\sigma^2 + w^H w)^{1/2}$. But since U and V are unitary, $\|A_1\|_2 = \|A\|_2 = \sigma$, and thus $w = 0$. The proof can now be completed by an induction argument on the smallest dimension $\min(m, n)$. ∎

A rectangular matrix $A \in \mathbf{R}^{m\times n}$ represents a linear mapping from $\mathbf{C}^n$ to $\mathbf{C}^m$. The significance of Theorem 1.2.1 is that it shows that there is an orthogonal basis in each of these spaces, with respect to which this mapping is represented by a generalized diagonal matrix Σ with real elements. Methods for computing the SVD are described in Section 2.6.

The SVD of A can be written

$$A = U_1 \Sigma_1 V_1^H = \sum_{i=1}^{r} \sigma_i u_i v_i^H, \tag{1.2.3}$$

where

$$U_1 = (u_1, \ldots, u_r), \qquad V_1 = (v_1, \ldots, v_r). \tag{1.2.4}$$

By this a matrix A of rank r is decomposed into a sum of $r = \operatorname{rank}(A)$ matrices of rank one.

The singular values of A are unique. The singular vector v_j, $j \leq r$, will be unique only when σ_j^2 is a simple eigenvalue of $A^H A$. For multiple singular values, the corresponding singular vectors can be chosen as any orthonormal basis for the unique subspace that they span. Once the singular vectors v_j, $1 \leq j \leq r$, have been chosen, the vectors u_j, $1 \leq j \leq r$, are uniquely determined from

$$A v_j = \sigma_j u_j, \qquad j = 1, \ldots, r. \tag{1.2.5}$$

Similarly, given u_j, $1 \leq j \leq r$, the vectors v_j, $1 \leq j \leq r$, are uniquely determined from

$$A^H u_j = \sigma_j v_j, \qquad j = 1, \ldots, r. \tag{1.2.6}$$

The SVD gives complete information about the four fundamental subspaces associated with A. It is easy to verify that

$$\mathcal{N}(A) = \operatorname{span}[v_{r+1}, \ldots, v_n], \qquad \mathcal{R}(A) = \operatorname{span}[u_1, \ldots, u_r], \tag{1.2.7}$$

$$\mathcal{R}(A^H) = \operatorname{span}[v_1, \ldots, v_r], \qquad \mathcal{N}(A^H) = \operatorname{span}[u_{r+1}, \ldots, u_m], \tag{1.2.8}$$

and we find the well-known relations

$$\mathcal{N}(A)^\perp = \mathcal{R}(A^H), \quad \mathcal{R}(A)^\perp = \mathcal{N}(A^H).$$

Note that with $V = (V_1,\ V_2)$ and $z \in \mathbf{C}^{n-r}$ an arbitrary vector,

$$x = V_2 z = \sum_{j=r+1}^{n} z_j v_j \tag{1.2.9}$$

gives the general solution to the homogeneous linear system $Ax = 0$. This result is often useful in optimization problems.

1.2.2. Related eigenvalue decompositions. There is a close relationship between the SVD and the Hermitian (or real symmetric) eigenvalue problem from (1.2.1) it follows that

$$(1.2.10) \qquad A^HA = V\Sigma^T\Sigma V^H, \qquad AA^H = U\Sigma\Sigma^T U^H.$$

Here,

$$\Sigma^T\Sigma = \begin{pmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbf{R}^{n\times n}, \quad \Sigma\Sigma^T = \begin{pmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbf{R}^{m\times m},$$

and thus $\sigma_1^2, \dots, \sigma_r^2$, are the nonzero eigenvalues of the Hermitian and positive semidefinite matrices A^HA and AA^H, and v_j and u_j are the corresponding eigenvectors. For a proof of the SVD using this relationship, see Stewart [729, 1973, p. 319].

A matrix $A \in \mathbf{C}^{n\times n}$ is Hermitian if $A^H = A$. A Hermitian matrix A has real eigenvalues $\lambda_1, \dots, \lambda_n$, and then $A^HA = A^2$ as real nonnegative eigenvalues equal to λ_i^2, $i = 1, \dots, n$. Hence, (1.2.10) shows that for a Hermitian matrix the singular values are given by $\sigma_i = |\lambda_i|$, $i = 1, \dots, n$.

In principle, the SVD can be found from the eigenvalue decomposition of the two Hermitian matrices A^HA and AA^H. However, this does not lead to a stable algorithm for computing the SVD.

EXAMPLE 1.2.1. Consider the case $n = 2$,

$$A = (a_1, a_2) \in \mathbf{R}^{m\times 2}, \quad a_1^Ta_2 = \cos\gamma,$$

and $\|a_1\|_2 = \|a_2\|_2 = 1$. Here γ is the angle between the vectors a_1 and a_2. The matrix

$$A^TA = \begin{pmatrix} 1 & \cos\gamma \\ \cos\gamma & 1 \end{pmatrix}$$

has eigenvalues $\lambda_1 = 2\cos^2(\gamma/2)$, $\lambda_2 = 2\sin^2(\gamma/2)$, and so,

$$\sigma_1 = \sqrt{2}\cdot\cos\frac{\gamma}{2}, \quad \sigma_2 = \sqrt{2}\sin\frac{\gamma}{2}.$$

The eigenvectors of A^TA,

$$v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

are the right singular vectors of A. The left singular vectors can be determined from (1.2.5).

Numerically, if γ is less than the square root of machine precision then $\cos\gamma \approx 1 - \gamma^2/2 = 1$, and A^TA has only one nonzero eigenvalue equal to 2. Thus the smallest singular value of A has been lost! ∎

The following relationship between the SVD and a Hermitian eigenvalue problem, which can easily be verified, was exploited by Lanczos [513, 1961, Chap. 3].

THEOREM 1.2.2. *Let the SVD of* $A \in \mathbf{C}^{m\times n}$ *be* $A = U\Sigma V^H$, *where* $\Sigma = \operatorname{diag}(\Sigma_1,\ 0)$,

$$U = (U_1,\ U_2),\quad U_1 \in \mathbf{C}^{m\times r},\quad V = (V_1,\ V_2),\quad V_1 \in \mathbf{C}^{n\times r}.$$

Then

$$C = \begin{pmatrix} 0 & A \\ A^H & 0 \end{pmatrix} = P^H \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & -\Sigma_1 & 0 \\ 0 & 0 & 0 \end{pmatrix} P, \tag{1.2.11}$$

where P is unitary

$$P = \frac{1}{\sqrt{2}} \begin{pmatrix} U_1 & U_1 & \sqrt{2}U_2 & 0 \\ V_1 & -V_1 & 0 & \sqrt{2}V_2 \end{pmatrix}^H . \tag{1.2.12}$$

Hence the eigenvalues of C are $\pm\sigma_1, \pm\sigma_2, \dots, \pm\sigma_r$, *and zero repeated* $(m+n-2r)$ *times, where* $r = \operatorname{rank}(A)$.

1.2.3. Matrix approximations.

The singular value decomposition plays an important role in a number of matrix approximation problems. In the theorem below we consider the approximation of one matrix by another of lower rank. Several other results can be found in Golub [365, 1968] and in Golub and Van Loan [389, 1989, Chap. 12.4].

THEOREM 1.2.3. *Let* $A \in \mathbf{C}^{m\times n}$ *have* $\operatorname{rank}(A) = r$, *and the SVD*

$$A = U\Sigma V^H = \sum_{i=1}^{r} \sigma_i u_i v_i^H .$$

Let $B \in \mathcal{M}_k^{m\times n}$, *where* $\mathcal{M}_k^{m\times n}$ *is the set of matrices in* $\mathbf{C}^{m\times n}$ *of rank* $k < r$. *Then*

$$\min \|A - X\|_2, \quad X \in \mathcal{M}_k^{m\times n},$$

is obtained for $X = B$, *where*

$$B = \sum_{i=1}^{k} \sigma_i u_i v_i^H, \qquad \|A - B\|_2 = \sigma_{k+1}.$$

Proof. See Golub and Van Loan [389, 1989, Chap. 2.5.4] and Mirsky [578, 1960]. ∎

As a special case of this theorem it follows that if $\operatorname{rank}(A) = n$, then σ_n is the shortest distance from A to the set of singular matrices in the spectral norm.

REMARK 1.2.1. The theorem was originally proved for the Frobenius norm (see (1.4.7)). For this norm the minimum distance is

$$\|A - B\|_F = (\sigma_{k+1}^2 + \cdots + \sigma_r^2)^{1/2},$$

and the solution is unique; see Eckhart and Young [261, 1936]. A generalization of the Eckhart–Young theorem is given by Golub, Hoffman, and Stewart [369, 1987]. ∎

Closely related to the singular value decomposition is the **polar decomposition.**

THEOREM 1.2.4. Polar Decomposition. *Let $A \in \mathbf{C}^{m\times n}$, $m \geq n$. Then there exist a matrix $Q \in \mathbf{C}^{m\times n}$ and a unique Hermitian positive semidefinite matrix $H \in \mathbf{C}^{n\times n}$ such that*

$$A = QH, \qquad Q^H Q = I. \tag{1.2.13}$$

If rank$(A) = n$ *then H is positive definite and Q is uniquely determined.*

Proof. Let A have the singular value decomposition

$$A = U\begin{pmatrix}\Sigma_1\\ 0\end{pmatrix}V^H, \qquad \Sigma_1 = \text{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_n),$$

where U and V are unitary and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. It follows that $A = QH$, where

$$Q = U_1 V^H, \qquad H = V\Sigma V^H,$$

and $U_1 = (u_1, \ldots, u_n)$. ∎

The polar decomposition can be regarded as a generalization to matrices of the complex number representation $z = re^{i\theta}$, $r \geq 0$. Since $H^2 = V\Sigma^2 V^H = A^H A$ it follows that H equals the unique Hermitian positive semidefinite square root of $A^H A$,

$$H = (A^H A)^{1/2}.$$

The unitary factor U in the polar decomposition possesses a best approximation property described in the following theorem from Higham [453, 1986].

