2

Normwise Condition of Linear Equation Solving

The QR factorization is one of the main engines in numerical linear algebra. The following result, a backward analysis for linear equation solving using it, is a particular case of Theorem 19.3 of [12].

Theorem 2.1. *Let $A \in \mathbb{R}^{n\times n}$ be invertible and $b \in \mathbb{R}^n$. If the system $Ax = b$ is solved using the Householder QR factorization the computed solution $\tilde{x}$ satisfies*

$$\tilde{A}\tilde{x} = \tilde{b}$$

where $\tilde{A}$ and $\tilde{b}$ satisfy the relative error bounds

$$\|\tilde{A} - A\|_F \le n\gamma_{cn}\|A\|_F \quad \text{and} \quad \|\tilde{b} - b\| \le n\gamma_{cn}\|b\|$$

for a small constant c and with γ_{cn} as defined in (1.5). □

It follows from this backward stability result, (1.6), and Theorem 1.3 that the relative error for the computed solution $\tilde{x}$ satisfies

$$\frac{\|\tilde{x} - x\|}{\|x\|} \le cn^2\epsilon_{\mathsf{mach}}\mathsf{cond}(A, b) + o(\epsilon_{\mathsf{mach}}) \tag{2.1}$$

and the loss of precision is bounded by

$$\mathsf{LoP}(A^{-1}b) \le 2\log_\beta n + \log_\beta \mathsf{cond}(A, b) + \log_\beta c + o(1), \tag{2.2}$$

where $\mathsf{cond}(A, b)$ is the normwise condition number for linear equation solving

$$\mathsf{cond}(A, b) = \lim_{\delta\to 0} \sup_{\max\{\mathsf{RelError}(A),\mathsf{RelError}(b)\}\le\delta} \frac{\mathsf{RelError}(A^{-1}b)}{\delta}.$$

Inequality (2.1) calls for an understanding of what $\mathsf{cond}(A, b)$ is deeper than the equality above. The pursuit of this understanding is the goal of this chapter.

2.1 Turing's Condition Number

The condition number $\mathsf{cond}(A, b)$ in the introduction is a normwise one. For this reason, we begin by providing a brief review of norms (for a more detailed treatment we refer the reader to Higham [12, Chapter 6]).

The three most useful norms in error analysis on the real vector space $\mathbb{R}^n$ are the following:

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 := \left(\sum_{i=1}^{n} |x_i| \right)^{1/2}, \quad \|x\|_\infty := \max_{1 \le i \le n} |x_i|.$$

These are special cases of the Hölder r-norm

$$\|x\|_r := \left(\sum_{i=1}^{n} |x_i|^r \right)^{1/r}$$

defined for a real number $r \geq 1$. Even though we will only need the cases $r \in \{1, 2, \infty\}$, stating the results for general Hölder norms avoids case distinctions and thus saves space.

For a given $r \geq 1$ there is exactly one $r^* \geq 1$ such that $1/r + 1/r^* = 1$. The well-known Hölder inequality states that for $x, z \in \mathbb{R}^n$ we have

$$|x^{\mathrm{T}} z| \leq \|x\|_r \, \|z\|_{r^*}.$$

Moreover, equality holds if $(|x_i|^r)$ and $(|z_i|^{r^*})$ are linearly dependent. This easily implies that for any $x \in \mathbb{R}^n$

$$\max_{\|z\|_{r^*}=1} x^{\mathrm{T}} z = \|x\|_r. \tag{2.3}$$

For this reason, one calls $\| \ \|_{r^*}$ the *dual norm* of $\| \ \|_r$. In particular, for each $x \in \mathbb{R}^n$ with $\|x\|_r = 1$ there exists $z \in \mathbb{R}^n$ such that $\|z\|_{r^*} = 1$ and $z^{\mathrm{T}} x = 1$.

We will adopt the notational convention $\| \ \| := \| \ \|_2$ for the Euclidean vector norm. Note that this norm is dual to itself. Note as well that $\| \ \|_1$ and $\| \ \|_\infty$ are dual to each other.

To the vector norms $\| \ \|_r$ on a domain space $\mathbb{R}^n$ and $\| \ \|_s$ on a range space $\mathbb{R}^m$, one associates the *subordinate matrix norm* $\| \ \|_{rs}$ on the vector space of linear operators $A \colon \mathbb{R}^n \to \mathbb{R}^m$ defined by

$$\|A\|_{rs} := \sup_{\substack{x \in \mathbb{R}^p \\ x \neq 0}} \frac{\|Ax\|_s}{\|x\|_r} = \sup_{\|x\|_r = 1} \|Ax\|_s. \tag{2.4}$$

By compactness of the unit sphere the supremum is a minimum. In case $r = s$ we write $\| \ \|_r$ instead of $\| \ \|_{rr}$. It is easy to show that

$$\|A\|_1 = \max_j \sum_{i=1}^{m} |a_{ij}|, \quad \|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

(We recall that we already met $\| \ \|_\infty$ in Section 1.4.) Furthermore, when $r = 2$, $\| \ \|_2$ is called the *spectral norm* and it is written simply as $\| \ \|$.

We note that the following submultiplicativity property of matrix norms holds: for $r, s, t \geq 1$ and matrices A, B we have

$$\|AB\|_{rs} \leq \|A\|_{ts} \, \|B\|_{rt} \tag{2.5}$$

provided the matrix product is defined.

Most of what we will need about operator norms is stated in the following simple lemma.

Lemma 2.2. *1. For $y \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ we have $\|yv^{\mathrm{T}}\|_{rs} = \|y\|_s \, \|v\|_{r^*}$.*

2. Suppose that $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ satisfy $\|x\|_r = \|y\|_s = 1$. Then there exists $B \in \mathbb{R}^{m \times n}$ such that $\|B\|_{rs} = 1$ and $Bx = y$.

