Appendix A

Facts from Linear Algebra and Matrix Analysis

Linear algebra studies linear systems of equations, and their solutions. This study is extremely important for engineering applications. Linear models represent a simple, tractable first choice for modeling complicated systems. Moreover, many devices for measuring the physical world are designed to produce measurements that are as close as possible to linear functions of the signal to be measured. In this appendix, we review several fundamental definitions, constructions and facts from linear algebra and matrix analysis. For readers with background in engineering, statistics or applied mathematics, much of this material is likely to be familiar. Section A.9, contains a briefly review of norms on matrices and spectral functions of matrices, two more advanced topics which we use extensively throughout the book. We have attempted to make this introduction as simple and self-contained as possible; readers looking for a more thorough introduction to this area could consult the (excellent) books of Horn and Johnson [Horn and Johnson, 1985b] or Bhatia [Bhatia, 1996b].

A.1 Vector Spaces, Linear Independence, Bases and Dimension

We use the notation $\mathbb{R}$ for the real numbers, and $\mathbb{R}^n$ for the n-dimensional real vectors

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}. \tag{A.1.1}$$

The space $\mathbb{R}^n$ an example of a vector space – a space in which we can perform addition and scalar multiplication a way that conforms to our intuition from $\mathbb{R}^3$. More formally:

Definition A.1.1 (Vector space)**.** *A* vector space $\mathbb{V}$ *over a field of scalars* $\mathbb{F}$ *is a set* $\mathbb{V}$ *(with a special distinguished zero element* $\mathbf{0} \in \mathbb{V}$*) endowed with two operations:*

- **Vector addition** $+$*, which takes two vectors* $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{V}$ *and produces another vector* $\boldsymbol{v} + \boldsymbol{w} \in \mathbb{V}$*,*
- **Scalar multiplication***, which takes a vector* $\boldsymbol{v} \in \mathbb{V}$ *and a scalar* $\alpha \in \mathbb{F}$*, and produces a vector* $\alpha \boldsymbol{v} \in \mathbb{V}$*,*

such that (1) addition is associative: $\boldsymbol{v} + (\boldsymbol{w} + \boldsymbol{x}) = (\boldsymbol{v} + \boldsymbol{w}) + \boldsymbol{x}$*, (2) addition is commutative:* $\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{w} + \boldsymbol{v}$*, (3) zero is the additive identity:* $\boldsymbol{v} + \mathbf{0} = \boldsymbol{v}$*, (3) every element has an additive inverse: for each* $\boldsymbol{v} \in \mathbb{V}$*, there exists an element "*$-\boldsymbol{v}$*"* $\in \mathbb{V}$ *such that* $\boldsymbol{v} + -\boldsymbol{v} = \mathbf{0}$*, (5)* $\alpha(\beta \boldsymbol{v}) = (\alpha\beta)\boldsymbol{v}$*, (6) multiplicative identity:* $1\boldsymbol{v} = \boldsymbol{v}$*, where* $1 \in \mathbb{F}$ *is the multiplicative identity in* $\mathbb{F}$*, (7)* $\alpha(\boldsymbol{v} + \boldsymbol{w}) = \alpha\boldsymbol{v} + \alpha\boldsymbol{w}$*, (8)* $(\alpha + \beta)\boldsymbol{v} = \alpha\boldsymbol{v} + \beta\boldsymbol{v}$*.*

Example A.1.2. *The following are examples of vector spaces (check this!)*

- *The n-dimensional real vectors* $\mathbb{R}^n$*, over the scalar field* $\mathbb{F} = \mathbb{R}$*.*
- *The $m \times n$ real matrices*

  $$\mathbb{R}^{m \times n} = \left\{ \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{bmatrix} \middle| \; x_{ij} \in \mathbb{R} \right\}, \tag{A.1.2}$$

  over the scalar field $\mathbb{F} = \mathbb{R}$*.*
- *The complex vectors* $\mathbb{C}^n$ *or complex matrices* $\mathbb{C}^{m \times n}$*, over the scalar field* $\mathbb{F} = \mathbb{C}$*.*
- *Function spaces, e.g.,*

  $$C[0,1] = \{ f : [0,1] \to \mathbb{R} \mid \; f \text{ continuous} \}, \tag{A.1.3}$$

over $\mathbb{R}$. Vector spaces of functions defined on the continuum arise naturally in the study sampling problems, in which we wish to derive information about the physical world from digital measurements.

By itself, the notion of a vector space is not particularly rich: it is simply a space in which linear operations make sense. A vector space can be viewed as the "playing field" on which much more interesting models can be built, and much richer questions can be asked. As a step in this direction, we can note that it makes sense to take linear combinations of elements of a vector space. A *linear combination* is an expression of the form

$$\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \cdots + \alpha_k \boldsymbol{v}_k,$$

where $\alpha_1, \dots, \alpha_k \in \mathbb{F}$ and $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k \in \mathbb{V}$.

Definition A.1.3 (Linear independence)**.** *A set of vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ are* linearly independent *if*

$$\sum_{i=1}^{k} \alpha_i \boldsymbol{v}_i = \mathbf{0} \qquad \Longrightarrow \qquad \alpha_1 = 0, \dots, \alpha_k = 0.$$

If a collection of vectors are not linearly independent, then there exists some choice of (α_i) not all zero, for which $\sum_i \alpha_i \boldsymbol{v}_i = \mathbf{0}$. In this case, we say that the set $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ is *linearly dependent*.

Definition A.1.4 (Basis for a vector space)**.** *A* basis *B for the vector space $\mathbb{V}$ is defined as a* maximal, linearly independent set.

Here, *maximal* means that B is not contained in any larger linearly independent set. Any basis B for $\mathbb{V}$ *spans* $\mathbb{V}$, in the sense that every element of $\mathbb{V}$ can be written as a linear combination of elements of B:

$$\forall\, \boldsymbol{v} \in \mathbb{V},\ \exists\, \boldsymbol{b}_1, \dots, \boldsymbol{b}_k \in B,\ \alpha_1, \dots, \alpha_k \in \mathbb{F}, \quad \text{such that} \quad \boldsymbol{v} = \sum_{i=1}^{k} \alpha_i \boldsymbol{b}_i. \tag{A.1.4}$$

Moreover, if B is a basis, the coefficients $\alpha_1, \dots, \alpha_k$ in the above expression are unique.

Example A.1.5. *In $\mathbb{R}^n$, we often use the* standard basis $B = \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ *of coordinate vectors*

$$\boldsymbol{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \boldsymbol{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \dots \quad \boldsymbol{e}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}. \tag{A.1.5}$$

In $\mathbb{R}^{m \times n}$ we may work with the standard basis of coordinate matrices $\boldsymbol{E}_{ij}$ that are one in entry (i, j) and zero elsewhere.

Figure A.1. **Linear subspaces** of $\mathbb{R}^2$ and $\mathbb{R}^3$.

Every vector space $\mathbb{V}$ has a basis.[1] One very fundamental result in linear algebra states that every basis has the same size:

Theorem A.1.6 (Invariance of dimension)**.** *For any vector space* $\mathbb{V}$*, every basis* B *has the same cardinality, which we denote* $\dim(\mathbb{V})$*, and call the* dimension *of* $\mathbb{V}$*.*

The notion of dimension is especially useful for talking about subspaces of the vector space $\mathbb{V}$.

Definition A.1.7 (Linear subspace)**.** *A* linear subspace *of a vector space* $\mathbb{V}$ *is a set* $\mathbb{W} \subseteq \mathbb{V}$ *that is also a vector space.*

For $\mathbb{W} \subseteq \mathbb{V}$ to be a linear subspace, it is necessary and sufficient that $\mathbb{W}$ be stable under linear combinations: for all $\alpha, \beta \in \mathbb{F}$ and $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{W}$, $\alpha \boldsymbol{w}_1 + \beta \boldsymbol{w}_2 \in \mathbb{W}$. Linear subspaces play a very important dual role, both as cleanly characterizing the solvability of linear equations, and as geometric data models. Geometrically, we can visualize a subspace as a generalization of a line, or plane, which must pass the origin: $\mathbf{0} \in \mathbb{W}$ (see Figure A.1).