THEOREM 1.2.5. *Let $A, B \in \mathbf{C}^{m\times n}$ and let $B^H A \in \mathbf{C}^{n\times n}$ have the polar decomposition $B^H A = UH$. Then, for any unitary $Z \in \mathbf{C}^{n\times n}$,*

$$\|A - BU\|_F \leq \|A - BZ\|_F \leq \|A + BU\|_F, \tag{1.2.14}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. In the special case in which $m = n$ and $B = I$ we have

$$\|A - U\|_F \leq \|A - Z\|_F \leq \|A + U\|_F, \tag{1.2.15}$$

and the minimum is

$$\|A - U\|_F = \left(\sum_{i=1}^{n}(\sigma_i - 1)^2\right)^{1/2},$$

where $\sigma_i = \sigma_i(A)$.

Hence the nearest unitary matrix to $A \in \mathbf{C}^{n\times n}$ is the unitary factor of the polar decomposition. Fan and Hoffman [286, 1955] showed that (1.2.15) holds for any unitarily invariant norm. Higham [453, 1986] also discusses the approximation properties of the Hermitian factor H.

1.2.4. The sensitivity of singular values and vectors.

Like the eigenvalues of a real Hermitian matrix, the singular values of a general matrix have a minmax characterization.

THEOREM 1.2.6. *Let $A \in \mathbf{R}^{m\times n}$ have singular values*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0, \qquad p = \min(m,n),$$

and S be a linear subspace of $\mathbf{R}^n$. Then

$$\sigma_i = \min_{\dim(S)=n-i+1} \max_{\substack{x\in S\\ x\neq 0}} \frac{\|Ax\|_2}{\|x\|_2}. \tag{1.2.16}$$

Proof. The result is established in almost the same way as for the corresponding eigenvalue theorem, the Courant–Fischer theorem; see Wilkinson [836, 1965, pp. 99–101]. ∎

The minmax characterization of the singular values may be used to establish results on the sensitivity of the singular values of A to perturbations.

THEOREM 1.2.7. *Let A and $\tilde{A} = A+E \in \mathbf{R}^{m\times n}$, $m \geq n$, have singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ and $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_n$, respectively. Then*

$$|\sigma_i - \tilde{\sigma}_i| \leq \|E\|_2, \tag{1.2.17}$$

$$\sum_{i=1}^{n} |\sigma_i - \tilde{\sigma}_i|^2 \leq \|E\|_F^2. \tag{1.2.18}$$

Proof. See Stewart [729, 1973, pp. 321–322]. ∎

The result (1.2.18) is known as the Wielandt–Hoffman theorem for singular values. The theorem shows the important fact that the singular values of a matrix A are well-conditioned with respect to perturbations of A. Perturbations of the elements of a matrix produce perturbations of the same, or smaller, magnitude in the singular values. This is of great importance for the use of the SVD to determine the "numerical rank" of a matrix; see Section 2.7.1.

The next result gives a perturbation result for singular vectors.

THEOREM 1.2.8. *Let $A \in \mathbf{R}^{m\times n}$, $m \geq n$, have singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ and singular vectors u_i, v_i, $i = 1,\ldots,n$. Let $\tilde{\sigma}_i$, $\tilde{u}_i$, and $\tilde{v}_i$ be the corresponding values for $\tilde{A} = A + E$. Then if $\|E\|_2 < \gamma_i$ it holds*

$$\max(\sin\theta(u_i,\tilde{u}_i),\ \sin\theta(v_i,\tilde{v}_i)) \leq \frac{\|E\|_2}{\gamma_i - \|E\|_2}, \tag{1.2.19}$$

where γ_i is the **absolute gap** *between σ_i and the other singular values,*

$$\gamma_i = \min_{j\neq i} |\sigma_i - \sigma_j|. \tag{1.2.20}$$

Proof. A more general result is given in Golub and Van Loan [389, 1989, Thm. 8.3.5]. ∎

Sharper perturbation results can be given for singular values and vectors of bidiagonal matrices; see Theorem 2.6.2.

It is well known that the eigenvalues of the leading principal minor of order $(n-1)$ of a Hermitian matrix $A \in \mathbf{R}^{n\times n}$ interlace the eigenvalues of A; see Wilkinson [836, 1965, p. 103]. A similar theorem holds for singular values.

THEOREM 1.2.9. *Let A be bordered by a column $u \in \mathbf{R}^m$,*

$$\hat{A} = (A, u) \in \mathbf{R}^{m\times n}, \quad m \geq n.$$

Then the ordered singular values σ_i of A separate the ordered singular values $\hat{\sigma}_i$ of $\hat{A}$ as follows:

$$\hat{\sigma}_1 \geq \sigma_1 \geq \hat{\sigma}_2 \geq \sigma_2 \geq \cdots \geq \hat{\sigma}_{n-1} \geq \sigma_{n-1} \geq \hat{\sigma}_n.$$

Similarly, if A is bordered by a row $v \in \mathbf{R}^n$,

$$\hat{A} = \begin{pmatrix} A \\ v^H \end{pmatrix} \in \mathbf{R}^{m\times n}, \quad m \geq n,$$

$$\hat{\sigma}_1 \geq \sigma_1 \geq \hat{\sigma}_2 \geq \sigma_2 \geq \cdots \geq \hat{\sigma}_{n-1} \geq \sigma_{n-1} \geq \hat{\sigma}_n \geq \sigma_n.$$

Proof. The theorem is a consequence of the minmax characterization of the singular values in Theorem 1.2.6; cf. Lawson and Hanson [520, 1974, p. 26]. ∎

1.2.5. The SVD and pseudoinverse. The SVD is a powerful tool for solving the linear least squares problem. This is because the unitary matrices that transform A to diagonal form (1.2.1) do not change the l_2-norm of vectors. We have the following fundamental result, which applies to both overdetermined and underdetermined linear systems.

THEOREM 1.2.10. *Consider the general linear least squares problem*

$$\min_{x\in S} \|x\|_2, \qquad S = \{x \in \mathbf{R}^n \mid \|b - Ax\|_2 = \min\}, \tag{1.2.21}$$

where $A \in \mathbf{C}^{m\times n}$ and $\operatorname{rank}(A) = r \leq \min(m, n)$. *This problem always has a unique solution, which can be written in terms of the SVD of A as*

$$x = V \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^H b. \tag{1.2.22}$$

Proof. Let

$$z = V^H x = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \qquad c = U^H b = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

where $z_1, c_1 \in \mathbf{C}^r$. Then

$$\begin{aligned} \|b - Ax\|_2 &= \|U^H(b - AVV^H x)\|_2 \\ &= \left\| \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} - \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} c_1 - \Sigma_1 z_1 \\ c_2 \end{pmatrix} \right\|_2. \end{aligned}$$

Thus, the residual norm will be minimized for z_2 arbitrary and $z_1 = \Sigma_r^{-1} c_1$. The choice $z_2 = 0$ minimizes $\|z\|_2$, and therefore $\|x\|_2 = \|Vz\|_2$ as well. ∎

DEFINITION 1.2.1. *We write* (1.2.22) *as $x = A^\dagger b$, where*

$$A^\dagger = V \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^H \in \mathbf{C}^{n\times m} \tag{1.2.23}$$

is called the **pseudoinverse** *of A, and the solution* (1.2.22) *is called the pseudoinverse solution.*

It follows easily from Theorem 1.2.10 that $A^\dagger$ minimizes $\|AX - I\|_F$. For computing the pseudoinverse solution it suffices to compute Σ, V, and the vector $c = U^H b$. The pseudoinverse of a scalar is

$$\sigma^\dagger = \begin{cases} 1/\sigma, & \text{if } \sigma \neq 0, \\ 0, & \text{if } \sigma = 0. \end{cases} \tag{1.2.24}$$

This shows the important fact that the pseudoinverse $A^\dagger$ is not a continuous function of A, unless we allow only perturbations which do not change the rank of A. The pseudoinverse can also be uniquely characterized by the two geometrical conditions

$$A^\dagger b \perp \mathcal{N}(A), \quad (I - AA^\dagger)b \perp \mathcal{R}(A) \quad \forall b \in \mathbf{R}^m. \tag{1.2.25}$$

The matrix $A^\dagger$ is often called the **Moore–Penrose** inverse. E. H. Moore introduced the general reciprocal in 1920. It was rediscovered by Bjerhammar [83, 1951] and Penrose [655, 1955], who gave the following elegant algebraic characterization.

THEOREM 1.2.11. Penrose's conditions. *The pseudoinverse $X = A^\dagger$ is uniquely determined by the following four conditions.*

$$\begin{array}{ll} (1) \quad AXA = A, & (2) \quad XAX = X, \\ (3) \quad (AX)^H = AX, & (4) \quad (XA)^H = XA. \end{array} \tag{1.2.26}$$

It follows in particular that $A^\dagger$ in (1.2.23) does not depend on the particular choice of U and V in the SVD. It can be directly verified that $A^\dagger$ given by (1.2.23) satisfies these four conditions. If only part of the Penrose conditions hold, the corresponding matrix X is called a generalized inverse. Such inverses have been extensively analyzed; see Nashed [596, 1976].

The pseudoinverse can be shown to have the following properties.

THEOREM 1.2.12.

1. $(A^\dagger)^\dagger = A$;
2. $(A^\dagger)^H = (A^H)^\dagger$;
3. $(\alpha A)^\dagger = \alpha^\dagger A^\dagger$;
4. $(A^H A)^\dagger = A^\dagger (A^\dagger)^H$;
5. *if U and V are unitary* $(UAV^H)^\dagger = VA^\dagger U^H$;
6. *if $A = \sum_i A_i$, where $A_i A_j^H = 0$, $A_i^H A_j = 0$, $i \neq j$, then $A^\dagger = \sum_i A_i^\dagger$;*
7. *if A is normal ($AA^H = A^H A$) then $A^\dagger A = AA^\dagger$ and $(A^n)^\dagger = (A^\dagger)^n$;*
8. *A, A^H, $A^\dagger$, and $A^\dagger A$ all have rank equal to* trace$(A^\dagger A)$.