Proof. (1) We have

$$\|yv^{\mathrm{T}}\|_{rs} = \max_{\|x\|_r = 1} \|yv^{\mathrm{T}}x\|_s = \|y\|_s \max_{\|x\|_r = 1} |v^{\mathrm{T}}x| = \|y\|_s \|v\|_{r^*},$$

where the last equality holds due to (2.3).

(2) By (2.3) there exists $z \in \mathbb{R}^n$ such that $\|z\|_{r^*} = 1$ and $z^{\mathrm{T}}x = 1$. For $B := yz^{\mathrm{T}}$ we have $Bx = y$ and

$$\|B\|_{rs} = \max_{\|x'\|_r = 1} \|yz^{\mathrm{T}}x'\|_s = \|y\|_s \max_{\|x'\|_r = 1} |z^{\mathrm{T}}x'| = \|y\|_s \, \|z\|_{r^*} = 1,$$

where we again used (2.3) for the second last equality. □

We now proceed to exhibit a characterization of the normwise condition number for linear equation solving, pursuing the theme described in §1.5.2.

Let $m = n$ and fix norms $\| \ \|_r$ and $\| \ \|_s$ on $\mathbb{R}^n$. Also, let

$$\Sigma := \{A \in \mathbb{R}^{n \times n} \mid \det(A) = 0\}$$

denote the *set of ill-posed matrices* and put $\mathcal{D} := \mathbb{R}^{n \times n} \setminus \Sigma$. We define the map $\kappa_{rs} : \mathcal{D} \to \mathbb{R}$ by

$$\kappa_{rs}(A) := \|A\|_{rs} \|A^{-1}\|_{sr}.$$

Note that $\kappa_{rs}(A) \geq 1$, since $1 = \|\mathrm{I}\|_r \leq \|A\|_{rs} \|A^{-1}\|_{sr} = \kappa_{rs}(A)$.

Theorem 2.3. *Let $\varphi : \mathcal{D} \times \mathbb{R}^n \to \mathbb{R}^n$ be given by $\varphi(A, b) = A^{-1}b$. We measure the relative error in $\mathcal{D} \times \mathbb{R}^n$ by*

$$\mathsf{RelError}(A, b) = \max \left\{ \frac{\|\tilde{A} - A\|_{rs}}{\|A\|_{rs}}, \frac{\|\tilde{b} - b\|_s}{\|b\|_s} \right\}$$

and we measure the relative error in the solution space normwise with respect to $\| \ \|_r$. *Then*

$$\mathsf{cond}^{\varphi}(A,b) = \kappa_{rs}(A) + \frac{\|A^{-1}\|_{sr}\|b\|_s}{\|A^{-1}b\|_r}.$$

In particular, we have

$$\kappa_{rs}(A) \leq \mathsf{cond}^{\varphi}(A,b) \leq 2\kappa_{rs}(A).$$

Proof. Let $\tilde{A} = A - E$ and $\tilde{b} = b + f$. By definition, $\|E\|_{rs} \leq \mathsf{R}\|A\|_{rs}$ and $\|f\|_s \leq \mathsf{R}\|b\|_s$ where, for simplicity, $\mathsf{R} = \mathsf{RelError}(A,b)$. We have, for $\mathsf{R} \to 0$,

$$(A-E)^{-1} = A^{-1}(\mathrm{I}-EA^{-1})^{-1} = A^{-1}\big(\mathrm{I}+EA^{-1}+o(\mathsf{R})\big) = A^{-1}+A^{-1}EA^{-1}+o(\mathsf{R}).$$

This implies, writing $x := A^{-1}b$ and $\tilde{x} := \tilde{A}^{-1}\tilde{b}$,

$$\tilde{x} - x = (A-E)^{-1}(b+f) - x = A^{-1}Ex + A^{-1}f + o(\mathsf{R}). \tag{2.6}$$

Taking norms and using (2.5) we conclude

$$\begin{aligned} \|\tilde{x} - x\|_r &\leq \|A^{-1}\|_{sr}\|E\|_{rs}\|x\|_r + \|A^{-1}\|_{sr}\|f\|_s + o(\mathsf{R}) \\ &\leq \|A^{-1}\|_{sr}\|A\|_{rs}\|x\|_r\mathsf{R} + \|A^{-1}\|_{sr}\|b\|_s\mathsf{R} + o(\mathsf{R}), \end{aligned}$$

hence

$$\frac{\|\tilde{x} - x\|_r}{\mathsf{R}\|x\|_r} \leq \kappa_{rs}(A) + \frac{\|A^{-1}\|_{sr}\|b\|_s}{\|x\|_r},$$

which shows the upper bound in the claimed equality.

For the corresponding lower bound we choose $y \in \mathbb{R}^m$ such that $\|y\|_s = 1$ and $\|A^{-1}y\|_r = \|A^{-1}\|_{sr}$. Further, we choose $v \in \mathbb{R}^n$ such that $\|v\|_{r^*} = 1$ and $v^{\mathrm{T}}x = \|x\|_r$, which is possible by (2.3). Now we put

$$E := \mathsf{R}\|A\|_{rs}\, yv^{\mathrm{T}}, \qquad f := \pm\mathsf{R}\|b\|_s\, y. \tag{2.7}$$

We note that

$$\|E\|_{rs} = \mathsf{R}\|A\|_{rs}, \qquad \|f\|_s = \mathsf{R}\|b\|_s,$$

the first equality since, by Lemma 2.2(1), $\|yv^{\mathrm{T}}\|_{rs} = \|y\|_s\|v\|_{r^*} = 1$. We have

$$A^{-1}Ex = \mathsf{R}\|A\|_{rs}\, A^{-1}y\, v^{\mathrm{T}}x = \mathsf{R}\|A\|_{rs}\|x\|_r A^{-1}y$$

and hence $\|A^{-1}Ex\|_r = \kappa_{rs}(A)\|x\|_r\mathsf{R}$. Similarly, $A^{-1}f = \pm R\|b\|_s A^{-1}y$ and $\|A^{-1}f\|_r = \|A^{-1}\|_{sr}\|b\|_s\mathsf{R}$. Since $A^{-1}Ex$ and $A^{-1}f$ are both proportional to $A^{-1}y$, we obtain from (2.6)

$$\|\tilde{x} - x\|_r = \kappa_{rs}(A)\|x\|_r\mathsf{R} + \|A^{-1}\|_{sr}\|b\|_s\mathsf{R}$$

when choosing the sign for f in (2.7) appropriately. This proves the claimed lower bound. □

Theorem 2.3 —together with (2.1)— immediately yields a bound for the loss of precision in linear equation solving.