A.2 Inner Products

The most important geometric relationship between subspaces is that of *orthogonality*. To describe it clearly, we need the notion of an inner product. Below, we will assume that we are working with a vector space over either the real or complex numbers, and so the complex conjugate $\bar{\alpha}$ of $\alpha \in \mathbb{F}$ is well-defined.

Definition A.2.1 (Inner product)**.** *A function* $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{F}$ *is an* inner product *if it satisfies:*

- **Linearity***:* $\langle \alpha \boldsymbol{v} + \beta \boldsymbol{w}, \boldsymbol{x} \rangle = \alpha \langle \boldsymbol{v}, \boldsymbol{x} \rangle + \beta \langle \boldsymbol{w}, \boldsymbol{x} \rangle$*;*
- **Conjugate symmetry** $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \overline{\langle \boldsymbol{w}, \boldsymbol{v} \rangle}$*;*
- **Positive definiteness** $\langle \boldsymbol{v}, \boldsymbol{v} \rangle \geq 0$*, with equality iff* $\boldsymbol{v} = \mathbf{0}$*.*

We then say that $\boldsymbol{v}$ and $\boldsymbol{w}$ are *orthogonal* (with respect to inner product $\langle \cdot, \cdot \rangle$) if $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = 0$. In this case, we write $\boldsymbol{v} \perp \boldsymbol{w}$. For a given set $S \subseteq$

[1] This statement may seem obvious, but is tricky: it turns out to be equivalent to the *axiom of choice* in set theory, and hence is best viewed as an assumption. For the vector spaces we consider in this course ($\mathbb{R}^n$, $\mathbb{C}^n$, ect.), it will be very easy to construct a basis, and so for our purposes, the question is essentially moot.

$\mathbb{V}$, we define its orthogonal complement as the set of all vectors that are orthogonal to every element of S:

Definition A.2.2 (Orthogonal complement)**.** *For* $S \subseteq \mathbb{V}$, $S^\perp = \{\boldsymbol{v} \in \mathbb{V} \mid \langle \boldsymbol{v}, \boldsymbol{s} \rangle = 0 \; \forall \, s \in S\}$.

It is worth noting that for any set S, $S^\perp \subseteq \mathbb{V}$ is a linear subspace. This holds even if S is not a subspace itself.

We will use (and return to) two main examples of inner products. The first is the canonical inner product on $\mathbb{R}^n$, which simply sets

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{z}_i. \tag{A.2.1}$$

This extends to a canonical inner product on $\mathbb{R}^{m \times n}$, which is sometimes called the Frobenius inner product:

$$\langle \boldsymbol{X}, \boldsymbol{Z} \rangle \doteq \sum_{i=1}^{m} \sum_{j=1}^{n} \boldsymbol{X}_{ij} \boldsymbol{Z}_{ij}. \tag{A.2.2}$$

Recall that the trace of a square matrix is simply the sum of its diagonal elements:

Definition A.2.3. *For* $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, $\operatorname{trace}(\boldsymbol{M}) = \sum_{i=1}^{n} \boldsymbol{M}_{ii}$.

Using the trace, we can give an expression for the Frobenius inner product which appears more complicated, but actually turns out to be tremendously useful:

$$\langle \boldsymbol{X}, \boldsymbol{Z} \rangle = \operatorname{trace}(\boldsymbol{X}^* \boldsymbol{Z}) = \operatorname{trace}(\boldsymbol{X} \boldsymbol{Z}^*). \tag{A.2.3}$$

For manipulating this expression, it is worth noting that the trace is invariant under a cyclic permutation of its argument:

Theorem A.2.4. *For any matrices* $\boldsymbol{A}$, $\boldsymbol{B}$ *of compatible size,* $\operatorname{trace}(\boldsymbol{AB}) = \operatorname{trace}(\boldsymbol{BA})$. *More generally, if* $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n$ *are matrices of compatible size, and* π *is a cyclic permutation on* $\{1, \dots, n\}$,

$$\operatorname{trace}(\boldsymbol{A}_1 \boldsymbol{A}_2 \dots \boldsymbol{A}_n) \;=\; \operatorname{trace}\left(\boldsymbol{A}_{\pi(1)} \boldsymbol{A}_{\pi(2)} \dots \boldsymbol{A}_{\pi(n)}\right). \tag{A.2.4}$$

A.3 Linear Transformations and Matrices

A mapping $\mathcal{L}$ between vector spaces $\mathbb{V}$ and $\mathbb{V}'$ over a common field $\mathbb{F}$ is a *linear transformation* (or linear map) if it respects the vector space operations:

Definition A.3.1 (Linear map)**.** *A* linear map *is a function* $\mathcal{L} : \mathbb{V} \to \mathbb{V}'$ *such for all* $\alpha, \beta \in \mathbb{F}$ *and* $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{V}$, $\mathcal{L}[\alpha \boldsymbol{v} + \beta \boldsymbol{w}] = \alpha \mathcal{L}[\boldsymbol{v}] + \beta \mathcal{L}[\boldsymbol{w}]$.

If $\mathbb{V}' = \mathbb{V}$ then we call $\mathcal{L}$ a *linear operator*.

Example A.3.2. *Let $\mathbb{V} = \mathbb{R}^{m\times n}$, and $\Omega \subseteq \{1,\dots,m\} \times \{1,\dots,n\}$. Let $\mathcal{P}_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ via*

$$(\mathcal{P}_\Omega[\boldsymbol{X}])_{ij} = \begin{cases} \boldsymbol{X}_{ij} & (i,j) \in \Omega, \\ 0 & \textit{else}, \end{cases} \tag{A.3.1}$$

i.e., the restriction of $\boldsymbol{X}$ to Ω. Then $\mathcal{P}_\Omega$ is a linear operator.

The special case of $\mathbb{V} = \mathbb{R}^n$, $\mathbb{V}' = \mathbb{R}^m$ is of special importance. It turns out that there is a bijective correspondence between linear operators $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ and $m \times n$ matrices:

Theorem A.3.3. *For $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, let*

$$(\boldsymbol{A}\boldsymbol{x})_i = \sum_j \boldsymbol{A}_{ij}\boldsymbol{x}_j. \tag{A.3.2}$$

Then for every $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the mapping $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Conversely for every linear map $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ there exists a unique $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ such that for every $\boldsymbol{x}$, $\mathcal{L}[\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{x}$.

This fact justifies the seemingly awkward standard definition of matrix multiplication – it is simply the correct way of representing the composition of two linear maps:

Theorem A.3.4. *If $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^p$ and $\mathcal{L}' : \mathbb{R}^p \to \mathbb{R}^m$ are linear maps, with corresponding matrix representations $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ and $\boldsymbol{A}' \in \mathbb{R}^{m\times p}$, and $\mathcal{L}' \circ \mathcal{L}$ denotes the composition $\mathcal{L}' \circ \mathcal{L}(\boldsymbol{x}) = \mathcal{L}'[\mathcal{L}[\boldsymbol{x}]]$, then $\mathcal{L}' \circ \mathcal{L}$ is a linear map, and its matrix representation is given by the matrix product $\boldsymbol{A}'\boldsymbol{A}$ whose (i,j) entry is*

$$\left(\boldsymbol{A}'\boldsymbol{A}\right)_{ij} = \sum_{k=1}^{p} \boldsymbol{A}'_{ik}\boldsymbol{A}_{kj}. \tag{A.3.3}$$

The (conjugate) transpose of a matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is the $n \times m$ matix $\boldsymbol{A}^* \in \mathbb{C}^{n\times m}$ given by:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \Rightarrow \boldsymbol{A}^* = \begin{bmatrix} \overline{a_{11}} & \dots & \overline{a_{m1}} \\ \vdots & \ddots & \vdots \\ \overline{a_{1n}} & \dots & \overline{a_{mn}} \end{bmatrix} \tag{A.3.4}$$

When $\boldsymbol{A}$ is real, this is just the transpose. Transposition is a very simple operation on the entries of a matrix, but it has a basic reason for existing:

Theorem A.3.5. *Let $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$, with corresponding matrix $\boldsymbol{A}$. Its adjoint map is the unique linear map $\mathcal{L}^* : \mathbb{R}^m \to \mathbb{R}^n$ satisfying*

$$\forall\, \boldsymbol{x}, \boldsymbol{y}, \qquad \langle \boldsymbol{y}, \mathcal{L}[\boldsymbol{x}] \rangle = \langle \mathcal{L}^*[\boldsymbol{y}], \boldsymbol{x} \rangle\,. \tag{A.3.5}$$

The matrix $\boldsymbol{A}^*$ is the matrix representation of the adjoint map $\mathcal{L}^*$.