Proof. The statements easily follow from (1.2.23). See also Penrose [655, 1955]. ■

The pseudoinverse does not share some other properties of the ordinary inverse. For example, in general

$$(AB)^\dagger \neq B^\dagger A^\dagger \quad \text{and} \quad AA^\dagger \neq A^\dagger A.$$

Example 1.2.2. If we take $A = (1 \quad 0)$ and $B = (1 \quad 1)^T$, then $AB = 1$,

$$1 = (AB)^\dagger \neq B^\dagger A^\dagger = \frac{1}{2}(1 \quad 1)\begin{pmatrix}1\\0\end{pmatrix} = \frac{1}{2},$$

and

$$AA^\dagger = (1 \quad 0)\begin{pmatrix}1\\0\end{pmatrix} = 1, \quad A^\dagger A = \begin{pmatrix}1\\0\end{pmatrix}(1 \quad 0) = \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}. \quad ∎$$

In the special case in which $A \in \mathbf{C}^{m\times n}$, $\operatorname{rank}(A) = n$,

$$A^\dagger = (A^HA)^{-1}A^H, \qquad (A^H)^\dagger = A(A^HA)^{-1}. \tag{1.2.27}$$

The first expression follows from Theorems 1.1.2 and 1.1.3 and relates to the least squares solution in the case of full column rank. The second expression follows using property 2 in Theorem 1.2.12 and relates to the minimum norm solution to an underdetermined system of full row rank.

Necessary and sufficient conditions for the relation $(AB)^\dagger = B^\dagger A^\dagger$ to hold have been given by Greville [399, 1966]. The following theorem gives useful *sufficient* conditions.

Theorem 1.2.13. *Assume that* $A \in \mathbf{C}^{m\times r}$, $B \in \mathbf{C}^{r\times n}$, *where* $\operatorname{rank}(A) = \operatorname{rank}(B) = r$. *Then it holds that*

$$(AB)^\dagger = B^\dagger A^\dagger = B^H(BB^H)^{-1}(A^HA)^{-1}A^H. \tag{1.2.28}$$

Proof. The last equality follows from (1.2.27). The first equality is verified by showing that the four Penrose conditions are satisfied. ∎

1.2.6. Orthogonal projectors and angles between subspaces.

An important property of the pseudoinverse is that it gives simple expressions for the orthogonal projections onto the four fundamental subspaces of A:

$$\begin{aligned} P_{\mathcal{R}(A)} &= AA^\dagger, & P_{\mathcal{N}(A^H)} &= I - AA^\dagger,\\ P_{\mathcal{R}(A^H)} &= A^\dagger A, & P_{\mathcal{N}(A)} &= I - A^\dagger A. \end{aligned} \tag{1.2.29}$$

These expressions are easily verified using the Penrose conditions (1.2.26).

If the columns of a matrix U are orthonormal then $U^HU = I$, and $P_{\mathcal{R}(U)} = UU^H$ satisfies (1.1.17). Using (1.2.10) we can therefore express the projections (1.2.29) in terms of the singular vectors of A as

$$\begin{aligned} P_{\mathcal{R}(A)} &= U_1U_1^H, & P_{\mathcal{N}(A^H)} &= U_2U_2^H,\\ P_{\mathcal{R}(A^T)} &= V_1V_1^H, & P_{\mathcal{N}(A)} &= V_2V_2^H, \end{aligned} \tag{1.2.30}$$

where $U_1 = (u_1, \dots, u_r)$ and $V_1 = (v_1, \dots, v_r)$.

DEFINITION 1.2.2. *Let $S_A = \mathcal{R}(A)$ and $S_B = \mathcal{R}(B)$ be two subspaces of $\mathbf{C}^m$ where without restriction we assume that $p = \dim(S_A) \geq \dim(S_B) = q \geq 1$. The* **principal angles** *θ_k, between S_A and S_B and the corresponding* **principal vectors** *u_k, v_k, $k = 1, \ldots, q$, are recursively defined by*

$$(1.2.31) \qquad \cos\theta_k = \max_{u \in S_A} \max_{v \in S_B} u^H v, \quad \|u\|_2 = \|v\|_2 = 1,$$

subject to the constraints

$$(1.2.32) \qquad u \perp u_j, \qquad v \perp v_j, \qquad j = 1, \ldots, k-1.$$

Note that for $k = 1$, the constraints are empty, and θ_1 is the smallest principal angle between S_A and S_B. The principal vectors need not be uniquely defined, but the principal angles always are. Principal angles and vectors have important applications, e.g., in statistics.

If $p = q$ the subspaces have the same dimension. In this case the distance between the subspaces S_A and S_B is defined to be

$$(1.2.33) \qquad \operatorname{dist}(S_A, S_B) = |\sin(\theta_p)| = (1 - \sigma_p^2)^{1/2},$$

where θ_p is the *largest* principal angle.

The relationship between principal angles and the SVD is given in the following theorem.

THEOREM 1.2.14. *Assume that $Q_A \in \mathbf{R}^{m \times p}$ and $Q_B \in \mathbf{R}^{m \times q}$ form unitary bases for the two subspaces S_A and S_B. Consider the SVD*

$$(1.2.34) \qquad M = Q_A^H Q_B = YCZ^H, \qquad C = \operatorname{diag}(\sigma_1, \ldots, \sigma_q),$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q$, $Y^H Y = Z^H Z = I_q$. Then the principal angles and principal vectors are given by

$$(1.2.35) \qquad \cos\theta_k = \sigma_k, \quad U = Q_A Y, \quad V = Q_B Z.$$

Proof. The proof follows from the minmax characterization of the singular values and vectors; see Theorem 1.2.6.

It can also be shown that the nonzero singular values of $(P_{S_A} - P_{S_B})$, where $P_{S_A} = Q_A Q_A^H$ and $P_{S_B} = Q_B Q_B^H$ are the orthogonal projectors, equal $\sin(\theta_k)$, $k = 1, \ldots, q$. This gives the alternative definition

$$\operatorname{dist}(S_A, S_B) = \|P_{S_A} - P_{S_B}\|_2.$$

Methods for computing principal angles and vectors, and applications are discussed in Björck and Golub [111, 1973] and Golub and Zha [394, 1994].

1.3. The QR Decomposition

1.3.1. The full rank case. The SVD of A gives the solution of the general rank deficient least squares problem (1.2.21). However, in many applications it is too expensive to compute the SVD, and one has to use simpler decompositions. Among these the most important are the QR and related decompositions.

Let $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$, and let $Q \in \mathbf{R}^{m\times m}$ be an orthogonal matrix. Since orthogonal transformations preserve the Euclidean length it follows that the linear least squares problem

$$\min_x \|Q^T(Ax-b)\|_2 \tag{1.3.1}$$

is equivalent to (1.1.1). We now show how to choose Q so that the problem (1.3.1) becomes simple to solve.

THEOREM 1.3.1. QR Decomposition. *Let $A \in R^{m\times n}, m \ge n$. Then there is an orthogonal matrix $Q \in R^{m\times m}$ such that*

$$A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{1.3.2}$$

where R is upper triangular with nonnegative diagonal elements. The decomposition (1.3.2) *is called the QR decomposition of A, and the matrix R will be called the R-factor of A.*

Proof. The proof is by induction on n. Let A be partitioned in the form $A = (a_1, A_2)$, $a_1 \in R^m$, and put $\rho = \|a_1\|_2$. Let $U = (y, U_1)$ be an orthogonal matrix with $y = a_1/\rho$ if $a_1 \neq 0$, and $y = e_1$ otherwise. Since $U_1^T y = 0$ it follows that

$$U^TA = \begin{pmatrix} \rho & r^T \\ 0 & B \end{pmatrix}, \qquad B = U_1^T A_2 \in R^{(m-1)\times(n-1)},$$

where $\rho = \|a_1\|_2$, $r = A_2^T y$.

For $n = 1$, A_2 is empty and the theorem holds with $Q = U$ and $R = \rho$, a scalar. Assume now that the induction hypothesis holds for $n-1$. Then there is an orthogonal matrix $\bar{Q}$ such that $\bar{Q}^T B = \begin{pmatrix} \bar{R} \\ 0 \end{pmatrix}$, and (1.3.2) will hold if we define

$$Q = U\begin{pmatrix} 1 & 0 \\ 0 & \bar{Q} \end{pmatrix}, \quad R = \begin{pmatrix} \rho & r^T \\ 0 & \bar{R} \end{pmatrix}. \quad \blacksquare$$

The proof of Theorem 1.3.1 gives a way to compute Q and R, provided we can construct an orthogonal matrix $U = (y,\ U_1)$ given its first column. Several ways to perform this construction using elementary orthogonal transformations are given in Section 2.2.1. The systematic use of orthogonal transformations to reduce matrices to simpler form was initiated by Givens [361, 1958] and Householder [475, 1958]. The application to linear least squares problems is due to Golub [364, 1965], although Householder [475] discussed least squares.

Note that from the form of the decomposition (1.3.2) it follows immediately that R has the same singular values and right singular vectors as A. A relationship between the Cholesky factorization of A^TA and the QR decomposition of A is given next.

THEOREM 1.3.2. *Let $A \in \mathbf{R}^{m\times n}$ have rank n. Then if the R-factor in the QR decomposition of A has positive diagonal elements it equals the Cholesky factor of A^TA.*

Proof. If rank $(A) = n$, then by Theorem 1.1.4 the Cholesky factor of A^TA is unique. Now from (1.3.2) it follows that

$$A^TA = (R^T\ 0)Q^TQ\begin{pmatrix} R \\ 0 \end{pmatrix} = R^TR,$$

which concludes the proof. ∎

Assume that rank$(A) = n$, and partition Q in the form

$$Q = (Q_1, Q_2), \qquad Q_1 \in \mathbf{R}^{m\times n}, \quad Q_2 \in \mathbf{R}^{m\times(m-n)}. \tag{1.3.3}$$

Then by (1.3.2) and nonsingularity of R we have

$$A = (Q_1, Q_2)\begin{pmatrix} R \\ 0 \end{pmatrix} = Q_1R, \quad Q_1 = AR^{-1}. \tag{1.3.4}$$

Hence we can express Q_1 uniquely in terms of A and R. However the matrix Q_2 will not, in general, be uniquely determined.