Corollary 2.4. *Let $A \in \mathbb{R}^{n\times n}$ be invertible and $b \in \mathbb{R}^n$. If the system $Ax = b$ is solved using the Householder QR factorization, then the computed solution $\tilde{x}$ satisfies, for a small constant c,*

$$\mathsf{LoP}(A^{-1}b) \leq 2\log_\beta n + \log_\beta \kappa_{rs}(A) + \log_\beta c + o(1)$$

where $o(1)$ is for $\epsilon_{\mathsf{mach}} \to 0$. □

The next result shows that κ_{rs} actually coincides with the condition number for the problem of matrix inversion.

Theorem 2.5. *Let $\psi\colon \mathcal{D} \to \mathbb{R}^{n\times n}$ be given by $\psi(A) = A^{-1}$. We measure the relative error on the data space and solution space with respect to $\|\ \|_{rs}$ and $\|\ \|_{sr}$, respectively. Then we have*

$$\mathsf{cond}^{\psi}(A) = \kappa_{rs}(A).$$

Proof. Let $E \in \mathbb{R}^{n\times n}$ be such that $\tilde{A} = A - E$. Then $\mathsf{RelError}(A) = \frac{\|E\|_{rs}}{\|A\|_{rs}}$. As in the proof of Theorem 2.3 we have for $\|E\| \to 0$,

$$\|\tilde{A}^{-1} - A^{-1}\|_{sr} = \|A^{-1}EA^{-1}\|_{sr} + o\big(\|E\|\big). \tag{2.8}$$

Hence, $\|A^{-1}EA^{-1}\|_{sr} \leq \|A^{-1}\|_{sr}\|E\|_{rs}A^{-1}\|_{sr}$. Consequently, we obtain

$$\mathsf{RelError}(A^{-1}) = \frac{\|\tilde{A}^{-1} - A^{-1}\|_{sr}}{\|A^{-1}\|_{sr}} \leq \|A^{-1}\|_{sr}\|E\|_{rs} + o\big(\|E\|\big).$$

We conclude that

$$\frac{\mathsf{RelError}(A^{-1})}{\mathsf{RelError}(A)} \leq \|A\|_{rs}\|A^{-1}\|_{sr} + o(1)$$

and hence $\mathsf{cond}^{\psi}(A) \leq \kappa_{rs}(A)$.

To prove the reversed inequality it is enough to find arbitrary small matrices E such that $\|A^{-1}EA^{-1}\|_{sr} = \|A^{-1}\|_{sr}^2\|E\|_{rs}$ since then we can proceed from (2.8) as we did in Theorem 2.3 from(2.6).

To do so let $y \in \mathbb{R}^n$ be such that $\|y\|_s = 1$ and $\|A^{-1}y\|_r = \|A^{-1}\|_{sr}$. Define $x := \frac{1}{\|A^{-1}\|_{sr}}A^{-1}y$ so that $A^{-1}y = \|A^{-1}\|_{sr}x$ and $\|x\|_r = \|y\|_s = 1$. For any $B \in \mathbb{R}^{n\times n}$ we have

$$\|A^{-1}BA^{-1}\|_{sr} \ \geq\ \|A^{-1}BA^{-1}y\|_r \ =\ \|A^{-1}\|_{sr} \cdot \|A^{-1}Bx\|_r.$$

By Lemma 2.2(2) there exists $B \in \mathbb{R}^{n\times n}$ such that $Bx = y$ and $\|B\|_{rs} = 1$. Therefore,

$$\|A^{-1}BA^{-1}\|_{sr} \ \geq\ \|A^{-1}\|_{sr} \cdot \|A^{-1}y\|_r \ =\ \|A^{-1}\|_{sr}^2.$$

Taking $E = \delta B$ with arbitrarily small δ finishes the proof. □

The most often considered case is $r = s = 2$, that is, when measuring the error in both input and output space with the Euclidean norm. The resulting condition number $\kappa(A) := \kappa_{22}(A)$ is so pervasive in numerical linear algebra that it is commonly referred to as "the condition number of A" —without mention to the function for which we want to measure condition. We remark that $\kappa(A)$ was originally introduced by Turing [19] and by von Neumann and Goldstine [21] (Turing actually considered norms other than the spectral).

2.2 Condition and Distance to Ill-posedness

A goal of this section, now revisiting the discussion in §1.5.4, is to show that the condition number $\kappa_{rs}(A)$ can be expressed as the relativized inverse of the distance from the square matrix A to the set Σ of singular matrices: a large $\kappa_{rs}(A)$ means that A is close to a singular matrix. In order to make this precise we introduce the distance of $A \in \mathbb{R}^{n\times n}$ to the set Σ of singular matrices

$$d_{rs}(A, \Sigma) \;:=\; \min\{\|A - B\|_{rs} \mid B \in \Sigma\} \tag{2.9}$$

defined with respect to the norm $\|\ \|_{rs}$. For the spectral norm we just write $d(A, \Sigma) := d_{22}(A, \Sigma)$.

The following result was proved by Kahan, who attributes it to Gastinel (cf. Higham [12, Thm.6.5]). For the special case of spectral norms, this fundamental result had been obtained much earlier, in 1936, by Eckart and Young [6].