A linear map $\mathcal{L} : \mathbb{V} \to \mathbb{V}'$ is *invertible* if for every $\boldsymbol{y} \in \mathbb{V}'$, there is a unique $\boldsymbol{x} \in \mathbb{V}$ such that $\mathcal{L}[\boldsymbol{x}] = \boldsymbol{y}$. In particular, if $\mathbb{V} = \mathbb{V}' = \mathbb{R}^n$, we call $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ invertible if it corresponds to an invertible linear map. This means that the system of equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{A.3.6}$$

always has a unique solution

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}. \tag{A.3.7}$$

It is not too difficult to show that if $\mathcal{L}$ is a linear map, its inverse $\mathcal{L}^{-1}$ is also linear. So, the notation $\boldsymbol{A}^{-1}$ above can be taken to mean "the matrix representation of the inverse mapping $\mathcal{L}^{-1}$". Fortunately, there are much more concrete criteria for determining if a given matix $\boldsymbol{A}$ is invertible, and if so, for calculating $\boldsymbol{A}^{-1}$.

Definition A.3.6 (Determinant)**.** *The determinant of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is the signed volume of the parallelepiped defined by the columns of $\boldsymbol{A}$:*

$$\det(\boldsymbol{A}) \;=\; \sum_{\pi \text{ a permutation on } \{1,\dots,n\}} \operatorname{sgn}(\pi) \times \prod_{i=1}^{n} A_{i,\pi(i)}, \tag{A.3.8}$$

The explicit expression (A.3.8) is not usually of direct use. More important is the geometric intuition: if $\det(\boldsymbol{A})$ is zero, the columns of $\boldsymbol{A}$ span a parallelpiped of zero volume, and so they lie on some lower dimensional subspace of $\mathbb{R}^n$. Vectors $\boldsymbol{y}$ that do not reside in this subspace cannot be generated as linear combinations of the columns of $\boldsymbol{A}$, and $\boldsymbol{A}$ is not invertible. Conversely, if $\det \boldsymbol{A} \neq 0$, the columns of $\boldsymbol{A}$ span all of $\mathbb{R}^n$, and $\boldsymbol{A}$ is invertible. Making this reasoning formal, one obtains

Theorem A.3.7. *A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is invertible if and only if $\det \boldsymbol{A} \neq 0$. If $\boldsymbol{A}$ is invertible, we can express its inverse as $\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})}\boldsymbol{C}$, where $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ is the companion matrix:*

$$\begin{bmatrix} (-1)^{1+1}\det(\boldsymbol{A}_{\setminus 1,\setminus 1}) & (-1)^{1+2}\det(\boldsymbol{A}_{\setminus 2,\setminus 1}) & \dots & (-1)^{1+n}\det(\boldsymbol{A}_{\setminus n,\setminus 1}) \\ (-1)^{2+1}\det(\boldsymbol{A}_{\setminus 1,\setminus 2}) & (-1)^{2+2}\det(\boldsymbol{A}_{\setminus 2,\setminus 2}) & \dots & (-1)^{2+n}\det(\boldsymbol{A}_{\setminus n,\setminus 2}) \\ \vdots & \vdots & \ddots & \vdots \\ (-1)^{n+1}\det(\boldsymbol{A}_{\setminus 1,\setminus n}) & (-1)^{n+2}\det(\boldsymbol{A}_{\setminus 2,\setminus n}) & \dots & (-1)^{n+n}\det(\boldsymbol{A}_{\setminus n,\setminus n}) \end{bmatrix},$$

where the matrix $\boldsymbol{A}_{\setminus i,\setminus j}$ is constructed from $\boldsymbol{A}$ by removing the i-th row and j-th column.

Again, the above expression for $\boldsymbol{A}^{-1}$ is of little use computationally, but is conceptually helpful, since it shows that the entries of the inverse are rational functions of the entries of $\boldsymbol{A}$.

It is worth noting that for any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\det(\boldsymbol{A}\boldsymbol{B}) = \det(\boldsymbol{A})\det(\boldsymbol{B}). \tag{A.3.9}$$

This corroborates the fact that a product of invertible linear maps is invertible, and a product of invertible matrices is invertible. In particular,

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}. \tag{A.3.10}$$

It is also useful to note that for every matrix $\boldsymbol{A}$,

$$\det(\boldsymbol{A}) = \det(\boldsymbol{A}^*). \tag{A.3.11}$$

A.4 Matrix Groups

Because the product of two $n \times n$ matrices is again an $n \times n$ matrix, this operation can produce objects with interesting algebraic structure. We will not emphasize the algebra of matrix groups – or even formally define a group. Rather, we just recall the names of several groups that will recur throughout the course:

- **The General Linear Group** $GL(n, \mathbb{R})$ consists of the invertible matrices:

  $$GL(n, \mathbb{R}) = \left\{ \boldsymbol{A} \in \mathbb{R}^{n \times n} \mid \det(\boldsymbol{A}) \neq 0 \right\}. \tag{A.4.1}$$

  Similarly, $GL(n, \mathbb{C})$ denotes the $n \times n$ invertible matrices with complex entries.

- **The Orthogonal Group** $O(n)$ consists of the real $n \times n$ matrices that satisfy $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{I}$:

  $$O(n) = \left\{ \boldsymbol{A} \in \mathbb{R}^{n \times n} \mid \boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{I} \right\}. \tag{A.4.2}$$

  The expression $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{I}$ implies that $\boldsymbol{A}$ is invertible, and that $\boldsymbol{A}^{-1} = \boldsymbol{A}^*$. Hence, $O(n) \subset GL(n)$. Two notes are in order: first, since $\boldsymbol{I} = \boldsymbol{I}^* = (\boldsymbol{A}^*\boldsymbol{A})^* = \boldsymbol{A}\boldsymbol{A}^*$, it is enough to keep only the expression $\boldsymbol{A}^*\boldsymbol{A}$ in the definition. Second, because $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^*)$, we have $\det(\boldsymbol{A})^2 = \boldsymbol{I}$, and so every $\boldsymbol{A} \in O(n)$ has determinant ± 1.

- **The Special Orthogonal Group** $SO(n)$ consists of the $n \times n$ matrices that satisfy $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{I}$, *and* $\det(\boldsymbol{A}) = +1$:

  $$SO(n) = \left\{ \boldsymbol{A} \in \mathbb{R}^{n \times n} \mid \boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{I},\ \det(\boldsymbol{A}) = +1 \right\}. \tag{A.4.3}$$

  Clearly, $SO(n) \subset O(n) \subset GL(n, \mathbb{R})$. In $\mathbb{R}^3$, the group $SO(3)$ corresponds to the rotation matrices; $O(3)$ contains rotations and reflections.

- **The Unitary and Special Unitary Groups** are subgroups of $GL(n, \mathbb{C})$. The unitary group $U(n)$ contains those matrix $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ satisfying $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{I}$. The special unitary group $SU(n)$ contains those $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ satisfying $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{I}$ and $\det(\boldsymbol{A}) = 1$. So, $SU(n) \subset U(n) \subset GL(n, \mathbb{C})$.