From (1.3.4) it follows that

$$\mathcal{R}(A) = \mathcal{R}(Q_1), \quad \mathcal{R}(A)^\perp = \mathcal{R}(Q_2), \tag{1.3.5}$$

which shows that the columns of Q_1 and Q_2 form orthonormal bases for $\mathcal{R}(A)$ and its complement. It follows that the corresponding orthogonal projections are

$$P_{\mathcal{R}(A)} = Q_1Q_1^T, \quad P_{\mathcal{R}(A)^\perp} = Q_2Q_2^T. \tag{1.3.6}$$

We now show how to use the QR decomposition (1.3.2) to solve the augmented system (1.1.24). As shown in Theorem 1.1.5, this includes as special cases both the solution of the linear least squares problem ($b = 0$) and the minimum norm solution of an underdetermined system ($c = 0$).

THEOREM 1.3.3. *Let $A \in R^{m\times n}, m \geq n$, $b \in R^m$, and $c \in R^n$ be given. Assume that* rank$(A) = n$, *and let the QR decomposition of A be given by* (1.3.2). *Then the solution to the augmented system*

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix}\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

can be computed from

$$z = R^{-T}c, \qquad \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = Q^Tb, \tag{1.3.7}$$

$$x = R^{-1}(d_1 - z), \qquad y = Q\begin{pmatrix} z \\ d_2 \end{pmatrix}. \tag{1.3.8}$$

Proof. The augmented system can be written $y + Ax = b$, $A^Ty = c$, and using the factorization (1.3.2),

$$y + Q\begin{pmatrix} R \\ 0 \end{pmatrix} x = b, \quad \begin{pmatrix} R^T & 0 \end{pmatrix} Q^T y = c.$$

Multiplying the first equation with Q^T and the second with R^{-T} we get

$$Q^T y + \begin{pmatrix} R \\ 0 \end{pmatrix} x = Q^T b, \quad \begin{pmatrix} I_n & 0 \end{pmatrix} Q^T y = R^{-T} c.$$

Using the second equation to eliminate the first n components of $Q^T y$ in the first equation, we can solve for x. The last $m-n$ components of $Q^T y$ are obtained from the last $m-n$ equations in the first block. ∎

Taking $c = 0$ and $r = y = b - Ax$ in (1.3.7)-(1.3.8) it follows that the solution to the least squares problem $\min_x \|Ax - b\|_2$ is obtained from

$$\text{(1.3.9)} \qquad \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = Q^T b, \quad Rx = d_1, \quad r = Q\begin{pmatrix} 0 \\ d_2 \end{pmatrix}.$$

In particular, $\|r\|_2 = \|d_2\|_2$. Taking $b = 0$, we find that the solution to the problem $\min \|y\|_2$ such that $A^T y = c$ is obtained from

$$\text{(1.3.10)} \qquad R^T z = c, \qquad y = Q\begin{pmatrix} z \\ 0 \end{pmatrix}.$$

It follows that when A has full rank n the pseudoinverses of A and A^T are given by the expressions

$$\text{(1.3.11)} \qquad A^\dagger = R^{-1} Q_1^T, \qquad (A^T)^\dagger = Q_1 R^{-T}.$$

1.3.2. Rank revealing QR decompositions. According to Theorem 1.3.1 any matrix $A \in R^{m\times n}$ has a QR decomposition. However, as illustrated in the following example, if rank$(A) < n$, then the decomposition is not unique.

EXAMPLE 1.3.1. For any c and s such that $c^2 + s^2 = 1$ we have

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} 0 & s \\ 0 & c \end{pmatrix} = QR.$$

Here rank$(A) = 1 < 2 = n$. Note that the columns of Q no longer provide orthogonal bases for $R(A)$ and its complement. ∎

We now show how the QR decomposition can be modified for the case when rank$(A) < n$. (Note that this includes the case when $m < n$.)

THEOREM 1.3.4. *Given $A \in R^{m\times n}$ with* rank$(A) = r$ *there is a permutation matrix Π and an orthogonal matrix $Q \in R^{m\times m}$ such that*

$$\text{(1.3.12)} \qquad A\Pi = Q\begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix} \begin{matrix} \}r \\ \}m-r \end{matrix},$$

where $R_{11} \in R^{r\times r}$ is upper triangular with positive diagonal elements.

Proof. Since rank$(A) = r$, we can always choose a permutation matrix Π such that $A\Pi = (A_1, A_2)$, where $A_1 \in R^{m\times r}$ has linearly independent columns. Let

$$Q^T A_1 = \begin{pmatrix} R_{11} \\ 0 \end{pmatrix}, \qquad Q = (Q_1, Q_2),$$

be the QR decomposition of A_1, where $Q_1 \in \mathbf{R}^{m\times r}$. By Theorem 1.3.2, Q_1 and R_{11} are uniquely determined, and R_{11} has positive diagonal elements. Put

$$Q^T A\Pi = (Q^T A_1, Q^T A_2) = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}.$$

From rank$(Q^T A\Pi) =$ rank$(A) = r$ it follows that $R_{22} = 0$, since otherwise $Q^T A\Pi$ would have more than r linearly independent rows. Hence the decomposition must have the form (1.3.12). ∎

The decomposition (1.3.12) is not in general unique. Several strategies for determining a suitable column permutation Π are described in Section 2.7. When Π has been chosen, Q_1, R_{11}, and R_{12} are uniquely determined.

Also, when A has full column rank, but is close to a rank deficient matrix, QR decompositions with column permutations are of interest.

THEOREM 1.3.5. (See H. P. Hong and C. T. Pan [473, 1992].) *Let $A \in \mathbf{R}^{m\times n}$, $(m \geq n)$, and r be any integer $0 < r < n$. Then there exists a permutation matrix Π such that the QR factorization has the form*

$$A\Pi = Q\begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}, \tag{1.3.13}$$

with $R_{11} \in \mathbf{R}^{r\times r}$ upper triangular, $c = \sqrt{r(n-r) + \min(r, n-r)}$, and

$$\sigma_r(R_{11}) \geq \frac{1}{c}\sigma_r(A), \qquad \|R_{22}\|_2 \leq c\sigma_{r+1}(A). \tag{1.3.14}$$

A QR decomposition of the form (1.3.13)–(1.3.14) is called a **rank revealing** QR (RRQR) decomposition. If $\sigma_{r+1} = 0$ we recover the decomposition (1.3.12). Although the existence of a column permutation so that the corresponding QR decomposition satisfies (1.3.14) has been proved, it is still an open question if an algorithm of polynomial complexity exists for finding such a permutation; see Section 2.7.5. (Note that an exhaustive search for Π has combinatorial complexity!)

REMARK 1.3.1. From (1.3.13) it follows that

$$A\Pi \begin{pmatrix} R_{11}^{-1}R_{12} \\ -I \end{pmatrix} = Q\begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}\begin{pmatrix} R_{11}^{-1}R_{12} \\ -I \end{pmatrix} = Q\begin{pmatrix} 0 \\ R_{22} \end{pmatrix}.$$

Hence if $R_{22} = 0$, a dimensional argument shows that the nullspace of $A\Pi$ is given by

$$\mathcal{N}(A\Pi) = \mathcal{R}\begin{pmatrix} R_{11}^{-1}R_{12} \\ -I_{n-r} \end{pmatrix}. \tag{1.3.15}$$

1.3.3. The complete orthogonal decomposition.

For some applications it will be useful to carry the reduction in (1.3.12) one step further, using orthogonal transformations from the right as well. By performing a QR decomposition of the transpose of the triangular factor, the off-diagonal block can be eliminated:

$$\begin{pmatrix} R_{11}^T & 0 \\ R_{12}^T & 0 \end{pmatrix} = \hat{Q} \begin{pmatrix} \hat{R}_{11} & 0 \\ 0 & 0 \end{pmatrix}. \tag{1.3.16}$$

We then obtain a decomposition of the following form.

DEFINITION 1.3.1. *A* **complete orthogonal decomposition** *of $A \in R^{m\times n}$ with* rank $(A) = r$ *is a decomposition of the form*

$$A = Q \begin{pmatrix} T & 0 \\ 0 & 0 \end{pmatrix} V^T, \tag{1.3.17}$$

where $Q \in R^{m\times m}$ and $V \in R^{n\times n}$ are orthogonal matrices and $T \in R^{r\times r}$ is upper or lower triangular with positive diagonal elements.

Obviously, a decomposition of the form (1.3.17) is not unique. For example, the SVD of A is one example of a complete orthogonal decomposition. The form closest to the SVD that can be achieved by a finite algorithm is the **bidiagonal decomposition**

$$A = Q \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} V^T, \tag{1.3.18}$$

where B is a bidiagonal matrix with nonnegative diagonal elements; see Section 2.6.2.

From Theorem 1.2.12 it follows that the pseudoinverse of A can be expressed in terms of the decomposition (1.3.17) as

$$A^\dagger = V \begin{pmatrix} T^{-1} & 0 \\ 0 & 0 \end{pmatrix} Q^T. \tag{1.3.19}$$

Further, partitioning the orthogonal matrices in (1.3.17) by rows we have

$$A = (Q_1, Q_2) \begin{pmatrix} T & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}.$$

It follows that the complete orthogonal decomposition, like the SVD, provides orthogonal bases for the fundamental subspaces of A.

For many computational purposes the complete QR decomposition (1.3.17) is as useful as the SVD. An important advantage over the SVD is that the complete QR decomposition can be updated much more efficiently than the SVD when A is subject to a change of low rank; see Section 3.5. Different methods for computing the various QR decompositions are described in Sections 2.4 and 2.7.

1.4. Sensitivity of Least Squares Solutions

In this section we give results on the sensitivity of pseudoinverses and least squares solutions to perturbations in A and b. Many of the results below were first given by Wedin [824, 1973]. Stewart in [731, 1977] gives a unified treatment with interesting historical comments on the perturbation theory for pseudoinverses and least squares solutions. A more recent and excellent source of information is Stewart and Sun [754, 1990, Chap. 3].

1.4.1. Vector and matrix norms. In perturbation and error analyses it is useful to have a measure of the size of a vector or a matrix. Such measures are provided by vector and matrix norms, which can be regarded as generalizations of the absolute value.