Theorem 2.6. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. Then*

$$d_{rs}(A, \Sigma) \;=\; \frac{1}{\|A^{-1}\|_{sr}}.$$

Proof. Let A be nonsingular and $A + E$ be singular. Then there exists an $x \in \mathbb{R}^n \setminus \{0\}$ such that $(A+E)x = 0$. This means that $x = -A^{-1}Ex$ and hence

$$\|x\|_r \le \|A^{-1}E\|_{rr} \cdot \|x\|_r \le \|A^{-1}\|_{sr} \cdot \|E\|_{rs} \cdot \|x\|_r,$$

which implies $\|E\|_{rs} \ge \|A^{-1}\|_{sr}^{-1}$. Therefore $d_{rs}(A, \Sigma) \ge \|A^{-1}\|_{sr}^{-1}$.

To show the other inequality it suffices to find a singular matrix $\tilde{A}$ with $d_{rs}(A, \tilde{A}) \le \|A^{-1}\|_{sr}^{-1}$. Let $y \in \mathbb{R}^n$ be such that $\|A^{-1}\|_{sr} = \|A^{-1}y\|_r$ and $\|y\|_s = 1$. Writing $x := A^{-1}y$ we have $\|x\|_r = \|A^{-1}\|_{sr}$, in particular $x \neq 0$. By Lemma 2.2(2) there exists $B \in \mathbb{R}^{n\times n}$ such that $\|B\|_{rs} = 1$ and

$$B\,\frac{x}{\|x\|_r} = -y.$$

Hence $E := \|x\|_r^{-1}B$ satisfies $Ex = -y$ and hence $(A + E)x = 0$. So the matrix $\tilde{A} := A + E$ must be singular. In addition, we have

$$d_{rs}(A, \tilde{A}) = \|E\|_{rs} = \|x\|_r^{-1}\|B\|_{rs} = \|A^{-1}\|_{sr}^{-1} \cdot \|B\|_{rs} = \|A^{-1}\|_{sr}^{-1},$$

which finishes the proof. □

Defining $\kappa_{rs}(A) := \infty$ for a singular matrix, we immediately obtain the following result which is known as the "Condition Number Theorem."

Corollary 2.7. *For nonzero $A \in \mathbb{R}^{n\times n}$ we have*

$$\kappa_{rs}(A) \;=\; \frac{\|A\|_{rs}}{d_{rs}(A,\Sigma)}. \qquad \square$$

Thus the condition number $\kappa_{rs}(A)$ can be seen as the inverse of a normalized distance of A to the set of ill-posed inputs Σ.

Notation. In this book we will consider matrices both as given by its columns or by its rows. In order to emphasize this distinction, and avoid ambiguities, given vectors $a_1, \ldots, a_n \in \mathbb{R}^m$ we write $(a_1, \ldots, a_n)$ for the matrix in $\mathbb{R}^{n\times m}$ whose rows are $a_1, \ldots, a_n$ and $[a_1, \ldots, a_n]$ for the matrix in $\mathbb{R}^{m\times n}$ whose columns are these vectors. Note that this notation dispense us with transposing $(x_1, \ldots, x_n)$ when we want to emphasize that this is a column vector.

For a matrix $A \in \mathbb{R}^{n\times m}$, a vector $c \in \mathbb{R}^n$, and an index $j \in [m]$, we denote by $A(j : c)$ the matrix obtained by replacing the jth row of A by c. The meaning of $A[j : c]$ is defined similarly.

We draw now a consequence of Theorem 2.6 that will be used in several variations throughout the book.

Proposition 2.8. *For $A \in \mathbb{R}^{n\times n}$ and $r, s \geq 1$ there exists $j \in [n]$ and $c \in \mathbb{R}^n$ such that $A[j : c] \in \Sigma$ and $\|a_j - c\|_s \leq n^{1/r}\, d_{rs}(A,\Sigma)$.*

Proof. Theorem 2.6 states that $\|A^{-1}\|_{sr} = \epsilon^{-1}$, when writing $\epsilon := d_{rs}(A,\Sigma)$. There exists $b \in \mathbb{R}^n$ such that $\|b\|_s = 1$ and $\|A^{-1}b\|_r = \|A^{-1}\|_{sr}$. So if we put $v := A^{-1}b$, then $\|v\|_r \geq \epsilon^{-1}$. This implies $\|v\|_\infty \geq n^{-1/r}\|v\|_r \geq n^{-1/r}\epsilon^{-1}$. W.l.o.g. we may asssume that $|v_n| = \|v\|_\infty$.

Since $Av = b$, we can express v_n by Cramer's rule as follows

$$v_n \;=\; \frac{\det([a_1, \ldots, a_{n-1}, b])}{\det(A)}.$$

This implies

$$0 \;=\; \det(A) - v_n^{-1}\det([a_1, \ldots, a_{n-1}, b]) \;=\; \det\left([a_1, \ldots, a_{n-1}, a_n - v_n^{-1}b]\right).$$

Thus if we put $c := a_n - v_n^{-1}b$ we have $A[i : c] \in \Sigma$ and

$$\|a_n - c\|_s = |v_n|^{-1}\|b\|_s = |v_n|^{-1} \leq n^{1/r}\,\epsilon. \qquad \square$$

2.3 The Singular Value Decomposition

The singular value decomposition of a matrix is the numerically appropriate way to discuss matrix rank. It also leads to a natural generalization of Theorem 2.6.

In this section we mainly work with the spectral norm and the *Frobenius norm* of a matrix $A = (a_{ij}) \in \mathbb{R}^{m\times n}$, which is defined as

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

Clearly, $\|A\|_F$ equals the Euclidean norm of A when interpreted as an element of $\mathbb{R}^{mn}$. The advantage of the Frobenius norm is that it is induced by an inner product on $\mathbb{R}^{m\times n}$.