A.5 Subspaces Associated with a Matrix

To each linear operator $\mathcal{L} : \mathbb{V} \to \mathbb{V}'$, we associate two important subspaces, the range and the null space:

Definition A.5.1 (Range, null space)**.** *For* $\mathcal{L} : \mathbb{V} \to \mathbb{V}'$*,*

$$\begin{aligned} \mathrm{range}(\mathcal{L}) &= \{\mathcal{L}[\boldsymbol{x}] \mid \boldsymbol{x} \in \mathbb{V}\} \subseteq \mathbb{V}', && (A.5.1) \\ \mathrm{null}(\mathcal{L}) &= \{\boldsymbol{x} \in \mathbb{V} \mid \mathcal{L}[\boldsymbol{x}] = \boldsymbol{0}\} \subseteq \mathbb{V}. && (A.5.2) \end{aligned}$$

The range is a linear subspace of $\mathbb{V}'$*, while the null space is a linear subspace of* $\mathbb{V}$*.*

Specializing these definitions to $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$, we obtain

$$\begin{aligned} \mathrm{null}(\boldsymbol{A}) &= \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}, && (A.5.3) \\ \mathrm{range}(\boldsymbol{A}) &= \{\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^n\} \;=\; \mathrm{col}(\boldsymbol{A}), && (A.5.4) \\ \mathrm{row}(\boldsymbol{A}) &= \{\boldsymbol{w}^*\boldsymbol{A} \mid \boldsymbol{w} \in \mathbb{R}^m\}. && (A.5.5) \end{aligned}$$

The sets $\mathrm{null}(\boldsymbol{A})$, $\mathrm{range}(\boldsymbol{A})$ and $\mathrm{row}(\boldsymbol{A})$ are all linear subspaces. They satisfy several very important relationships:

Theorem A.5.2. *For* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$*, the following relationships hold:*

- $\mathrm{null}(\boldsymbol{A})^\perp = \mathrm{range}(\boldsymbol{A}^*)$.
- $\mathrm{range}(\boldsymbol{A})^\perp = \mathrm{null}(\boldsymbol{A}^*)$.
- $\mathrm{null}(\boldsymbol{A}^*) = \mathrm{null}(\boldsymbol{A}\boldsymbol{A}^*)$.
- $\mathrm{range}(\boldsymbol{A}) = \mathrm{range}(\boldsymbol{A}\boldsymbol{A}^*)$.

From this, we obtain that $\dim(\mathrm{row}(\boldsymbol{A})) + \dim(\mathrm{null}(\boldsymbol{A})) = n$.

Theorem A.5.3 (Matrix rank)**.** *For any* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$*,* $\dim(\mathrm{row}(\boldsymbol{A})) = \dim(\mathrm{range}(\boldsymbol{A}))$*. We call the common value the* rank *of* $\boldsymbol{A}$*. It is equal to the maximum size of a set of linearly independent rows, which is in turn equal to the maximum size of a set of linearly independent columns.*

The rank satisfies many useful properties:

Theorem A.5.4 (Facts about the rank)**.** *The rank satisfies:*

- $\mathrm{rank}\,(\boldsymbol{A}\boldsymbol{B}) \le \min\{\mathrm{rank}\,(\boldsymbol{A}), \mathrm{rank}\,(\boldsymbol{B})\}$.
- **Sylvester's inequalty** *For* $\boldsymbol{A} \in \mathbb{R}^{m \times p}$*,* $\boldsymbol{B} \in \mathbb{R}^{p \times n}$*,* $\mathrm{rank}\,(\boldsymbol{A}\boldsymbol{B}) \ge \mathrm{rank}\,(\boldsymbol{A}) + \mathrm{rank}\,(\boldsymbol{B}) - p$.
- **Subadditivity** $\forall\ \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$, $\mathrm{rank}\,(\boldsymbol{A} + \boldsymbol{B}) \le \mathrm{rank}\,(\boldsymbol{A}) + \mathrm{rank}\,(\boldsymbol{B})$.
- $\mathrm{rank}\,(\boldsymbol{A}) = \mathrm{rank}\,(\boldsymbol{A}\boldsymbol{A}^*) = \mathrm{rank}\,(\boldsymbol{A}^*\boldsymbol{A})$.

Figure A.2. Geometry of solution sets for linear equations.

A.6 Linear Systems of Equations

Using the range and null space, we can decide if the system $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ has a solution, and how many solutions it has:

Theorem A.6.1. *Consider a linear system of equations* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$.

- **Existence**: *The system* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ *has a solution* $\boldsymbol{x}$ *if and only if* $\boldsymbol{y} \in \mathrm{range}(\boldsymbol{A})$.
- **Uniqueness**: *Suppose that* $\boldsymbol{x}_0$ *satisfies* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$. *Every solution to the equation* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ *can be generated as* $\boldsymbol{x}_0 + \boldsymbol{v}$, *where* $\boldsymbol{v} \in \mathrm{null}(\boldsymbol{A})$. *The solution* $\boldsymbol{x}_0$ *is* unique *if and only if the null space is trivial* ($\mathrm{null}(\boldsymbol{A}) = \{\boldsymbol{0}\}$).

The last point means that whenever $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ has a solution $\boldsymbol{x}_0$, the solution set has the form

$$\boldsymbol{x}_0 + \mathrm{null}(\boldsymbol{A}). \tag{A.6.1}$$

The "+" here is "in the sense of Minkowski", which just means that $\boldsymbol{x} + S = \{\boldsymbol{x} + \boldsymbol{s} \mid \boldsymbol{s} \in S\}$. Since $\mathrm{null}(\boldsymbol{A})$ is a linear subspace, the resulting set is a translate of a linear subspace. We call such a set an *affine subspace*. Unlike a linear subspace, an affine subspace need not contain $\boldsymbol{0}$.

Definition A.6.2 (Affine combination and affine subspace)**.** *Let* $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k \in \mathbb{V}$. *An* affine combination *is an expression of the form* $\sum_i \alpha_i \boldsymbol{v}_i$, *with* $\sum_i \alpha_i = 1$. *An* affine subspace *is a set* $A \subset \mathbb{V}$ *which is stable under affine combinations.*

It is easy to check that A is an affine subspace if and only if $A = \boldsymbol{x} + S$ for some linear subspace S. So, geometrically, we can visualize the solution set of $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ as living on a plane which does not contain $\boldsymbol{0}$ – see Figure A.2.

Invertible systems.

If $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ is square, and has full rank m, then for every $\boldsymbol{y}$, the system $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ has exactly one solution $\hat{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{y}$. In practice, we very frequently encounter linear systems of equations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ for which $\boldsymbol{A}$ is not invertible. We describe two important cases below:

Overdetermined systems.

Suppose that $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $m > n$. Since $\mathrm{rank}(\boldsymbol{A}) \le \min\{m, n\} < m$, the range of $\boldsymbol{A}$ is a lower-dimensional subspace of $\mathbb{R}^m$. Hence, in general, the system of equations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ will not have a solution. Hence, we resort to seeking an approximate solution. Classically, this was often done via the method of *least squares.* Define the Euclidean length $\|\boldsymbol{z}\|_2 = \sqrt{\sum_i z_i^2}$ of a vector $\boldsymbol{z} \in \mathbb{R}^n$. A least-squares solution solves

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 . \tag{A.6.2}$$

If $\boldsymbol{A}$ has full column rank n, the solution $\hat{\boldsymbol{x}}_{LS}$ to this problem is unique, and is given by

$$\hat{\boldsymbol{x}}_{LS} = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*\boldsymbol{y}. \tag{A.6.3}$$

We sometimes write $\boldsymbol{A}^\dagger = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*$, and call this matrix the *pseudoinverse* of $\boldsymbol{A}$. Notice that

$$\begin{aligned} \boldsymbol{A}\hat{\boldsymbol{x}}_{LS} &= \boldsymbol{A}(\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*\boldsymbol{y} && \text{(A.6.4)} \\ &= \boldsymbol{P}_{\mathrm{range}(\boldsymbol{A})}\boldsymbol{y} && \text{(A.6.5)} \end{aligned}$$

is the orthogonal projection of $\boldsymbol{y}$ onto range($\boldsymbol{A}$); the matrix $\boldsymbol{P}_{\mathrm{range}(\boldsymbol{A})} = \boldsymbol{A}(\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*$ is the projection matrix onto this space. The optimal value of the least squares problem is

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}_{LS}\|_2^2 &= \left\|(\boldsymbol{I} - \boldsymbol{P}_{\mathrm{range}(\boldsymbol{A})})\boldsymbol{y}\right\|_2^2 && \text{(A.6.6)} \\ &= \left\|\boldsymbol{P}_{\mathrm{range}(\boldsymbol{A})^\perp}\boldsymbol{y}\right\|_2^2 . && \text{(A.6.7)} \end{aligned}$$

This is just the squared (Euclidean) distance from the observation $\boldsymbol{y}$ to range($\boldsymbol{A}$).

