Mathias Richter
Stefan Schäffler

Inverse Probleme mit stochastisch modellierten Messdaten

Stochastische und numerische Methoden der Diskretisierung und Optimierung

Springer Spektrum

Inverse Probleme mit stochastisch modellierten Messdaten

Mathias Richter · Stefan Schäffler

Inverse Probleme mit stochastisch modellierten Messdaten

Stochastische und numerische Methoden der Diskretisierung und Optimierung

Springer Spektrum

Mathias Richter
Fakultät für Elektro- und
Informationstechnik (EIT)
Universität der Bundeswehr München
Neubiberg, Deutschland

Stefan Schäffler
Fakultät für Elektro- und
Informationstechnik (EIT)
Universität der Bundeswehr München
Neubiberg, Deutschland

ISBN 978-3-662-66342-4 ISBN 978-3-662-66343-1 (eBook)
https://doi.org/10.1007/978-3-662-66343-1

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Planung/Lektorat: Nikoo Azarm
Springer Spektrum ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Vorwort

In vielen wissenschaftlichen Teildisziplinen sowie in vielen technisch-industriellen Fragestellungen spielen inverse Probleme eine zentrale Rolle; dabei ist man vor die Entscheidungssituation gestellt, aus einer (im Allgemeinen durch Messungen) beobachteten Wirkung auf die entsprechende Ursache zurückschließen zu müssen. Das klassische Beispiel in diesem Zusammenhang ist sicher die Computertomographie, bei der die Wirkung aus der Ablenkung und veränderten Intensität von Strahlen besteht, die zum Beispiel durch spezielles Gewebe im Körper verursacht werden. Ziel ist es, dieses Gewebe im Körper zu identifizieren und durch bildgebende Verfahren zu visualisieren. Jedes Handy, das ein durch die Übertragung gestörtes Signal empfängt (Wirkung), benötigt Methoden, um das ursprünglich gesendete Signal (Ursache) zu rekonstruieren, um die vom Sender gewünschte Information an den Empfänger weitergeben zu können. In der Volkswirtschaft ist es von großer Bedeutung, aus der Beobachtung gewisser Kenngrößen auf ihre Ursachen schließen zu können, um so rechtzeitig Fehlentwicklungen vorbeugen zu können. Das mathematische Teilgebiet der „inversen Probleme“ gehört somit zu den für die Anwendungen wichtigsten mathematischen Disziplinen.

Obwohl sich zufällig gestörte Messungen als Beobachtung einer Wirkung in den Anwendungen nicht vermeiden lassen, finden sie in den mathematischen Abhandlungen bisher praktisch keine Beachtung. Das vorliegende Buch bindet zum ersten Mal die stochastische Modellierung von Messdaten in alle Aspekte der Analyse inverser Probleme mit ein; dies erfordert natürlich eine wesentliche Erweiterung der benötigten mathematischen Grundlagen. Dem wurde auch im Hinblick darauf, dass dieses Buch zum Selbststudium für theoretisch Interessierte und für Anwender geeignet sein soll, entsprechend Rechnung getragen.

München, im September 2022 *Mathias Richter, Stefan Schäffler*

Einleitung

Der Begriff „inverses Problem“ basiert nicht auf einem mathematischen, sondern auf einem physikalisch-technischen Hintergrund. Ein inverses Problem liegt immer dann vor,

- wenn durch eine gegebene Abbildung $F : \mathcal{U} \to \mathcal{W}$ etwa im Sinne der Physik der Kausalzusammenhang zwischen einer Ursache $u \in \mathcal{U}$ und der entsprechenden Wirkung $F(u) \in \mathcal{W}$ modelliert wird
- und wenn die mathematische Aufgabe darin besteht, aus einer (gemessenen oder gewünschten) Wirkung $w \in \mathcal{W}$ auf die entsprechende Ursache $\hat{u} \in \mathcal{U}$ zu schließen.