Both the spectral norm and the Frobenius norm are invariant under orthogonal transformations.

Lemma 2.9. *For $A \in \mathbb{R}^{m\times n}$ and orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ we have $\|UAV\|_F = \|A\|_F$ and $\|UAV\| = \|A\|$.*

Proof. For the first assertion let $s_1, \ldots, s_n$ denote the columns of A. Then Us_i is the ith column of UA. Since U is orthogonal we have $\|Us_i\| = \|s_i\|$ and therefore

$$\|UA\|_F^2 = \sum_{i\le n} \|Us_i\|^2 = \sum_{i\le n} \|s_i\|^2 = \|A\|_F^2.$$

In the same way one shows that $\|AV\|_F = \|A\|_F$. The second assertion is shown as follows

$$\begin{aligned}\|UAV\| &= \sup_{\|x\|=1} \|UAVx\| = \sup_{\|x\|=1} \|U(AVx)\| \\ &= \sup_{\|x\|=1} \|AVx\| = \sup_{\|x\|=1} \|A(Vx)\| \\ &= \sup_{\|x'\|=1} \|Ax'\| = \|A\|.\end{aligned}$$

□

For conveniently stating the singular value decomposition, we extend the usual notation for diagonal matrices from square to rectangular $m \times n$-matrices. We put $p := \min\{n, m\}$ and define, for $a_1, \ldots, a_p \in \mathbb{R}$,

$$\mathsf{diag}_{m,n}(a_1, \ldots, a_p) := (b_{ij}) \in \mathbb{R}^{m\times n} \quad \text{with} \quad b_{ij} := \begin{cases} a_i & \text{if } i = j \\ 0 & \text{otherwise.}\end{cases}$$

For notational convenience we usually drop the index, the format being clear from the context.

The next result is known as the "Singular Value Decomposition Theorem" (or, in short, the "SVD Theorem").

Theorem 2.10. *For $A \in \mathbb{R}^{m\times n}$ there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that*

$$U^{\mathrm{T}} A V \;=\; \mathsf{diag}(\sigma_1, \ldots, \sigma_p),$$

with $p = \min\{m, n\}$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$.

Proof. Let $x \in \mathbb{R}^n$, $\|x\| = 1$ be such that $\sigma := \|A\| = \|Ax\|$ and define $y := \sigma^{-1}Ax \in \mathbb{R}^m$, so that $\|y\| = 1$ and $Ax = \sigma y$. There exist matrices $V_2 \in \mathbb{R}^{n\times(n-1)}$ and $U_2 \in \mathbb{R}^{m\times(m-1)}$ such that $V := [x, V_2]$ and $U := [y, U_2]$ are orthogonal.

We have for some $w \in \mathbb{R}^{n-1}$ and $B \in \mathbb{R}^{(m-1)\times(n-1)}$ that

$$\begin{aligned} U^{\mathrm{T}} A V &= \begin{bmatrix} y^{\mathrm{T}} \\ U_2^{\mathrm{T}} \end{bmatrix} A [x, V_2] = \begin{bmatrix} y^{\mathrm{T}} \\ U_2^{\mathrm{T}} \end{bmatrix} [\sigma y, AV_2] \\ &= \begin{bmatrix} \sigma & w^{\mathrm{T}} \\ 0 & B \end{bmatrix} =: A_1. \end{aligned}$$

Note that $\|A_1\| = \|A\|$ by the orthogonal invariance of the spectral norm. Moreover, we have for $v \in \mathbb{R}^{n-1}$

$$\|Bv\| = \left\| \begin{bmatrix} \sigma & w^{\mathrm{T}} \\ 0 & B \end{bmatrix} \begin{bmatrix} 0 \\ v \end{bmatrix} \right\| = \left\| U^{\mathrm{T}} A V \begin{bmatrix} 0 \\ v \end{bmatrix} \right\| \leq \|U^{\mathrm{T}} A V\| \, \|v\| \leq \|A\| \, \|v\|,$$

hence $\|B\| \leq \|A\|$.

We claim that $w = 0$. To see this, note

$$A_1 \begin{bmatrix} \sigma \\ w \end{bmatrix} = \begin{bmatrix} \sigma^2 + w^{\mathrm{T}} w \\ * \end{bmatrix}$$

and therefore

$$\left\| A_1 \begin{bmatrix} \sigma \\ w \end{bmatrix} \right\| \geq \sigma^2 + \|w\|^2.$$

On the other hand,

$$\left\| A_1 \begin{bmatrix} \sigma \\ w \end{bmatrix} \right\| \leq \|A\|(\sigma^2 + \|w\|^2)^{1/2} \;=\; \sigma(\sigma^2 + \|w\|^2)^{1/2}.$$

It follows that $w = 0$. The argument can now be completed by induction. □

The nonnegative numbers σ_i in Theorem 8.14 are called the *singular values* of A and are sometimes written $\sigma_i(A)$. We will see soon enough (Corollary 2.15) that they are uniquely determined by A. Sometimes one writes $\sigma_{\max}$ and $\sigma_{\min}$ for σ_1 and σ_p, respectively. The ith columns u_i and v_i of U and V in Theorem 8.14 are called ith *left singular vector* and ith *right singular vector* of A, respectively (in general, those are not uniquely determined).

Remark 2.11. If $A \in \mathbb{R}^{n\times n}$ is symmetric, then the singular values of A are the absolute values of its eigenvalues.

The following result summarizes the main properties of the singular value decomposition.