Underdetermined systems.

If on the other hand $m < n$, as discussed above, the solution is not unique – if any solution $\boldsymbol{x}_0$ exists, then there is an entire affine space $\boldsymbol{x}_0 + \mathrm{null}(\boldsymbol{A})$ of solutions. A classical approach to handling such underdetermined systems is to look for the $\boldsymbol{x}$ of smallest length that is consistent with the system. Formally,

$$\text{minimize} \quad \|\boldsymbol{x}\|_2^2 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{A.6.8}$$

If $\boldsymbol{A}$ has full row rank (i.e., rank($\boldsymbol{A}$) $= m$), this problem has a unique solution:

Theorem A.6.3. *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ have full row rank (i.e.,* $\mathrm{rank}\,(\boldsymbol{A}) = m$*). Then for any $\boldsymbol{y} \in \mathbb{R}^m$, the optimization problem*

$$\textit{minimize} \quad \|\boldsymbol{x}\|_2^2 \quad \textit{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \tag{A.6.9}$$

has a unique optimal solution,

$$\hat{\boldsymbol{x}}_{\ell^2} = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}\boldsymbol{y}. \tag{A.6.10}$$

Proof. The following inequality can be checked by directly expanding the right hand side:[2]

$$\forall\, \boldsymbol{x},\, \boldsymbol{x}' \in \mathbb{R}^n, \quad \|\boldsymbol{x}'\|_2^2 \;\geq\; \|\boldsymbol{x}\|_2^2 + \langle 2\boldsymbol{x}, \boldsymbol{x}' - \boldsymbol{x}\rangle + \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2\,. \tag{A.6.11}$$

If $\boldsymbol{x}$ and $\boldsymbol{x}'$ are feasible for our problem, then $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x}' = \boldsymbol{y}$, and so $\boldsymbol{x}' - \boldsymbol{x} \in \mathrm{null}(\boldsymbol{A})$. For any feasible $\boldsymbol{x}' \neq \hat{\boldsymbol{x}}_{\ell^2}$, we have

$$\begin{aligned}
\|\boldsymbol{x}'\|_2^2 \;&\geq\; \|\hat{\boldsymbol{x}}_{\ell^2}\|_2^2 + 2\,\langle \hat{\boldsymbol{x}}_{\ell^2}, \boldsymbol{x}' - \hat{\boldsymbol{x}}_{\ell^2}\rangle + \|\boldsymbol{x}' - \hat{\boldsymbol{x}}_{\ell^2}\|_2^2 \\
&=\; \|\hat{\boldsymbol{x}}_{\ell^2}\|_2^2 + 2\left\langle \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}\boldsymbol{y}, \boldsymbol{x}' - \hat{\boldsymbol{x}}_{\ell^2}\right\rangle + \|\boldsymbol{x}' - \hat{\boldsymbol{x}}_{\ell^2}\|_2^2 \\
&=\; \|\hat{\boldsymbol{x}}_{\ell^2}\|_2^2 + \underbrace{2\left\langle (\boldsymbol{A}\boldsymbol{A}^*)^{-1}\boldsymbol{y}, \boldsymbol{A}(\boldsymbol{x}' - \hat{\boldsymbol{x}}_{\ell^2})\right\rangle}_{=0} + \|\boldsymbol{x}' - \hat{\boldsymbol{x}}_{\ell^2}\|_2^2 \\
&>\; \|\hat{\boldsymbol{x}}_{\ell^2}\|_2^2\,.
\end{aligned} \tag{A.6.12}$$

□

The matrix $\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}$ is also called a pseudo-inverse of $\boldsymbol{A}$, and also denoted by $\boldsymbol{A}^\dagger$.[3]

A.7 Eigenvectors and Eigenvalues

Definition A.7.1 (Eigenvalue, eigenvector)**.** *Let* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$*. We say that* $\lambda \in \mathbb{C}$ *is an eigenvalue of* $\boldsymbol{A}$ *if there exists some nonzero vector* $\boldsymbol{v} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}$ *such that*

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}. \tag{A.7.1}$$

If we view $\boldsymbol{A}$ as corresponding to a linear map $\mathcal{L} : \mathbb{C}^n \to \mathbb{C}^n$, the definition says that $\mathcal{L}$ preserves the direction of the vector $\boldsymbol{v}$. If λ is an eigenvector of $\boldsymbol{A}$, with corresponding eigenvector $\boldsymbol{v}$, then $\boldsymbol{v} \in \mathrm{null}(\boldsymbol{A} - \lambda\boldsymbol{I})$, and hence $\mathrm{rank}\,(\boldsymbol{A} - \lambda\boldsymbol{I}) < n$. Using the determinant criterion for singularity, we obtain

Theorem A.7.2. $\lambda \in \mathbb{C}$ *is an eigenvalue of* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *if and only if it is a root of the characteristic polynomial*

$$\chi(\lambda) = \det(\boldsymbol{A} - \lambda\boldsymbol{I}), \tag{A.7.2}$$

i.e., $\chi(\lambda) = 0$.

---

[2] If you have background in convexity, this follows immediately from the strong convexity of $\|\cdot\|_2^2$.

[3] The fact that we have apparently used the notation $\boldsymbol{A}^\dagger$ for two different things is resolved if we consider the general form of the pseudo-inverse, which is written in terms of the singular value decomposition (SVD). We will do this after reviewing the SVD in Section A.8.

This implies that every matrix $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ has n complex eigenvalues, counted with multiplicity. Often we are interested in real matrices $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. It is important to note that *the eigenvalues of a real matrix are not necessarily real*. There is one important special case in which the eigenvalues are guaranteed to be real: symmetric matrices. A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is *symmetric* if

$$\boldsymbol{A} = \boldsymbol{A}^*. \tag{A.7.3}$$

The eigenvalues of a symmetric matrix are necessarily real, with corresponding real eigenvectors. Moreover, it is not difficult to prove that if $\boldsymbol{v}$ and $\boldsymbol{v}'$ are eigenvectors of a symmetric matrix corresponding to *distinct* eigenvalues $\lambda \neq \lambda'$, then they are orthogonal: $\boldsymbol{v} \perp \boldsymbol{v}'$. From this, we obtain the eigenvector decomposition of a symmetric matrix:

Theorem A.7.3 (Eigenvector decomposition)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be symmetric. Then there exist orthonormal vectors $\boldsymbol{v}_1 \dots \boldsymbol{v}_n \in \mathbb{R}^n$ and real scalars $\lambda_1 \geq \cdots \geq \lambda_n$, such that if we write*

$$\boldsymbol{V} = [\boldsymbol{v}_1 \mid \cdots \mid \boldsymbol{v}_n] \in O(n), \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \in \mathbb{R}^{n \times n}, \tag{A.7.4}$$

we have

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^*, \tag{A.7.5}$$

The expression $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^*$ is sometimes written as $\boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^*$. Theorem A.7.3 leads to the following variational characterization of the eigenvalues, which is useful both for analytical purposes and for identifying optimization problems that can be solved directly via eigenvector decomposition:

Theorem A.7.4 (Variational characterization of eigenvalues)**.** *The first eigenvalue λ_1 of a symmetric matrix $\boldsymbol{A}$ is the optimal value of the problem*

$$\begin{array}{ll} \text{maximize} & \boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{x}\|_2^2 = 1. \end{array} \tag{A.7.6}$$