A vector norm is a function $\|\cdot\| : \mathbf{C}^n \to \mathbf{R}$ that satisfies the following three conditions:

1. $\|x\| > 0 \quad \forall x \in \mathbf{C}^n, \quad x \neq 0$ (definiteness);
2. $\|\alpha x\| = |\alpha|\|x\| \quad \forall \alpha \in \mathbf{C}, \quad x \in \mathbf{C}^n$ (homogeneity);
3. $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathbf{C}^n$ (triangle inequality).

The most common vector norms are the Hölder p-norms

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}, \qquad 1 \leq p < \infty. \tag{1.4.1}$$

The l_p-norms have the property that $\|x\|_p = \|\,|x|\,\|_p$. Vector norms with this property are said to be **absolute**. The three most important particular cases are $p = 1, 2$ and the limit when $p \to \infty$:

$$\begin{aligned} \|x\|_1 &= |x_1| + \cdots + |x_n|, \\ \|x\|_2 &= (|x_1|^2 + \cdots + |x_n|^2)^{1/2} = (x^H x)^{1/2}, \\ \|x\|_\infty &= \max_{1 \leq i \leq n} |x_i|. \end{aligned} \tag{1.4.2}$$

The vector 2-norm is the Euclidean length of the vector, and is invariant under unitary transformations, i.e.,

$$\|Qx\|_2^2 = x^H Q^H Q x = x^H x = \|x\|_2^2$$

if Q is unitary. Another important property is the Hölder inequality

$$|x^H y| \leq \|x\|_p \|y\|_q, \qquad \frac{1}{p} + \frac{1}{q} = 1.$$

The special case with $p = q = 2$ is called the Cauchy–Schwarz inequality.

A matrix norm is a function $\|\cdot\| : \mathbf{C}^{m \times n} \to \mathbf{R}$ that satisfies analogues of the three vector norm properties. A matrix norm can be constructed from any vector norm by defining

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|. \tag{1.4.3}$$

This norm is called the matrix norm **subordinate** to the vector norm. From the definition it follows directly that

$$\|Ax\| \le \|A\| \, \|x\|, \qquad x \in \mathbf{C}^n.$$

It is an easy exercise to show that subordinate matrix norms are submultiplicative, i.e., whenever the product AB is defined it satisfies the condition $\|AB\| \le \|A\|\|B\|$.

The matrix norms subordinate to the vector p-norms are especially important. For these it holds that $\|I_n\|_p = 1$. Formulas for $\|A\|_p$ are known only for $p = 1, 2, \infty$. It can be shown that

$$(1.4.4) \qquad \|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|, \qquad \|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|,$$

respectively. Hence these norms are easily computable, and it holds that $\|A\|_1 = \|A^H\|_\infty$. The 2-norm, also called the **spectral norm**, is given by

$$(1.4.5) \qquad \|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \sigma_1(A).$$

Since the nonzero singular values of A and A^H are the same it follows that $\|A\|_2 = \|A^H\|_2$. The spectral norm is expensive to compute, but a useful upper bound is

$$(1.4.6) \qquad \|A\|_2 \le (\|A\|_1\|A\|_\infty)^{1/2}.$$

It is well known that on a finite-dimensional space two norms differ by at most a positive constant, which only depends on the dimension. For the vector p-norms it holds that

$$\|x\|_{p_2} \le \|x\|_{p_1} \le n^{\left(\frac{1}{p_1}-\frac{1}{p_2}\right)} \|x\|_{p_2}, \qquad p_1 \le p_2.$$

Another way to proceed in defining norms for matrices is to regard $\mathbf{C}^{m\times n}$ as an mn-dimensional vector space and apply a vector norm over that space. With the exception of the **Frobenius norm** derived from the vector 2-norm,

$$(1.4.7) \qquad \|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2},$$

such norms are not much used. Note that $\|A^H\|_F = \|A\|_F$. Useful alternative characterizations of the Frobenius norm are

$$(1.4.8) \qquad \|A\|_F^2 = \operatorname{trace}(A^H A) = \sum_{i=1}^{k} \sigma_i^2(A), \qquad k = \min(m, n).$$

The Frobenius norm is submultiplicative, but is often larger than necessary, e.g., $\|I\|_F = n^{1/2}$. This tends to make bounds derived in terms of the Frobenius norm not as sharp as they might be. From (1.4.8) we also get lower and upper bounds for the matrix 2-norm

$$\frac{1}{\sqrt{n}}\|A\|_F \leq \|A\|_2 \leq \|A\|_F.$$

An important property of the Frobenius norm and the 2-norm is that they are invariant with respect to orthogonal transformations, i.e., for all unitary matrices Q and P ($Q^HQ = I$ and $P^HP = I$) of appropriate dimensions we have $\|QAP^H\| = \|A\|$. We finally remark that the 1-,∞-, and Frobenius norms satisfy

$$\|\,|A|\,\| = \|A\|, \qquad |A| = (|a_{ij}|),$$

but for the 2-norm the best result is that $\|\,|A|\,\|_2 \leq n^{1/2}\|A\|_2$.

1.4.2. Perturbation analysis of pseudoinverses. We first give some perturbation bounds for the pseudoinverse. We consider a matrix $A \in \mathbf{R}^{m\times n}$ and let $B = A + E$ be the perturbed matrix. The theory is complicated by the fact that $A^\dagger$ varies discontinuously when the rank of A changes; cf. (1.2.24).

THEOREM 1.4.1. *If* rank $(A + E) \neq$ rank (A) *then*

$$\|(A+E)^\dagger - A^\dagger\|_2 \geq 1/\|E\|_2.$$

Proof. See Wedin [824, 1973]. ∎

EXAMPLE 1.4.1. By Theorem 1.4.1, when the rank changes the perturbation in $A^\dagger$ may be unbounded when $\|E\|_2 \to 0$. A trivial example of this is obtained by taking

$$A = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}, \quad E = \begin{pmatrix} 0 & 0 \\ 0 & \epsilon \end{pmatrix},$$

where $\sigma > 0$, $\epsilon \neq 0$. Then $1 = \text{rank}\,(A) \neq \text{rank}\,(A+E) = 2$,

$$A^\dagger = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \quad (A+E)^\dagger = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & \epsilon^{-1} \end{pmatrix},$$

and $\|(A+E)^\dagger - A^\dagger\|_2 = |\epsilon|^{-1} = 1/\|E\|_2$. ∎

In case the perturbation E does not change the rank of A, such unbounded growth of $(A+E)^\dagger$ cannot occur.

THEOREM 1.4.2. *If* rank $(A + E) =$ rank $(A) = r$, *and* $\eta = \|A^\dagger\|_2\|E\|_2 < 1$, *then*

$$\|(A+E)^\dagger\|_2 \leq \frac{1}{1-\eta}\|A^\dagger\|_2. \tag{1.4.9}$$

Proof. From the assumption and Theorem 1.2.7 it follows that

$$1/\|(A+E)^\dagger\|_2 = \sigma_r(A+E) \geq \sigma_r(A) - \|E\|_2 = 1/\|A^\dagger\|_2 - \|E\|_2 > 0,$$

which implies (1.4.9). ∎

We now characterize perturbations for which the pseudoinverse is well behaved. An **acute** perturbation of A is a perturbation such that the column and row spaces of A do not alter fundamentally.

DEFINITION 1.4.1. *The subspaces $\mathcal{R}(A)$ and $\mathcal{R}(B)$ are said to be acute if the corresponding orthogonal projections satisfy*

$$\|P_{\mathcal{R}(A)} - P_{\mathcal{R}(B)}\|_2 < 1.$$

Further, the matrix $B = A + E$ is said to be an acute perturbation of A if $\mathcal{R}(A)$ and $\mathcal{R}(B)$ a well as $\mathcal{R}(A^T)$ and $\mathcal{R}(B^T)$ are acute.

Acute perturbations can be characterized by the following theorem.

THEOREM 1.4.3. *The matrix B is an acute perturbation of A if and only if*

$$\text{rank}\,(A) = \text{rank}\,(B) = \text{rank}\,(P_{\mathcal{R}(A)} B P_{\mathcal{R}(A^T)}). \tag{1.4.10}$$

Proof. See Stewart [731, 1977]. ∎

Let A and $B = A + E$ be square nonsingular matrices. Then, from the well-known identity $B^{-1} - A^{-1} = -B^{-1}EA^{-1}$, it follows that

$$\|A^{-1} - B^{-1}\| \leq \|A^{-1}\|\,\|B^{-1}\|\,\|E\|.$$

The following generalization of this result can be proved by expressing the projections in terms of pseudoinverses using the relations in (1.2.29):

$$B^\dagger - A^\dagger = -B^\dagger E A^\dagger + (B^T B)^\dagger E^T P_{\mathcal{N}(A^T)} - P_{\mathcal{N}(B)} E^T (AA^T)^\dagger. \tag{1.4.11}$$

This identity can be used to obtain bounds for $\|B^\dagger - A^\dagger\|$ in the general case. For the case when $\text{rank}\,(B) = \text{rank}\,(A)$ the following theorem applies.

THEOREM 1.4.4. *If $B = A + E$ and $\text{rank}\,(B) = \text{rank}\,(A)$, then*

$$\|B^\dagger - A^\dagger\| \leq \mu \|B^\dagger\|\,\|A^\dagger\|\,\|E\| \tag{1.4.12}$$

where $\mu = 1$ for the Frobenius norm $\|\cdot\|_F$, and for the spectral norm $\|\cdot\|_2$,

$$\mu = \begin{cases} \frac{1+\sqrt{5}}{2} & \textit{if}\ \text{rank}\,(A) < \min(m,n), \\ \sqrt{2} & \textit{if}\ \text{rank}\,(A) = \min(m,n). \end{cases}$$

Proof. For the $\|\cdot\|_2$ norm, see Wedin [824, 1973]. The result that $\mu = 1$ for the Frobenius norm is due to van der Sluis and Veltkamp [784, 1979]. ∎

From the results above we deduce the following corollary.

COROLLARY 1.4.1. *A necessary and sufficient condition that*

$$\lim_{E \to 0} (A + E)^\dagger = A^\dagger \tag{1.4.13}$$

is that $\lim_{E\to 0} \text{rank}\,(A + E) = \text{rank}\,(A)$.