Proposition 2.12. *Suppose that* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0$ *are the singular values of* $A \in \mathbb{R}^{m\times n}$ *and* u_i, v_i *are left and right singular vectors of* A*. Then:*

1. $A = \sum_{i=1}^{r} \sigma_i u_i v_i^{\mathrm{T}}$ *(singular value decomposition of* A*),*
2. $\mathsf{rank}(A) = r$,
3. $\ker(A) = \mathsf{span}\{v_{r+1}, \ldots, v_n\}$, $\mathsf{Im}(A) = \mathsf{span}\{u_1, \ldots, u_r\}$,
4. $\|A\| = \sigma_1$, $\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_p^2$,
5. $\min_{\|x\|=1} \|Ax\| = \sigma_n$ *, if* $m \geq n$,
6. $\kappa(A) = \sigma_1/\sigma_n$ *, if* $m = n$, $A \neq 0$,
7. A *and* A^{T} *have the same singular values, in particular* $\|A\| = \|A^{\mathrm{T}}\|$,
8. $\|A\| \leq \|A\|_F \leq \sqrt{\mathsf{rank}(A)}\|A\|$.

Proof. In the case $p = m \leq n$ we have

$$A = U \cdot \mathsf{diag}_{m,n}(\sigma_1, \ldots, \sigma_m) \cdot V^{\mathrm{T}} = [u_1 \ldots u_m] \begin{bmatrix} \sigma_1 v_1^{\mathrm{T}} \\ \vdots \\ \sigma_m v_m^{\mathrm{T}} \end{bmatrix} = \sum_{i=1}^{m} \sigma_i u_i v_i^{\mathrm{T}}. \tag{2.10}$$

The case $n > m$ is treated similarly, which proves the first assertion. The second assertion is immediate from the diagonal form of $U^{\mathrm{T}}AV$.

For showing (3) note that

$$(Av_1, \ldots, Av_n) = AV = U\,\mathsf{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) = (\sigma_1 u_1, \ldots, \sigma_r u_r, 0, \ldots, 0)$$

implies the inclusions $\mathsf{span}\{v_{r+1}, \ldots, v_n\} \subseteq \ker(A)$ and $\mathsf{span}\{u_1, \ldots, u_r\} \subseteq \mathsf{Im}(A)$. Equality follows by comparing the dimensions.

Assertion (4) is an immediate consequence of the orthogonal invariance of the spectral norm and the Frobenius norm, cf. Lemma 2.9. For (5) note that

$$\min_{\|x\|=1} \|Ax\| = \min_{\|x\|=1} \big\| \mathsf{diag}_{m,n}(\sigma_1, \ldots, \sigma_p)x \big\| = \begin{cases} \sigma_n & \text{if } m \geq n \\ 0 & \text{otherwise.} \end{cases}$$

For proving (6) suppose $m = n$ and $A \in \mathbb{R}^{n\times n}$ invertible. Then

$$V^{\mathrm{T}} A^{-1} U \;=\; \mathsf{diag}(\sigma_1^{-1}, \ldots, \sigma_n^{-1}).$$

Hence $\sigma_n^{-1} \geq \sigma_{n-1}^{-1} \geq \ldots \geq \sigma_1^{-1}$ are the singular values of A^{-1}. Assertion (4) implies $\|A^{-1}\| = \sigma_n^{-1}$. Hence

$$\kappa(A) \;=\; \|A\| \cdot \| A^{-1}\| \;=\; \frac{\sigma_1}{\sigma_n}.$$

The first part of Assertion (7) is trivial, the second easily follows from (4). Finally, assertion (8) follows from (4) by noting $\sigma_1^2 + \cdots + \sigma_r^2 \leq r\sigma_1^2$. □

We draw now some conclusions from the singular value decomposition. For a square matrix we always have $\kappa(A) \geq 1$. So the best condition one can hope for is $\kappa(A) = 1$. Orthogonal matrices A satisfy this property, since $\|A\| = 1$ (and A^{-1} is orthogonal as well). Of course, any nonzero multiple λA of an orthogonal matrix A also satisfies

$$\kappa(A) = \|\lambda A\| \cdot \| \lambda^{-1}A^{-1}\| = \lambda\lambda^{-1}\|A\| = 1.$$

Proposition 2.12(6) implies that these are all matrices with $\kappa(A) = 1$.

Corollary 2.13. *If $\kappa(A) = 1$ then $\sigma_1 = \ldots = \sigma_n$. This implies that $U^{\mathrm{T}}AV = \sigma_1 \mathrm{I}$ and hence $\sigma_1^{-1}A$ is orthogonal.* □

The following results extend Theorem 2.6 in the case of spectral norms.

Theorem 2.14. *Let $A = \sum_{i=1}^{r} \sigma_i u_i v_i^{\mathrm{T}}$ be a singular value decomposition of $A \in \mathbb{R}^{m\times n}$ and $0 \leq k < r = \mathsf{rank}(A)$. Then we have*

$$\min_{\mathsf{rank}(B)\leq k} \|A - B\| \;=\; \|A - A_k\| \;=\; \sigma_{k+1},$$

where $A_k := \sum_{i=1}^{k} \sigma_i u_i v_i^{\mathrm{T}}$.

Proof. As in (2.10) we get $U^{\mathrm{T}}A_kV = \mathsf{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$, which implies that $\mathsf{rank}(A_k) = k$. Moreover, $U^{\mathrm{T}}(A - A_k)V = \mathsf{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)$, which implies that $\|A - A_k\| = \sigma_{k+1}$.