Moreover, every optimizer $\hat{\boldsymbol{x}}$ is an eigenvector corresponding to λ_1. Similarly, the optimal value of

$$\begin{array}{ll} \text{minimize} & \boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{x}\|_2^2 = 1 \end{array} \tag{A.7.7}$$

is $\lambda_n(\boldsymbol{A})$. For the intermediate eigenvalues, if $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1}$ are any mutually orthogonal eigenvectors corresponding to $\lambda_1, \dots, \lambda_{k-1}$, we have that λ_k

is the optimal value for

$$\begin{array}{ll} \text{maximize} & \boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{x}\|_2^2 = 1,\ \boldsymbol{x} \perp \boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1}. \end{array} \tag{A.7.8}$$

From the previous result, it seems the eigenvector decomposition is a very useful tool for studying quadratic forms $q(\boldsymbol{x}) = \boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x}$. Matrices $\boldsymbol{A}$ for which $q(\boldsymbol{x})$ is always positive are especially important:

Definition A.7.5 (Positive definiteness)**.** *A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is* positive definite *if for all nonzero $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x} > 0$. It is* positive semidefinite *if for all $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x} \geq 0$.*

If $\boldsymbol{A}$ is positive definite, we write

$$\boldsymbol{A} \succ \boldsymbol{0}. \tag{A.7.9}$$

If $\boldsymbol{A}$ is positive semidefiniite, we write

$$\boldsymbol{A} \succeq \boldsymbol{0}. \tag{A.7.10}$$

More generally, for symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we write $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is semidefinite, i.e., $\boldsymbol{A} - \boldsymbol{B} \succeq \boldsymbol{0}$. This defines a *partial order* on the symmetric matrices, which we call the *semidefinite order*.

Theorem A.7.6. *A symmetric matrix $\boldsymbol{A}$ is positive definite (resp. semidefinite) if and only if all of its eigenvalues are positive (resp. nonnegative).*

Location of eigenvalues.

It is often useful to be able to characterize, in terms of the properties of $\boldsymbol{A}$, where the eigenvalues $\lambda \in \mathbb{C}$ are located. For example, we saw that if $\boldsymbol{A}$ is a symmetric matrix, the eigenvalues lie on the real axis. For general $\boldsymbol{A}$, the situation is more complicated. However, we do have the following result of Gershgorin, which states that the eigenvalues must live in a union of discs, centered about the diagonal elements $\boldsymbol{A}_{ii}$ of $\boldsymbol{A}$:

Theorem A.7.7 (Gershgorin disc theorem)**.** *Let $\boldsymbol{A} \in \mathbb{C}^{n \times n}$, and let $\lambda \in \mathbb{C}$ and $\boldsymbol{v} \in \mathbb{C}^n$ be an eigenvalue-eigenvector pair. Then there exists some $i \in \{1, \dots, n\}$ such that*

$$|\lambda - \boldsymbol{A}_{ii}| \ \leq\ \sum_{j \neq i} |\boldsymbol{A}_{ij}|. \tag{A.7.11}$$

This result is called the Gershgorin disc theorem, because it implies that in the complex plane $\boldsymbol{C}$, each eigenvalue λ lies in a union of discs D_i with centers $\boldsymbol{A}_{ii}$ and radii $r_i = \sum_{j \neq i} |\boldsymbol{A}_{ij}|$. It is most powerful when the off-diagonal elements of $\boldsymbol{A}$ are small. Numerous variants and refinements are known.

A.8 The Singular Value Decomposition (SVD)

Definitions.

The eigenvector decomposition $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^*$ defined in Theorem A.7.3 provides an essential tool for studying symmetric matrices $\boldsymbol{S}$. In particular, it shows that with an appropriate rotation of the space, a symmetric matrix acts like a diagonal matrix. It would be very useful to have a similar representation for general matrices, including non-symmetric square matrices, and rectangular matrices. The *singular value decomposition* goes much of the way, allowing us to find bases for the domain and range of a linear map with respect to which it becomes quite simple:

Theorem A.8.1 (Compact SVD, existence)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, with* $\mathrm{rank}(\boldsymbol{A}) = r$*. There exist scalars $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and matrices $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times r}$ with orthonormal columns ($\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{I}$) such that if we set*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix} \in \mathbb{R}^{r\times r}, \tag{A.8.1}$$

we have

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*. \tag{A.8.2}$$

The σ_i are called singular values *of $\boldsymbol{A}$, while the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are called the (left and right, respectively) singular vectors.*

The expression in Theorem A.8.1 can be used to express $\boldsymbol{A}$ as a sum of r orthogonal rank-one matrices:

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*. \tag{A.8.3}$$

The compact SVD immediately reveals several important properties of $\boldsymbol{A}$:

Theorem A.8.2 (Properties of the compact SVD)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, with compact SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. Then*

- $\mathrm{range}(\boldsymbol{A}) = \mathrm{range}(\boldsymbol{U})$*. The columns of $\boldsymbol{U}$ are an orthonormal basis for the range of $\boldsymbol{A}$.*
- $\mathrm{range}(\boldsymbol{A}^*) = \mathrm{range}(\boldsymbol{V})$*. The columns of $\boldsymbol{V}$ are an orthonormal basis for the row space of $\boldsymbol{A}$.*

Occasionally it is useful to extend $\boldsymbol{U}$ and $\boldsymbol{V}$ to orthogonal matrices, giving the full singular value decomposition:

Theorem A.8.3 (Full SVD)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then there exist $\boldsymbol{U} \in O(m)$, $\boldsymbol{V} \in O(n)$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ such that*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*, \tag{A.8.4}$$

$\boldsymbol{\Sigma}$ is diagonal (i.e., $\boldsymbol{\Sigma}_{ij} = 0$ for $i \neq j$), and $\boldsymbol{\Sigma}_{11} \geq \boldsymbol{\Sigma}_{22} \geq \cdots \geq \boldsymbol{\Sigma}_{\min\{m,n\},\min\{m,n\}} \geq 0$.

It is sometimes a point of confusion that the notation for the full SVD and the compact SVD coincide. In this course, we will mostly work with the compact SVD.

Approximation Properties.

The SVD provides an immediate solution to several approximation problems. Most fundamentally, it gives a way of forming a best rank-r approximation to $\boldsymbol{A}$:

Theorem A.8.4 (Best rank-r approximation)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ have singular value decomposition*

$$\boldsymbol{A} = \sum_{i=1}^{\min\{m,n\}} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*. \tag{A.8.5}$$

Then an *optimal solution to the rank-r approximation problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X} - \boldsymbol{A}\|_F \\ \text{subject to} & \operatorname{rank}(\boldsymbol{X}) \leq r \end{array} \tag{A.8.6}$$

is the truncated SVD

$$\widehat{\boldsymbol{A}}_r = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*. \tag{A.8.7}$$

If $\sigma_r(\boldsymbol{A}) > \sigma_{r+1}(\boldsymbol{A})$, then the solution is unique.

Interestingly, if we change $\|\cdot\|_F$ to other unitary invariant matrix norms (such as the operator norm), the above result remains unchanged. The SVD also gives a way of optimally approximating a given square matrix with an orthogonal matrix:

Theorem A.8.5 (Best orthogonal approximation)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, and let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be any full singular value decomposition of $\boldsymbol{A}$. Then an optimal solution to the problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X} - \boldsymbol{A}\|_F \\ \text{subject to} & \boldsymbol{X} \in O(n) \end{array} \tag{A.8.8}$$

is given by $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^*$.

A.9 Vector and Matrix Norms

Norms on vector spaces.