Die Berechnung der Wirkung $F(u)$ einer gegebenen Ursache heißt „direktes Problem“. Inverse Probleme lassen sich in zwei Gruppen gliedern:

- Identifikationsprobleme: Die beobachtete Wirkung w wurde durch Messungen gewonnen (z.B. medizinische Diagnostik: Computertomographie).
- Steuerungsprobleme: Die Wirkung w ist gewünscht (z.B. optimale Flugbahn einer Raumsonde) und es stellt sich die Frage, ob und welche Ursachen diese Wirkung erzielen.

Da für die Untersuchung inverser Probleme bei stochastisch modellierten Daten nur Messdaten in Frage kommen, werden im Folgenden nur Identifikationsprobleme behandelt. Betrachten wir dazu ein Beispiel.

Bei der Übertragung eines analogen Signals

$$u : [t_1, t_2] \to \mathbb{R}$$

von einem Sender zu einem Empfänger kommt es dort abhängig von den Eigenschaften des Übertragungskanals zum Empfang zeitlich verzögerter Kopien des Signals u zum Beispiel durch Reflexion an Gebäuden. Verwendet man eine Funktion $g : \mathbb{R} \to \mathbb{R}$ als Modell für den Übertragungskanal, so wird das empfangene Signal w durch

$$w : \mathbb{R} \to \mathbb{R}, \quad s \mapsto \int_{t_1}^{t_2} g(s-t)u(t)dt$$

beschrieben. Somit läßt sich das direkte Problem folgendermaßen formulieren:

- *Ursache: Das zu übertragende Signal u*
- *Wirkung: Das empfangene Signal w*

- *Modellierung:*

$$w(s) = F(u)(s) = \int_{t_1}^{t_2} g(s-t)u(t)dt, \quad s \in \mathbb{R},$$

unter Verwendung des Übertragungskanals g.

Das inverse Problem besteht nun darin, aus den zu verschiedenen Zeitpunkten gemessenen Amplituden des empfangenen Signals w das gesendete Signal u zu rekonstruieren. Betrachtet man etwa das Signal

$$u : [-0.5, 0.5] \to \mathbb{R}, \quad t \mapsto 0.25 - t^2,$$

und als Modell für den Übertragungskanal die Funktion

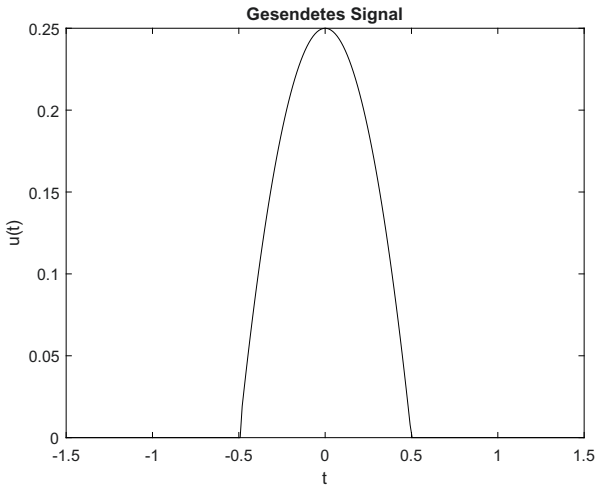

Abb. 0.1 *Gesendetes Signal*

$$g : \mathbb{R} \to \mathbb{R}, \quad t \mapsto e^{-10t^2},$$

so ergeben sich die verschiedenen Amplituden von w durch

$$w(t_i) = \int_{-0.5}^{0.5} e^{-10(t_i - t)^2}(0.25 - t^2)dt, \quad t_1 < \ldots < t_n.$$