Let now $B \in \mathbb{R}^{m\times n}$ with $\mathsf{rank}(B) \leq k$. Then $\dim(\ker B) \geq n - k$ and therefore $\mathsf{span}\{v_1, \ldots, v_{k+1}\} \cap \ker B \neq 0$. Let z be an element of this intersection with $\|z\| = 1$. Then

$$Az \;=\; \sum_{i=1}^{r} \sigma_i u_i v_i^{\mathrm{T}} z \;=\; \sum_{i=1}^{r} \sigma_i \langle v_i, z\rangle u_i \;,$$

and hence

$$\|Az\|^2 = \sum_{i=1}^{r} \sigma_i^2 \langle v_i, z\rangle^2 \geq \sum_{i=1}^{k+1} \sigma_i^2 \langle v_i, z\rangle^2 \geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} \langle v_i, z\rangle^2 = \sigma_{k+1}^2.$$

Therefore,

$$\|A - B\|^2 \geq \|(A - B)z\|^2 = \|Az\|^2 \geq \sigma_{k+1}^2$$

completing the proof of the theorem. □

Corollary 2.15. *The singular values σ_i of A are uniquely determined.* □

We can now extend some of the discussion in Section 2.1 from square to rectangular matrices. Put $p := \min\{m, n\}$ and consider the *set of ill-posed matrices*

$$\Sigma := \{A \in \mathbb{R}^{m\times n} \mid \mathsf{rank}(A) < p\}.$$

We may measure the distance to ill-posedness from a matrix $A \in \mathbb{R}^{m\times n}$, similarly as in (2.9), by the spectral norm, resulting in $d(A, \Sigma)$. Alternatively, we may also measure the distance from A to Σ with respect to the Frobenius norm and define

$$d_F(A, \Sigma) \;:=\; \min\{\|A - B\|_F \mid B \in \Sigma\}.$$

It turns out that this gives the same distance as when using the spectral norm.

Corollary 2.16. *For $A \in \mathbb{R}^{m\times n}$ we have $d(A, \Sigma) = d_F(A, \Sigma) = \sigma_{\min}(A)$.*

Proof. It is sufficient to show that $d_F(A, \Sigma) \le d(A, \Sigma)$ as the other inequality is obvious. Theorem 2.6 with $k = p - 1$ tells us that $d(A, \Sigma)$ equals to the smallest singular value σ_p of A. Let now $A = \sum_{i=1}^{p} \sigma_i u_i v_i^{\mathrm{T}}$ be a singular value decomposition of A. Then $B = \sum_{i=1}^{p-1} \sigma_i u_i v_i^{\mathrm{T}}$ lies in Σ and $A - B = \sigma_n u_n v_n^{\mathrm{T}}$ has Frobenius norm σ_n. Therefore $d_F(A, \Sigma) \le \sigma_p$, completing the proof. □

Remark 2.17. The singular value decomposition has a natural extension to complex matrices and so have all the results in this and the previous sections.

We finish this section with two results that will be needed in Chapter 8. Recall that $\sigma_{\min}(A)$ denotes the smallest singular values of A.

Lemma 2.18. *Let $A \in \mathbb{R}^{m\times n}$ with $n \ge m$ and $\sigma_{\min}(A) > 0$. Denote by B_m and B_n the closed unit balls in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. Then we have*

$$\sigma_{\min}(A) = \sup\{\lambda > 0 \mid \lambda B_m \subseteq A(B_n)\}.$$

Proof. By Theorem 8.14 we assume w.l.o.g. that $A = \mathsf{diag}(\sigma_1, \ldots, \sigma_m)$. It follows that

$$A(B_n) = \Big\{y \in \mathbb{R}^m \mid \tfrac{y_1^2}{\sigma_1^2} + \ldots + \tfrac{y_m^2}{\sigma_m^2} \le 1\Big\},$$

which is a hyperellipsoid with semi-axes σ_i. This shows the assertion (see Figure 8.2). □

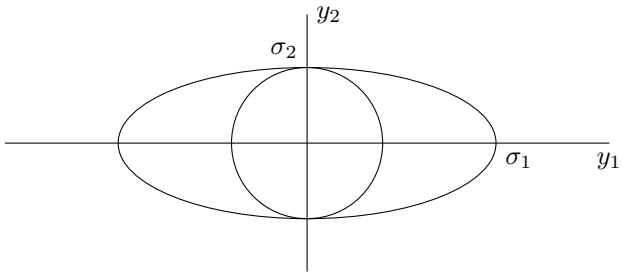

Figure 2.1: Ball of maximal radius σ_2 contained in an ellipse.

Remark 2.19. It is sometimes useful to visualize the singular values of A as the lengths of the semi-axes of the hyperellipsoid $\{Ax \mid \|x\| = 1\}$.

We will also need the following perturbation result.

Lemma 2.20. *For $A, B \in \mathbb{R}^{m\times n}$ we have*

$$|\sigma_{\min}(A+B) - \sigma_{\min}(A)| \leq \| B\|.$$

Proof. Since A and A^{T} have the same singular values, we assume w.l.o.g. that $n \geq m$. According to the characterization of $\sigma_{\min}$ in Proposition 2.12 there exists $x \in \mathbb{R}^n$ with $\|x\| = 1$ such that $\|Ax\| = \sigma_{\min}(A)$. Then

$$\sigma_{\min}(A+B) \leq \|(A+B)x\| \leq \|Ax\| + \|Bx\| \leq \sigma_{\min}(A) + \|B\|.$$

Since A, B were arbitrary we also get

$$\sigma_{\min}(A) = \sigma_{\min}((A+B)+(-B)) \leq \sigma_{\min}(A+B) + \|B\|.$$

This proves the assertion. □

2.4 Least Squares and the Moore-Penrose Inverse

In Section 2.1 we studied the condition of solving a square system of linear equations. If, instead, there are more equations than variables (overdetermined case) or less equations than variables (underdetermined case), the Moore-Penrose inverse and its condition naturally enter the game.

Let $A \in \mathbb{R}^{m\times n}$ be of maximal rank $p = \min\{m, n\}$ with a singular value decomposition

$$U^{\mathrm{T}} A V = \mathsf{diag}_{m,n}(\sigma_1, \ldots, \sigma_p),$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p > 0$. We define the *Moore-Penrose inverse* of A to be the matrix

$$A^{\dagger} = V\, \mathsf{diag}_{n,m}(\sigma_1^{-1}, \ldots, \sigma_p^{-1}) U^{\mathrm{T}}.$$

From the geometric description of $A^\dagger$ given below it follows that $A^\dagger$ is in fact independent of the choice of the orthogonal matrices U and V.