1.4.3. Perturbation analysis of least squares solutions.

We now consider the effect of perturbations of A and b upon the pseudoinverse solution $x = A^\dagger b$. In this analysis the **condition number** of a rectangular matrix $A \in \mathbf{R}^{m\times n}$ plays a significant role. The following definition generalizes the condition number of a square nonsingular matrix.

DEFINITION 1.4.2. *The condition number of $A \in \mathbf{R}^{m\times n}$ $(A \neq 0)$ is*

$$\kappa(A) = \|A\|_2\|A^\dagger\|_2 = \sigma_1/\sigma_r, \tag{1.4.14}$$

where $0 < r = \operatorname{rank}(A) \leq \min(m,n)$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, are the nonzero singular values of A.

The last equality in (1.4.14) follows from the relations $\|A\|_2 = \sigma_1$, $\|A^\dagger\|_2 = \sigma_r^{-1}$.

In the following we denote the perturbed A and b by

$$\tilde{A} = A + \delta A, \quad \tilde{b} = b + \delta b, \tag{1.4.15}$$

and the perturbed solution $\tilde{x} = \tilde{A}^\dagger\tilde{b} = x + \delta x$. We start by deriving the first-order perturbation estimate for least squares solutions when $\operatorname{rank}(A) = n$. Then if $\|\delta A\|_2 < \sigma_n$ we have $\operatorname{rank}(A + \delta A) = n$, and the perturbed solution $x + \delta x$ satisfies the normal equations

$$(A + \delta A)^T\Big((A + \delta A)(x + \delta x) - (b + \delta b)\Big) = 0.$$

Subtracting $A^T(Ax - b) = 0$, and neglecting second-order terms, we get

$$\delta x = A^\dagger(\delta b - \delta Ax) + (A^TA)^{-1}\delta A^T r, \tag{1.4.16}$$

where $A^\dagger = (A^TA)^{-1}A^T$ and $r = b - Ax$.

In the special case when only the right-hand side is perturbed, i.e., $\delta A = 0$, no second-order terms occur, and the exact perturbation equals

$$\delta x = A^\dagger\delta b = A^\dagger AA^\dagger\delta b = A^\dagger A_{\mathcal{R}(A)}\delta b = A^\dagger\delta b_1, \tag{1.4.17}$$

where we have split δb into orthogonal components

$$\delta b = \delta b_1 + \delta b_2, \qquad \delta b_1 = A_{\mathcal{R}(A)}\delta b, \qquad \delta b_2 = A_{\mathcal{N}(A^T)}\delta b.$$

Hence the perturbation δx depends only on the component of δb in $\mathcal{R}(A)$.

From the SVD of A we have $A^\dagger = V\Sigma^\dagger U^T$, $(A^TA)^{-1} = V(\Sigma^T\Sigma)^{-1}V^T$, and it follows that

$$\|(A^TA)^{-1}A^T\|_2 = \frac{1}{\sigma_n}, \quad \|(A^TA)^{-1}\|_2 = \frac{1}{\sigma_n^2}.$$

Using (1.4.17) and taking norms in (1.4.16) we obtain the first-order result

$$\|\delta x\|_2 \leq \frac{1}{\sigma_n}(\|\delta b_1\|_2 + \|\delta A\|_2\|x\|_2) + \frac{1}{\sigma_n^2}\|\delta A\|_2\|r\|_2. \tag{1.4.18}$$

Since $1/\sigma_n = \kappa(A)/\|A\|_2$ the last term here is proportional to $\kappa^2(A)$. Golub and Wilkinson [393, 1966] were the first to note that such a term occurs when $r \neq 0$. In van der Sluis [781, 1975] a geometrical explanation for the occurrence of this term is given, and lower bounds for the worst perturbation are also derived.

EXAMPLE 1.4.2. (See van der Sluis [781, 1975] and Figure 1.4.1.) Let $A = (a_1, a_2)$ be the matrix in Example 1.2.1, and assume that the angle

$\gamma = \arccos(a_1^T a_2)$ is small. Choose perturbations δa_1 and δa_2 of size $\|\delta a_1\|_2 = \|\delta a_2\|_2 = \epsilon$, so that the plane $\hat{S} = \text{span}(a_1+\delta a_1, a_2+\delta a_2)$ is obtained by rotation of the plane $S = \text{span}(a_1, a_2)$ around the bisector $u_1 = \frac{1}{2}(a_1 + a_2)$, which according to Example 1.2.1 is an approximate left singular vector. If δa_1 and δa_2 are orthogonal to S and of opposite direction, then the angle of rotation will be $\theta \approx \epsilon/(\frac{1}{2}\gamma)$. Now let $c = P_S b$ be the orthogonal projection of b onto S and assume that the approximate direction of c is along u_1. Then $\hat{c} = P_{\hat{S}} b$ is obtained by rotating the residual vector r through the angle θ and hence

$$\|\hat{c} - c\|_2 \approx \sin\theta \|r\|_2 \approx 2\epsilon\|r\|_2/\gamma.$$

Further, the direction of $\hat{c} - c$ will be approximately along $u_2 = \frac{1}{\gamma}(a_2 - a_1)$. Since $\delta a_1 + \delta a_2 = 0$ we have $\delta A x \approx 0$ and hence

$$A\delta x \approx \hat{c} - c, \qquad \|\delta x\|_2 \approx \frac{1}{\sigma_2}\|\hat{c} - c\|_2.$$

It follows that

$$\|\delta x\|_2 \approx \epsilon 2\sqrt{2}\|r\|_2/\gamma^2 = \frac{1}{\sqrt{2}} \cdot \epsilon \|r\|_2 \kappa^2,$$

which is what we wished to show. This example illustrates that the occurrence of κ^2 is due to two coinciding events: rotation of the projection plane around a dominant left singular vector produces a large change in r, and this has the direction of the minimal left singular vector. ∎

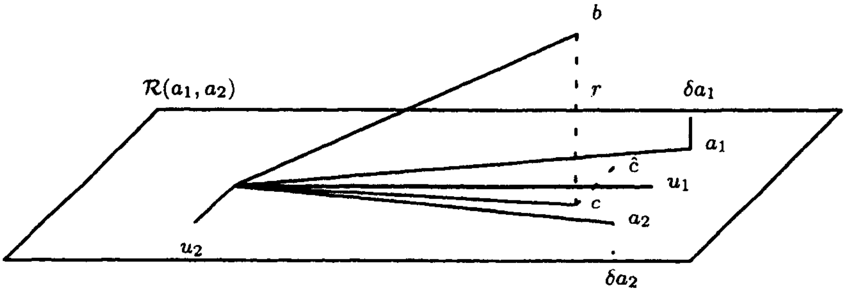

Fig. 1.4.1. *Exhibiting the squaring of* $\kappa(A)$.

We now give a more refined perturbation analysis, which follows that of Wedin [824, 1973] and applies to both overdetermined and underdetermined systems. In order to be able to prove any meaningful result we assume that the two conditions

$$\text{rank}\,(A + \delta A) = \text{rank}\,(A), \qquad \eta = \|A^\dagger\|_2 \|\delta A\|_2 < 1 \tag{1.4.19}$$

are satisfied. Note that if $\text{rank}\,(A) = \min(m, n)$ then the condition $\eta < 1$ suffices to guarantee that $\text{rank}\,(A + \delta A) = \text{rank}\,(A)$.

In the analysis we will need an estimate for the largest principal angle between the fundamental subspaces of $\tilde{A}$ and A. (For a definition of the principal angles between two subspaces see Definition 1.2.2.)

THEOREM 1.4.5. *Let $\tilde{A} = A + \delta A$ and assume that the conditions in* (1.4.19) *are satisfied. Then if $\chi(\cdot)$ denotes any of the four fundamental subspaces,*

$$\sin\theta_{\max}(\chi(\tilde{A}), \chi(A)) \le \eta < 1. \tag{1.4.20}$$

Proof. The result follows from Lemma 4.1 in Wedin [824, 1973]. ▮

We decompose the error δx as follows:

$$\delta x = \tilde{x} - x = \tilde{A}^\dagger \tilde{b} - x = \tilde{A}^\dagger(Ax + r + \delta b) - x,$$

or, using $P_{\mathcal{N}(\tilde{A})} = I - \tilde{A}^\dagger \tilde{A}$,

$$\delta x = \tilde{A}^\dagger P_{\mathcal{R}(\tilde{A})}(\delta b - \delta A x) + \tilde{A}^\dagger r - P_{\mathcal{N}(\tilde{A})} x. \tag{1.4.21}$$

We separately estimate each of the three terms in this decomposition of δx. Using Theorem 1.4.2 and the assumption (1.4.25) it follows that

$$\|\tilde{A}^\dagger(\delta b - \delta A x)\|_2 \le \frac{1}{1-\eta}\|A^\dagger\|_2 \cdot (\|\delta A\|_2 \|x\|_2 + \|\delta b\|_2). \tag{1.4.22}$$

(We remark that a sharper estimate can be obtained by substituting for $\|\delta b\|_2$ in (1.4.22) $\|\delta b_1\|_2 + \eta\|\delta b_2\|_2$.)

Since $r \perp \mathcal{R}(A)$ we have $r = P_{\mathcal{N}(A^T)} r$, and, using (1.2.29), we can write the second term as

$$\tilde{A}^\dagger r = \tilde{A}^\dagger \tilde{A} \tilde{A}^\dagger r = \tilde{A}^\dagger P_{\mathcal{R}(\tilde{A})} P_{\mathcal{N}(A^T)} r. \tag{1.4.23}$$

Now, by definition, $\|P_{\mathcal{R}(\tilde{A})} P_{\mathcal{N}(A^T)}\|_2 = \sin\theta_{\max}(\mathcal{R}(\tilde{A}), \mathcal{R}(A))$, where $\theta_{\max}$ is the largest principal angle between the two subspaces $\mathcal{R}(\tilde{A})$ and $\mathcal{R}(A)$.

Similarly, since $x = P_{\mathcal{R}(A^T)} x$, we can write the third term

$$P_{\mathcal{N}(\tilde{A})} x = P_{\mathcal{N}(\tilde{A})} P_{\mathcal{R}(A^T)} x \tag{1.4.24}$$

and

$$\|P_{\mathcal{N}(\tilde{A})} P_{\mathcal{R}(A^T)}\|_2 = \sin\theta_{\max}(\mathcal{N}(\tilde{A}), \mathcal{N}(A)).$$

Using Theorem 1.4.5 to estimate (1.4.23) and (1.4.24) we arrive at the following result.