Seien nun Messwerte

$$w(t_1), \ldots, w(t_{257})$$

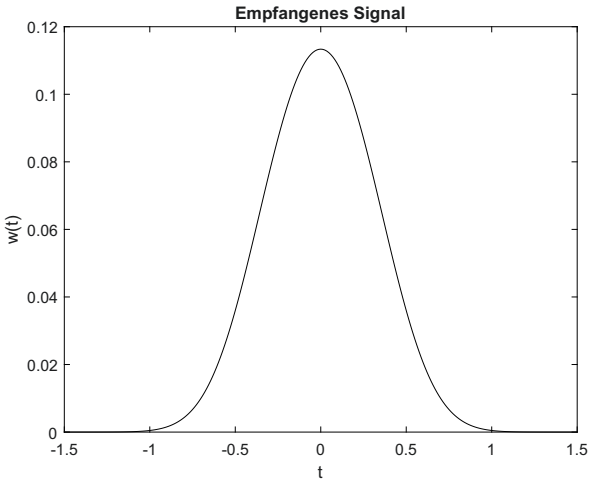

Abb. 0.2 *Empfangenes Signal*

zu äquidistanten Zeitpunkten

$$-1.5 = t_1 < \ldots < t_{257} = 1.5$$

des Signals w gegeben. Da die Signaldauer von u nicht a priori bekannt sein muss und da die Messwerte von w zu Zeitpunkten $t_i \in [-1.5, 1.5]$ *zur Verfügung stehen, bietet es sich an, das unbekannte Signal u im Intervall* $[-1.5, 1.5]$ *zu rekonstruieren und dafür eine Fourier-Entwicklung*

$$\hat{u} : \mathbb{R} \to \mathbb{R}, \quad t \mapsto \frac{a_0}{2} + \sum_{j=1}^{n} \left(a_j \cos\left(\frac{2\pi j}{3} t\right) + b_j \sin\left(\frac{2\pi j}{3} t\right) \right)$$

zu verwenden. Die Fourier-Koeffizienten werden durch die numerische Behandlung des resultierenden linearen Ausgleichsproblems bestimmt (siehe Abb. 0.3 für $n = 8$*).*

In diesem Beispiel ist die rechte Seite w der Gleichung

$$F(u) = w$$

mit $F : \mathcal{U} \to \mathcal{W}$ nicht vollständig gegeben, sondern nur partiell; im Falle eines Funktionenraumes $\mathcal{W}$ sind zum Beispiel häufig nur Funktionswerte $w(t_i)$ an gewissen Argumenten $t_1, \ldots, t_n$ gemessen worden. Formal wird deshalb ein Beobachtungsoperator

$$\Psi : \mathcal{W} \to \mathcal{M}$$

eingeführt, der jeder möglichen Wirkung w die entsprechende Beobachtung/Messung $\Psi(w) \in \mathcal{M}$ zuordnet.

Bei der Verwendung von Messdaten ist neben der Tatsache, dass im Allgemeinen die Wirkung nur partiell beobachtet werden kann, auch die Frage nach Rauscheffekten in den Messungen von zentraler Bedeutung. Betrachtet man die Messungen

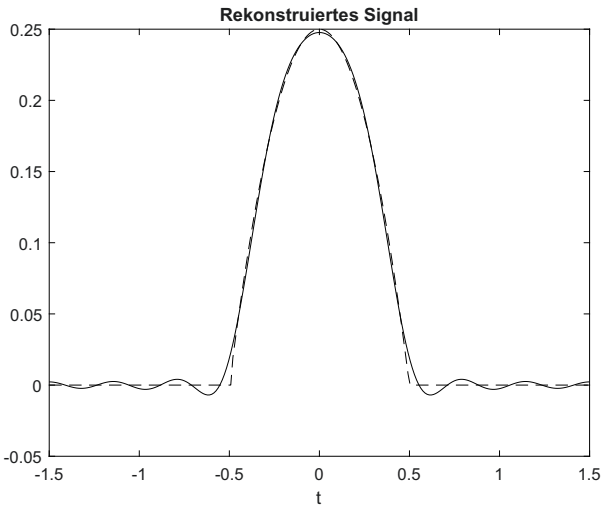

Abb. 0.3 *Rekonstruiertes Signal*

als exakt ohne Modellierung der Messfehler, so kann dies zu unbefriedigenden Ergebnissen führen.