A *norm* on a vector space $\mathbb{V}$ gives a way of measuring lengths of vectors, that conforms in important ways to our intuition from lengths in $\mathbb{R}^3$. Formally:

Definition A.9.1. *A norm on a real vector space $\mathbb{V}$ is a function $\| \cdot \| : \mathbb{V} \to \mathbb{R}$ that is*

1. **Nonnegatively homogeneous***: $\|\alpha \boldsymbol{x}\| = |\alpha| \|\boldsymbol{x}\|$ for all vectors $\boldsymbol{x} \in \mathbb{V}$, scalars $\alpha \in \mathbb{R}$,*
2. **Positive definite***: $\|\boldsymbol{x}\| \geq 0$, and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$*
3. **Subadditive***: $\|\cdot\|$ satisfies the triangle inequality $\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\|+\|\boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{V}$.*

One very important family of norms are the ℓ^p norms. If we take $\mathbb{V} = \mathbb{R}^n$, and $p \in [1, \infty)$, we can write

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p \right)^{1/p}. \tag{A.9.1}$$

The most familiar example is the ℓ^2 norm or "Euclidean norm"

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_i x_i^2} = \sqrt{\boldsymbol{x}^* \boldsymbol{x}}$$

which coincides with our usually way of measuring lengths. Two other cases are of almost equal importance: $p = 1$, and $p \to \infty$. Setting $p = 1$ in (A.9.1), we obtain

$$\|\boldsymbol{x}\|_1 = \sum_i |x_i|, \tag{A.9.2}$$

Finally, as p becomes larger, the expression in (A.9.1) accentuates large $|x_i|$. As $p \to \infty$, $\|\boldsymbol{x}\|_p \to \max_i |x_i|$. We extend the definition of the ℓ^p norm to $p = \infty$ by defining

$$\|\boldsymbol{x}\|_\infty = \max_i |x_i|. \tag{A.9.3}$$

However, the ℓ^p norms are far from the only norms on vectors.

Example A.9.2. *The following are examples of norms:*

- *For $p \geq 1$, $\|\boldsymbol{x}\|_p$ is a norm.*
- *Every positive definite matrix $\boldsymbol{P} \succ \boldsymbol{0}$ defines a norm, via $\|\boldsymbol{x}\|_{\boldsymbol{P}} = \sqrt{\boldsymbol{x}^* \boldsymbol{P} \boldsymbol{x}}$.*

- *For $\boldsymbol{x} \in \mathbb{R}^n$, $[\boldsymbol{x}]_{(k)}$ denote the k-th largest element of the sequence $|\boldsymbol{x}_1|, |\boldsymbol{x}_2|, \dots, |\boldsymbol{x}_n|$. Then*

$$\|\boldsymbol{x}\|_{[K]} = \sum_{k=1}^{K} [\boldsymbol{x}]_{(k)} \tag{A.9.4}$$

is a norm.

- *For $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, the Frobenius norm $\|\boldsymbol{X}\|_F = \sqrt{\langle \boldsymbol{X}, \boldsymbol{X} \rangle}$ is a norm.*

One fundamental result in the theory of normed spaces is that in finite dimensions, all norms are comparable:

Theorem A.9.3 (Equivalence of norms)**.** *Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two norms on a finite dimenisonal vector space $\mathbb{V}$. Then there exist $\alpha, \beta > 0$ such that for every $\boldsymbol{v} \in \mathbb{V}$,*

$$\alpha \|\boldsymbol{v}\|_a \;\le\; \|\boldsymbol{v}\|_b \;\le\; \beta \|\boldsymbol{v}\|_a . \tag{A.9.5}$$

It is important not to over-interpret this result. "Equivalence" here means that the values of the norms can be compared up to constants, as in (A.9.5). It does not mean that the norms behave in the same way – they may produce very different results when selected to define constraint sets, or as objective functions for optimization. For purposes of analysis, it is useful to note the following comparisons

Lemma A.9.4 (Comparisons between ℓ^p norms)**.** *For all $\boldsymbol{x} \in \mathbb{R}^n$,*

- $\|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1 \le \sqrt{n}\, \|\boldsymbol{x}\|_2$,
- $\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2 \le \sqrt{n}\, \|\boldsymbol{x}\|_\infty$,
- $\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_1 \le n\, \|\boldsymbol{x}\|_\infty$.

To each norm, we can associate a *dual norm.* To do this precisely, we need to define a normed linear space. If $\mathbb{V}$ is a vector space and $\|\cdot\|$ is a norm on $\mathbb{V}$, we call the pair $(\mathbb{V}, \|\cdot\|)$ a *normed linear space.* A *linear functional* is a linear map $\boldsymbol{\phi} : \mathbb{V} \to \mathbb{R}$. Since linear combinations of linear functionals are again linear functionals, the space of all linear functionals on a given vector space $\mathbb{V}$ is itself a vector space (called the "topological dual" of $\mathbb{V}$). On this space, we can define another function

$$\|\boldsymbol{\phi}\|^* = \sup_{\boldsymbol{v} \in \mathbb{V},\ \|\boldsymbol{v}\| \le 1} |\phi(\boldsymbol{v})|. \tag{A.9.6}$$

As the notation suggests, $\|\boldsymbol{\phi}\|^*$ is a norm, if we restrict to $\boldsymbol{\phi}$ for which the supremum is finite:

Definition A.9.5 (Dual space, dual norm)**.** *The normed dual of the space $(\mathbb{V}, \|\cdot\|)$ is the space $(\mathbb{V}^*, \|\cdot\|^*)$, were the dual norm $\|\cdot\|^*$ of a linear functional $\boldsymbol{\phi} : \mathbb{V} \to \mathbb{R}$ is defined as in (A.9.6) and*

$$\mathbb{V}^* = \left\{ \boldsymbol{\phi} : \mathbb{V} \to \mathbb{R} \textit{ linear } \mid \|\boldsymbol{\phi}\|^* < +\infty \right\}. \tag{A.9.7}$$

This definition may seem somewhat abstract; for our purposes, the dual spaces and dual norms we encounter will have fairly concrete descriptions:

Theorem A.9.6. *Let $\langle \cdot, \cdot \rangle$ denote the standard inner product on $\mathbb{R}^n$ (and by extension on $\mathbb{R}^{m\times n}$). Every linear functional $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}$ can be written as*

$$\boldsymbol{\phi}(\boldsymbol{x}) = \langle \boldsymbol{v}, \boldsymbol{x} \rangle \,, \tag{A.9.8}$$

for some vector $\boldsymbol{v} \in \mathbb{R}^n$. Similarly, every linear functional $\boldsymbol{\phi} : \mathbb{R}^{m\times n} \to \mathbb{R}$ can be written as

$$\boldsymbol{\phi}(\boldsymbol{X}) = \langle \boldsymbol{V}, \boldsymbol{X} \rangle \,, \tag{A.9.9}$$

for some matrix $\boldsymbol{V} \in \mathbb{R}^{m\times n}$.

The implication of this is that if we are considering a space $(\mathbb{R}^n, \|\cdot\|_\sharp)$, the dual space can be identified with $(\mathbb{R}^n, \|\cdot\|_\sharp^*)$, where

$$\|\boldsymbol{v}\|_\sharp^* = \sup_{\|\boldsymbol{x}\|_\sharp \le 1} \langle \boldsymbol{v}, \boldsymbol{x} \rangle \,. \tag{A.9.10}$$

In particular, we have the following examples:

Example A.9.7 (Duals of common norms)**.** *Check the following:*

- *The dual of the ℓ^∞ norm is the ℓ^1 norm.*
- *The dual of the ℓ^1 norm is the ℓ^∞ norm.*
- *The ℓ^2 and Frobenius norms are self-dual; i.e., $\|\cdot\|_2^* = \|\cdot\|_2$ and $\|\cdot\|_F^* = \|\cdot\|_F$.*
- *If $p, q \in [1, \infty)$, with $p^{-1} + q^{-1} = 1$, then $\|\cdot\|_p^* = \|\cdot\|_q$ and $\|\cdot\|_q^* = \|\cdot\|_p$.*

It is immediate from the definition that for any $\boldsymbol{x}, \boldsymbol{x}'$, and any norm $\|\cdot\|$,

$$\langle \boldsymbol{x}, \boldsymbol{x}' \rangle \;\le\; \|\boldsymbol{x}\| \, \|\boldsymbol{x}'\|^* \,. \tag{A.9.11}$$

If we take $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2$, we obtain the Cauchy-Schwarz inequality.

Matrix and operator norms.

Even more interesting structure can arise when $\mathbb{V}$ is a space of matrices, e.g., $\mathbb{V} = \mathbb{R}^{m\times n}$, due to the interpretation of a matrix as a linear operator. For square matrices, many authors reserve the term "matrix norm" for a function $\|\cdot\|$ that satisfies the three criteria in Definition A.9.1, *and* is *submultiplicative*

$$\|\boldsymbol{A}\boldsymbol{B}\| \;\le\; \|\boldsymbol{A}\| \, \|\boldsymbol{B}\| \,. \tag{A.9.12}$$

They use the term "vector norm on matrices" for functions on $\mathbb{V}$ that only satisfy Definition A.9.1. We will not emphasize this distinction in terminology. Nevertheless, the submultiplicative property (A.9.12) is often useful, and we will note it where it occurs.