Lemma 2.21. *1. Suppose that $m \geq n$ and $A \in \mathbb{R}^{m\times n}$ has rank n. Then the matrix A defines a linear isomorphism A_1 of $\mathbb{R}^n$ onto $\mathsf{Im}(A)$ and we have $A^\dagger = A_1^{-1}\circ\pi$, where $\pi\colon \mathbb{R}^m \to \mathsf{Im}(A)$ denotes the orthogonal projection. In particular, $A^\dagger A = \mathrm{I}$. Moreover, $A^{\mathrm{T}}A$ is invertible and $A^\dagger = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$.*

2. Suppose that $n \geq m$ and $A \in \mathbb{R}^{m\times n}$ has rank m. Then the matrix A defines an isomorphism $A_2\colon (\ker A)^\perp \to \mathbb{R}^m$ and we have $A^\dagger = \iota \circ A_2^{-1}$, where $\iota\colon (\ker A)^\perp \to \mathbb{R}^n$ denotes the embedding. In particular, $AA^\dagger = \mathrm{I}$. Moreover, AA^{T} is invertible and $A^\dagger = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}$.

Proof. The claims are obvious for the diagonal matrix $A = \mathsf{diag}_{m,n}(\sigma_1, \ldots, \sigma_p)$ and easily extend to the general case by orthogonal invariance. □

The following is obvious from the definition of $A^\dagger$.

Corollary 2.22. *We have $\|A^\dagger\| = \frac{1}{\sigma_{\min}(A)}$.* □

Suppose we are given a matrix $A \in \mathbb{R}^{m\times n}$, with $m > n$ and $\mathsf{rank}(A) = n$, as well as $b \in \mathbb{R}^m$. Since A, as a linear map, is not surjective, the system $Ax = b$ may have no solutions. We might therefore attempt to find the point $x \in \mathbb{R}^n$ with Ax closest to b. That is, to solve the problem

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|^2. \tag{2.11}$$

Since A is injective there is a unique minimizer x for (2.11) namely, the preimage of the projection c of b onto $\mathsf{Im}(A)$. From Lemma 2.21(1) it follows immediately that the minimizer can be expressed as $x = A^\dagger b$ (see Figure 2.2).

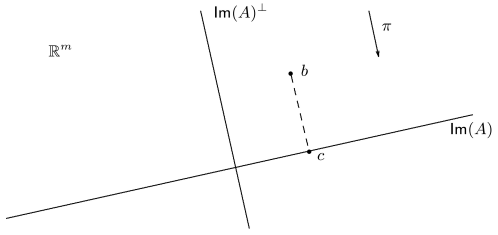

Figure 2.2: The spaces $\mathsf{Im}(A)$, $\mathsf{Im}(A)^\perp$ and the points b and c in $\mathbb{R}^m$.

For the case of underdetermined systems we consider instead the case $m < n$ and $\mathsf{rank}(A) = m$. For each $b \in \mathbb{R}^m$, the set of solutions of $Ax = b$ is an affine

subspace of $\mathbb{R}^n$ of dimension $n-m$ and therefore contains a unique point of minimal norm. We want to find this point, i.e., to solve

$$\min_{x|Ax=b} \|x\|^2. \tag{2.12}$$

Lemma 2.21(2) implies that the solution of (2.12) again satisfies $x = A^\dagger b$.

So the Moore-Penrose inverse naturally yields the solution of linear least squares problems and of underdetermined systems. What is the condition of computing the Moore-Penrose inverse? Theorem 2.5 has a natural extension showing that the quantity

$$\kappa_{rs}(A) := \|A\|_{rs}\|A^\dagger\|_{sr}$$

equals the normwise condition for the computation of the Moore-Penrose inverse.

Theorem 2.23. *Consider*

$$\psi\colon \{A \in \mathbb{R}^{m\times n} \mid \mathsf{rank}(A) = \min\{m,n\}\} \to \mathbb{R}^{m\times n}, A \mapsto A^\dagger.$$

Then we have $\mathsf{cond}^\psi(A) = \kappa_{rs}(A)$*, when measuring errors on the data space with respect to* $\|\ \|_{rs}$ *and on the solution space with respect to* $\|\ \|_{sr}$*.*

Proof. Let $\tilde{A} = A - E$. We claim that for $\|E\| \to 0$ we have

$$\tilde{A}^\dagger - A^\dagger = A^\dagger E A^\dagger + o(\|E\|).$$

For proving this we may assume w.l.o.g. that $m \geq n$, hence $A^\dagger = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$, and perform a computation similar as in the proof of Theorem 2.5. We leave the straightforward details to the reader. The remaining arguments then follow in exactly the same way as in the proof of Theorem 2.5, just by replacing A^{-1} by $A^\dagger$. □

We note that the solution of linear least squares problems and underdetermined systems has, in contrast with Moore-Penrose inversion, a normwise condition that is only losely approximated by $\kappa(A)$. Indeed, in 1973, P.-A. Wedin gave tight upper bounds for this normwise condition from which it follows that it varies between $\Theta(\kappa(A))$ and $\Theta(\kappa(A)^2)$ (see e.g., [12, Theorem 19.1] for the precise statement). Interestingly, unlike Theorem 2.3, the normwise condition for solving $\min \|Ax - b\|$ depends on b as well as on A.

We finally note that Theorem 2.6 has a natural extension: $\kappa(A)$ is again the relativized inverse of the distance to ill-posedness, where the latter now amounts to rank-deficiency. The following is an immediate consequence of Corollary 2.16.

Corollary 2.24. *For* $A \in \mathbb{R}^{m\times n}$ *we have*

$$\kappa(A) = \frac{\|A\|}{d(A,\Sigma)} = \frac{\|A\|}{d_F(A,\Sigma)},$$

where $\Sigma = \{A \in \mathbb{R}^{m\times n} \mid \mathsf{rank}(A) < \min\{m,n\}\}$. □