THEOREM 1.4.6. *Assume that* $\operatorname{rank}(A + \delta A) = \operatorname{rank}(A)$, *and let*

$$\|\delta A\|_2 / \|A\|_2 \le \epsilon_A, \quad \|\delta b\|_2 / \|b\|_2 \le \epsilon_b. \tag{1.4.25}$$

Then if $\eta = \kappa\epsilon_A < 1$ the perturbations δx and δr in the least squares solution x and the residual $r = b - Ax$ satisfy

$$\|\delta x\|_2 \le \frac{\kappa}{1-\eta}\left(\epsilon_A \|x\|_2 + \epsilon_b \frac{\|b\|_2}{\|A\|_2} + \epsilon_A \kappa \frac{\|r\|_2}{\|A\|_2}\right) + \epsilon_A \kappa \|x\|_2 \tag{1.4.26}$$

and

$$\|\delta r\|_2 \le \epsilon_A \|x\|_2 \|A\|_2 + \epsilon_b \|b\|_2 + \epsilon_A \kappa \|r\|_2. \tag{1.4.27}$$

Proof. The estimate (1.4.26) follows from above, and (1.4.27) is proved using the decomposition

$$\begin{aligned} \tilde{r} - r &= P_{\mathcal{N}(\tilde{A}^T)}(b + \delta b) - P_{\mathcal{N}(A^T)}b \\ &= P_{\mathcal{N}(\tilde{A}^T)}\delta b + P_{\mathcal{N}(\tilde{A}^T)}P_{\mathcal{R}(A)}b - P_{\mathcal{R}(\tilde{A}^T)}P_{\mathcal{N}(A^T)}r \end{aligned}$$

and using $P_{\mathcal{N}(\tilde{A}^T)}P_{\mathcal{R}(A)}b = P_{\mathcal{N}(\tilde{A}^T)}Ax = -P_{\mathcal{N}(\tilde{A}^T)}\delta Ax$. ∎

REMARK 1.4.1. The last term in (1.4.26) (and therefore also in (1.4.27)) vanishes if rank$(A) = n$, since then $\mathcal{N}(\tilde{A}) = \{0\}$. If the system is compatible, e.g., if rank$(A) = m$, then $r = 0$ and the term involving κ^2 in (1.4.26) vanishes. For rank$(A) = n$, and $\epsilon_b = 0$, the condition number of the least squares problem can be written as

$$\kappa_{LS}(A, b) = \kappa(A)\left(1 + \kappa(A)\frac{\|r\|_2}{\|A\|_2\|x\|_2}\right). \tag{1.4.28}$$

Note that the conditioning depends on r and therefore on the right-hand side!

REMARK 1.4.2. When rank$(A) = n$ there are perturbations δA and δb such that the estimates in Theorem 1.4.6 can almost be attained for an arbitrary matrix A and vector b. This can be shown by considering first-order approximations of the terms (see Wedin [824, 1973]).

REMARK 1.4.3. It should be stressed that the perturbation analysis above is based on the class of perturbations defined by (1.4.25) and relevant only if the errors in the components of A and b are roughly of equal magnitude. For example, if the columns of A have widely different norms, then a more relevant class of perturbations often is

$$\tilde{a}_j = a_j + \delta a_j, \quad \|\delta a_j\|_2 \le \epsilon_j \|a_j\|_2, \quad j = 1, 2, \ldots, n. \tag{1.4.29}$$

Similarly, if the norm of the perturbation bounds for the *rows* in A differ widely, then (1.4.26) and (1.4.27) may considerably overestimate the perturbations. However, scaling the rows in (A, b) will change the least squares problem. ∎

A sharper estimate is usually obtained by scaling the columns of A so that the relative perturbation bound in all columns is the same:

$$Ax = (AD^{-1})(Dx) = \tilde{A}\tilde{x}, \quad D = \text{diag}(\epsilon_1\|a_1\|_2, \ldots, \epsilon_n\|a_n\|_2).$$

In particular, if $\epsilon_j = \epsilon$, $\forall j$, then this scales the matrix so that all column norms are equal. The following result by van der Sluis [780, 1969, Thm. 4.3] shows that this scaling approximately minimizes $\kappa(\tilde{A})$ over all diagonal scalings. (Note, however, that scaling the columns changes the norm in which the error in x is measured.)

THEOREM 1.4.7. *Let $C \in \mathbf{R}^{n\times n}$ be symmetric and positive definite, and denote by $\mathcal{D}$ the set of $n \times n$ nonsingular diagonal matrices. Then if in C all diagonal elements are equal, and C has at most q nonzero elements in any row, it holds that*

$$\kappa(C) \le q \min_{D\in\mathcal{D}} \kappa(DCD).$$

If this result is applied with $q = n$ to the matrix of normal equations A^TA, it follows that if all columns in A have unit length, then

$$\kappa(A) \leq \sqrt{n} \min_{D \in \mathcal{D}} \kappa(AD). \tag{1.4.30}$$

1.4.4. Asymptotic forms and derivatives. Derivatives of orthogonal projectors and pseudoinverses were first considered by Golub and Pereyra [378, 1973]. Stewart [731, 1977] gives asymptotic forms and derivatives for orthogonal projectors, pseudoinverses, and least squares solutions.

If $A = A(\tau)$ is differentiable and varies without changing rank, then the projection $P_{\mathcal{R}(A)}$ is differentiable and

$$\frac{dP_{\mathcal{R}(A)}}{d\tau} = P_{\mathcal{N}(A^T)} \frac{dA}{d\tau} P_{\mathcal{R}(A^T)} A^\dagger + (A^\dagger)^T P_{\mathcal{R}(A^T)} \frac{dA^T}{d\tau} P_{\mathcal{N}(A^T)}.$$

For the pseudoinverse,

$$\begin{aligned} \frac{dA^\dagger}{d\tau} &= -A^\dagger P_{\mathcal{R}(A)} \frac{dA}{d\tau} P_{\mathcal{R}(A^T)} A^\dagger + (A^TA)^\dagger P_{\mathcal{R}(A^T)} \frac{dA^T}{d\tau} P_{\mathcal{N}(A^T)} \\ & \quad -P_{\mathcal{N}(A)} \frac{dA^T}{d\tau} P_{\mathcal{R}(A)} (AA^T)^\dagger. \end{aligned}$$

Finally, for the least squares solution $x = A^\dagger b$, we obtain

$$\begin{aligned} \frac{dx}{d\tau} &= -A^\dagger P_{\mathcal{R}(A)} \frac{dA}{d\tau} P_{\mathcal{R}(A^T)} x + (A^TA)^\dagger P_{\mathcal{R}(A^T)} \frac{dA^T}{d\tau} P_{\mathcal{N}(A^T)} b \\ & \quad -P_{\mathcal{N}(A)} \frac{dA^T}{d\tau} P_{\mathcal{R}(A)} (A^T)^\dagger x. \end{aligned}$$

1.4.5. Componentwise perturbation analysis. There are several drawbacks with a normwise perturbation analysis. As already mentioned, it can give huge overestimates when the corresponding problem is badly scaled. Using norms we ignore how the perturbation is distributed among the elements of the matrix and vector. For these reasons *componentwise perturbation analysis* is of interest. An excellent survey of the theory and history behind such an analysis is given by Higham [466, 1994].

In this section we derive perturbation results and condition numbers corresponding to componentwise errors in A and b for the least squares problem. A similar analysis is given in Arioli et al. in [21, 1989]. We assume that we have componentwise bounds on the perturbations in the data

$$|\delta a_{ij}| \leq \omega e_{ij}, \quad |\delta b_i| \leq \omega f_i, \quad i, j = 1, \ldots, n,$$

where $e_{ij} > 0$ and $f_i > 0$ are known. In order to write such componentwise bounds in a simple way we define the absolute value of a matrix A and vector b by

$$|A|_{ij} = (|a_{ij}|), \qquad |b|_i = (|b_i|).$$

We introduce the partial ordering "$\leq$" for matrices A, B and vectors x, y, which is to be interpreted *componentwise*:

$$A \leq B \iff a_{ij} \leq b_{ij}, \quad x \leq y \iff x_i \leq y_i.$$

It is easy to show that if $C = AB$, then

$$|c_{ij}| \leq \sum_{k=1}^{n} |a_{ik}|\,|b_{kj}|,$$

and hence $|C| \leq |A|\,|B|$. A similar rule holds for matrix-vector multiplication.

With these notations we can write the componentwise bounds above as

$$|\delta A| \leq \omega E, \quad |\delta b| \leq \omega f, \tag{1.4.31}$$

where $E > 0$, $f > 0$. Taking $E = |A|$ and $f = |b|$ in (1.4.31) corresponds to **componentwise relative error bounds** for A and b.