The most important source of norms on matrices comes from the notion of a matrix as a linear operator:

Definition A.9.8 (Operator norm)**.** *Let* $(\mathbb{W}, \|\cdot\|_a)$ *and* $(\mathbb{W}', \|\cdot\|_b)$ *be two normed linear spaces, and let* $\mathcal{L} : \mathbb{W} \to \mathbb{W}'$*. The* operator norm *of* $\mathcal{L}$ *is*

$$\|\mathcal{L}\|_{a \to b} = \sup_{\|\boldsymbol{w}\|_a \le 1} \|\mathcal{L}[\boldsymbol{w}]\|_b \,. \tag{A.9.13}$$

Specializing the definition a bit, for an $m \times n$ matrix $\boldsymbol{A}$, if $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, we write

$$\|\boldsymbol{A}\|_{a \to b} = \sup_{\|\boldsymbol{x}\|_a \le 1} \|\boldsymbol{A}\boldsymbol{x}\|_b \,. \tag{A.9.14}$$

The most important special case is

Theorem A.9.9. *The norm of a matrix* $\boldsymbol{A}$ *as an operator from* $\ell^2_n = (\mathbb{R}^n, \|\cdot\|_2)$ *to* $\ell^2_m = (\mathbb{R}^m, \|\cdot\|_2)$ *is*

$$\|\boldsymbol{A}\|_{2 \to 2} = \sigma_1(\boldsymbol{A}). \tag{A.9.15}$$

Several other cases are of interest:

Theorem A.9.10. *The norm of any matrix as an operator from* $(\mathbb{R}^n, \|\cdot\|_1)$ *to any normed space* $(\mathbb{R}^m, \|\cdot\|_\sharp)$ *is simply the largest* $\|\cdot\|_\sharp$ *of any column of* $\boldsymbol{A}$*:*

$$\|\boldsymbol{A}\|_{1 \to \sharp} = \max_{j=1,\dots,n} \|\boldsymbol{A}\boldsymbol{e}_j\|_\sharp \,. \tag{A.9.16}$$

The norm of any matrix as an operator from $(\mathbb{R}^n, \|\cdot\|_\flat)$ *for any norm* $\|\cdot\|_\flat$ *into* $(\mathbb{R}^m, \|\cdot\|_\infty)$ *is the largest dual norm of any of the rows:*

$$\|\boldsymbol{A}\|_{\sharp \to \infty} = \max_{i=1,\dots,m} \|\boldsymbol{e}_i^* \boldsymbol{A}\|_\sharp^* \,, \tag{A.9.17}$$

where the dual norm $\|\cdot\|_\flat^*$ *is*

$$\|\boldsymbol{v}\|_\flat^* = \sup_{\|\boldsymbol{u}\|_\flat \le 1} \langle \boldsymbol{u}, \boldsymbol{v} \rangle \,. \tag{A.9.18}$$

For example, $\|\boldsymbol{A}\|_{1 \to 1}$ is just the largest ℓ^1 norm of any column of $\boldsymbol{A}$.

Unitary invariant matrix norms.

It is interesting to note that the operator norm of a matrix $\boldsymbol{A}$ depends only on the singular values of $\boldsymbol{A}$:

$$\|\boldsymbol{A}\|_{2,2} = \sigma_1(\boldsymbol{A}) = \|\boldsymbol{\sigma}(\boldsymbol{A})\|_\infty \,, \tag{A.9.19}$$

where $\boldsymbol{\sigma}(\boldsymbol{A})$ is the vector of singular values. In fact, the Frobenius norm $\|\boldsymbol{A}\|_F$ depends only on the singular values as well:

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{\min m,n} \sigma_i(\boldsymbol{A})^2} = \|\boldsymbol{\sigma}(\boldsymbol{A})\|_2 \,. \tag{A.9.20}$$

This fact is not too difficult to observe from the orthogonal invariance of $\|\cdot\|_F$:

$$\forall\, \boldsymbol{A} \in \mathbb{R}^{m\times n},\ \boldsymbol{P} \in O(m),\ \boldsymbol{Q} \in O(n), \quad \|\boldsymbol{PAQ}\|_F = \|\boldsymbol{A}\|_F\,. \tag{A.9.21}$$

This suggests a pattern. In fact, any ℓ^p norm of the singular values is a norm on matrices $\boldsymbol{A}$:

Definition A.9.11 (Schatten p-norm)**.** *For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, let $\boldsymbol{\sigma}(\boldsymbol{A}) \in \mathbb{R}^{\min\{m,n\}}$ denote the vector of singular values. For $p \in [1,\infty]$, the function*

$$\|\boldsymbol{A}\|_{S_p} = \|\boldsymbol{\sigma}(\boldsymbol{A})\|_p \tag{A.9.22}$$

is a norm on $\mathbb{R}^{m\times n}$.

It is easy to recognize the operator norm and Frobenius norm as special cases. One other special case is of great interest – the Schatten 1-norm

$$\|\boldsymbol{A}\|_{S_1} = \sum_i \sigma_i(\boldsymbol{A}). \tag{A.9.23}$$

This is also sometimes called the *trace norm* or *nuclear norm*. We reserve a special notation

$$\|\boldsymbol{A}\|_* = \sum_i \sigma_i(\boldsymbol{A}) \tag{A.9.24}$$

for this norm. The operator norm $\|\cdot\|_{2,2}$ and the nuclear norm $\|\cdot\|_*$ are dual norms.

We have defined several interesting, useful norms on matrices $\boldsymbol{A}$, by applying different vector norms to the singular values $\boldsymbol{\sigma}(\boldsymbol{A})$. Because the singular values are orthogonal invariant, i.e., for $\boldsymbol{P} \in O(m)$, $\boldsymbol{Q} \in O(n)$, $\boldsymbol{\sigma}(\boldsymbol{PAQ}) = \boldsymbol{\sigma}(\boldsymbol{A})$, norms defined in this way are also orthogonal invariant. It is natural to ask whether every function $\|\boldsymbol{\sigma}(\boldsymbol{A})\|$ generates a valid norm on $\mathbb{R}^{m\times n}$. It turns out that with several restrictions, this is true.

Definition A.9.12 (Symmetric gauge function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is a* symmetric gauge function *if it satisfies the following three conditions:*

- **Norm:** *f is a norm on $\mathbb{R}^n$.*
- **Permutation invariance:** *For every $\boldsymbol{x} \in \mathbb{R}^n$ and permutation matrix $\boldsymbol{\Pi}$, $f(\boldsymbol{\Pi x}) = f(\boldsymbol{x})$.*
- **Symmetry:** *For every $\boldsymbol{x} \in \mathbb{R}^n$ and diagonal sign matrix $\boldsymbol{\Sigma}$ (i.e., matrix with diagonal entries ± 1), $f(\boldsymbol{\Sigma x}) = f(\boldsymbol{x})$.*

Theorem A.9.13 (Von Neumann's characterization of unitary invariant norms)**.** *Fix $m \geq n$. For $\boldsymbol{M} \in \mathbb{C}^{m\times n}$, let $\boldsymbol{\sigma}(\boldsymbol{M}) \in \mathbb{R}^n$ denote its vector of singular values. Then for every symmetric gauge function $f_\sharp$,*

$$\|\boldsymbol{M}\|_\sharp \doteq f_\sharp(\boldsymbol{\sigma}(\boldsymbol{M})) \tag{A.9.25}$$

defines a unitary invariant matrix norm on $\mathbb{C}^{m \times n}$. *Conversely, for every unitary invariant matrix norm* $\|\boldsymbol{M}\|_\flat$ *there exists a symmetric gauge function* $f_\flat$ *such that* $\|\boldsymbol{M}\|_\flat = f_\flat(\boldsymbol{\sigma}(\boldsymbol{M}))$.