We first derive estimates for the perturbations in the solution of a nonsingular square system $Ax = b$. The basic identity for this perturbation analysis is

$$\delta x = (I + A^{-1}\delta A)^{-1} A^{-1}(\delta A x + \delta b).$$

Assuming that $|A^{-1}||\delta A| < 1$, taking absolute values gives the inequality

$$|\delta x| \leq (I - |A^{-1}||\delta A|)^{-1}\,|A^{-1}|(|\delta A||x| + |\delta b|).$$

The matrix $(I - |A^{-1}||\delta A|)$ is guaranteed to be nonsingular if $\|\,|A^{-1}|\,|\delta A|\,\| < 1$. For perturbations satisfying (1.4.31) we obtain

$$|\delta x| \leq \omega(I - \omega|A^{-1}|E)^{-1}|A^{-1}|(E|x| + f). \tag{1.4.32}$$

Provided that $\omega\kappa_E(A) < 1$, $\kappa_E(A) = \|\,|A^{-1}|E\|$, we get from (1.4.32) for any absolute norm the perturbation bound

$$\|\delta x\| \leq \frac{\omega}{1 - \omega\kappa_E(A)}\|\,|A^{-1}|(|E\,|x| + f)\|. \tag{1.4.33}$$

For the special case of componentwise relative error bounds ($E = |A|$),

$$\kappa_{|A|}(A) = \|\,|A^{-1}||A|\,\| \tag{1.4.34}$$

is the **Bauer–Skeel condition number** of A (also denoted by cond$\,(A)$). It is possible for $\kappa_{|A|}(A)$ to be much smaller than $\kappa(A)$. It can be shown that $\kappa_{|A|}(A)$ and the bound (1.4.33) for $E = |A|$ are *invariant under row scalings.*

We now consider a componentwise error analysis for the linear least squares problem. For simplicity we will neglect error terms of order ω^2. From (1.4.16) we obtain

$$|\delta x| = \omega|A^{\dagger}|(f + E|x|) + \omega|(A^TA)^{-1}|E^T|r|, \tag{1.4.35}$$

and for the special case of componentwise relative perturbations,

$$|\delta x| = \omega|A^{\dagger}|(|x| + |A||x|) + \omega|(A^TA)^{-1}||A^T||r|.$$

It follows that

$$\|\delta x\|_\infty \leq \omega\Big(\|\,|A^{\dagger}|(|A||x| + |b|)\,\|_\infty + \|\,|(A^TA)^{-1}||A|^T|r|\,\|_\infty\Big). \tag{1.4.36}$$

Hence

$$\mathrm{cond}(A) = \|\,|A^{\dagger}|\,|A|\,\|_\infty \tag{1.4.37}$$

can be taken as an approximate condition number for componentwise relative perturbations. In the general case when $m > n$, cond(A) often depends only weakly on the row scaling D, but in a way which is complicated to describe. For stiff problems, where some rows are scaled with a large weight w, cond(A) usually tends to a limit value when $w \to \infty$, whereas $\kappa(A)$ grows linearly with w; see Björck [97, 1991].

1.4.6. A posteriori estimation of errors. Let $\bar{x}$ be an approximate solution of the least squares problem $\min_x \|Ax - b\|_2$, where $A \in \mathbf{R}^{m\times n}$, $m \geq n$. Consider the problem of finding the smallest perturbation E such that $\bar{x}$ *exactly* solves the problem $\min_x \|(A + E)x - b\|_2$. For a consistent linear system $Ax = b$ Rigal and Gaches [686, 1967] showed that the perturbation E of smallest l_2-norm is given by the rank one perturbation

$$E_0 = \bar{r}\bar{x}/\|\bar{x}\|_2^2 = \bar{r}\bar{x}^{\dagger}. \tag{1.4.38}$$

The corresponding norm $\|E\|_2 = \|\bar{r}\|_2\|\bar{x}\|_2$ is called the **normwise backward error**.

How to find the normwise backward error for an inconsistent least squares problem was an open problem for a long time. Stewart [733, 1977] showed that for the two perturbations

$$E_1 = -\bar{r}\bar{r}^{\dagger}A, \qquad E_2 = (r - \bar{r})\bar{x}^{\dagger}, \tag{1.4.39}$$

$\bar{x}$ solves the perturbed least squares problems exactly. The corresponding norms are

$$\|E_1\|_2 = \|A^T\bar{r}\|_2/\|\bar{r}\|_2, \qquad E_2 = \|r - \bar{r}\|_2/\|\bar{x}\|_2.$$

The first is small when the residual $\bar{r}$ is almost orthogonal to the column space of A. The second is small when $\bar{r}$ is almost equal to the exact residual r. However, it is possible for $\bar{x}$ to be a solution of a slightly perturbed least squares problem and yet for both $\|E_1\|_2$ and $\|E_2\|_2$ to be orders of magnitude larger than the norm of the perturbation.

Recently Waldén, Karlsson, and Sun [811, 1995] gave an explicit representation for the set $\mathcal{E}$ of all perturbation matrices E such that $\bar{x}$ exactly solves

$$\min_x \|(A + E)x - b\|_2.$$

They also found an expression for the $E \in \mathcal{E}$ which minimizes $\|E\|_F$. The corresponding solution when perturbations in both A and b are allowed is given in the following theorem.

THEOREM 1.4.8. *Let $A \in \mathbf{R}^{m\times n}$, $m \geq n$, $b \in \mathbf{R}^m$, and $\bar{x}$ be an approximate least squares solution. The normwise backward error*

$$\eta_F(\bar{x}) = \min\{\|(E, \tau e)\|_F \mid \|(A+E)x - (b+e)\|_2 = \min\}$$

is given by

$$\eta_F(\bar{x}) = \begin{cases} \gamma\sqrt{\mu}, & \lambda_* \geq 0, \\ \left(\gamma^2\mu + \lambda_*\right)^{1/2}, & \lambda_* < 0, \end{cases} \tag{1.4.40}$$

where $\bar{r} = b - A\bar{x}$, $\gamma = \|\bar{r}\|_2/\|\bar{x}\|_2$, and

$$\lambda_* = \lambda_{\min}\left(AA^T - \mu\frac{\bar{r}\bar{r}^T}{\|\bar{x}\|_2^2}\right), \qquad \mu = \frac{\tau^2\|\bar{x}\|_2^2}{1+\tau^2\|\bar{x}\|_2^2} \leq 1.$$

The parameter τ in Theorem 1.4.8 allows some flexibility. For example, taking the limit $\tau \to \infty$ gives the case when only A is perturbed. Then $\mu = 1$, and (1.4.40) becomes

$$\eta_F(\bar{x}) = \begin{cases} \gamma, & \lambda_* \geq 0, \\ \left(\gamma^2 + \lambda_*\right)^{1/2}, & \lambda_* < 0. \end{cases}$$

Note that the required backward error is no larger than the backward error $\|E_0\|_F$ for a consistent system, where E_0 is given by (1.4.38). It is strictly smaller if $\lambda_* < 0$. Note that a sufficient condition for $\lambda_* < 0$ is $\hat{r} \not\in \mathcal{R}(A)$.

The expressions for η_F in the theorem are elegant but unsuitable for computation since they can suffer from cancellation when $\lambda_* < 0$. Higham [467, 1996, Chap. 15] has suggested the alternative formula,

$$\eta_F(\bar{x}) = \min\left(\eta_1, \sigma_{\min}(A, \eta_1 B)\right),$$

where

$$\eta_1 = \gamma\sqrt{\mu}, \qquad B = (I - \bar{r}\bar{r}^T/\|\bar{r}\|_2^2).$$

This is more computationally reliable, but still expensive to compute. Simpler lower and upper bounds are given in Waldén, Karlsson, and Sun [811, 1995]. Optimal backward error bounds for linear least squares problems with multiple right-hand sides have been given by Sun [767, 1995].

Given an arbitrary approximate least squares solution $\bar{x}$ the **componentwise backward error** is the smallest $\omega \geq 0$ in

$$|\delta A| \leq \omega|A|, \qquad |\delta b| \leq \omega|b|,$$

such that $\bar{x}$ is the *exact* solution of the perturbed problem

$$\min_x \|(A+\delta A)x - (b+\delta b)\|_2.$$

For a consistent linear system $b \in \mathcal{R}(A)$, Oettli and Prager [603, 1964] showed that

$$\omega_B = \max_{1\leq i\leq n} \frac{|A\bar{x} - b|_i}{(|A||\bar{x}| + |b|)_i}, \tag{1.4.41}$$

where 0/0 should be interpreted as 0, and $\zeta/0$ ($\zeta \neq 0$) as infinity. (The latter case means that no finite ω satisfying (1.4.41) exists.) Together with the perturbation result (1.4.36) this can be used to compute an a posteriori bound on the error in a given approximate solution $\bar{x}$.

Unfortunately there is no similar result for the inconsistent linear least squares problem. One approach could be to apply the Oettli–Prager bound to the augmented system (1.1.24). Here there are no perturbations in the diagonal blocks of the augmented system matrix and in the last zero vector in the augmented right-hand side. However, a result by Kiełbasiński and Schwetlick [505, 1988, Lemma 8.2.11] shows that allowing *unsymmetric* perturbations in the blocks A and A^T has little effect on the backward error. Also, by a slight modification of the perturbation analysis in the previous section we can accommodate a perturbation to the first diagonal block; see also Higham [461, 1990]. Hence, for an a posteriori error analysis, it makes sense to take the relative backward error of a computed solution $\bar{r}, \bar{x}$ to be the smallest nonnegative number ω such that

$$|\delta I| \leq \omega I, \quad |\delta A_i| \leq \omega E, \quad i = 1, 2, \qquad |\delta b| \leq \omega f,$$

and

$$\begin{pmatrix} I + \delta I & A + \delta A_1 \\ A^T + \delta A_2 & 0 \end{pmatrix} \begin{pmatrix} \bar{r} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} b + \delta b \\ 0 \end{pmatrix}. \tag{1.4.42}$$

Using the result of Oettli and Prager, $\omega(\bar{r}, \bar{x}) = \max(\omega_1, \omega_2)$, where

$$\omega_1 = \max_{1 \leq i \leq m} \frac{|\bar{r} + A\bar{x} - b|_i}{(|\bar{r}| + E|\bar{x}| + f)_i}, \quad \omega_2 = \max_{1 \leq i \leq n} \frac{|A^T \bar{r}|_i}{(E^T |\bar{r}|)_i} \tag{1.4.43}$$

gives the backward error for a computed solution $\bar{r}$ *and* $\bar{x}$.

If we only have a computed $\bar{x}$ it may be feasible to put $\bar{r} = b - A\bar{x}$ and apply the result above. With this choice we have $\omega_1 = 0$ (exactly) and hence

$$\omega(\bar{r}, \bar{x}) = \omega_2 = \max_{1 \leq i \leq n} \frac{|A^T (b - A\bar{x})|_i}{(E^T |b - A\bar{x}|)_i}.$$

However, in the case of a nearly consistent least squares problem, $fl(b - A\bar{x})$ will mainly consist of roundoff, and will not be accurately orthogonal to the range of A. Hence, although $\bar{x}$ may have a small relative backward error, ω_2 will not be small. This illustrates a fundamental problem in computing the backward error: for $\bar{x}$ to have a small backward error it is sufficient that *either* $(b - A\bar{x})$ *or* $A^T(b - A\bar{x})$ be small, but neither of these conditions is *necessary*.