Führen wir das obige Beispiel weiter und nehmen nun an, dass die 257 *Messwerte von w wie in Abbildung 0.4 vorliegen. Verwendet man erneut den Ansatz*

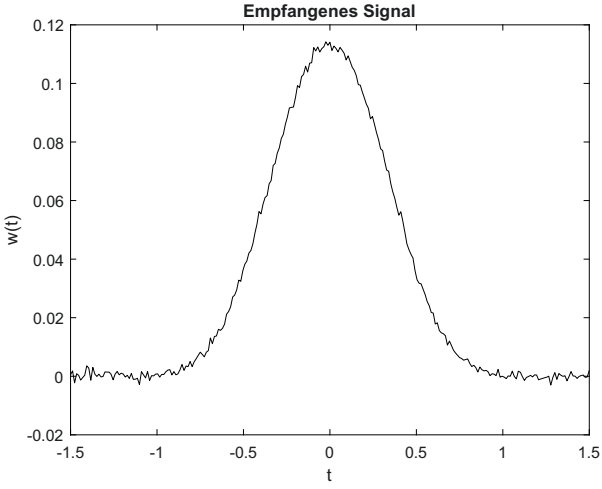

Abb. 0.4 *Empfangenes Signal*

$$\hat{u} : \mathbb{R} \to \mathbb{R}, \quad t \mapsto \frac{a_0}{2} + \sum_{j=1}^{8} \left(a_j \cos\left(\frac{2\pi j}{3} t\right) + b_j \sin\left(\frac{2\pi j}{3} t\right) \right)$$

und löst das resultierende lineare Ausgleichsproblem, so erhält man eine unbrauchbare Rekonstruktion von u (siehe Abb. 0.5). Dies resultiert aus der Tatsache, dass bei dieser Vorgehensweise unterstellt wird, dass û die Messfehler verursacht hat.

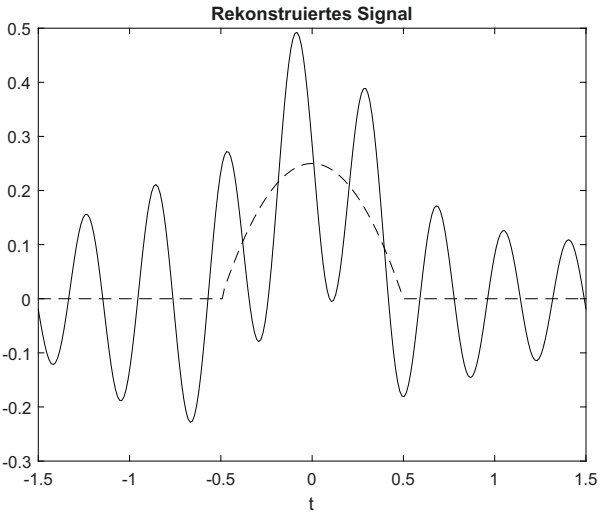

Abb. 0.5 *Rekonstruiertes Signal*

Wie dieses Beispiel zeigt, können Messfehler bei der Rekonstruktion einer Ursache u zu einer gemessenen Wirkung $\Psi(w)$ zu einem völlig unbrauchbaren Ergebnis führen. In diesem Buch werden daher die Messfehler stochastisch durch ein additives Rauschen modelliert; dies hat den Vorteil, dass man bei der Rekonstruktion der Ursache u die Varianz des Rauschens verwenden kann, um den Einfluss der Messfehler zu minimieren, was einerseits zu einer Glättung der Messdaten führt, andererseits eine problemadäquate Regularisierung des inversen Problems ergibt.

In unserem Beispiel modellieren wir die Messfehler additiv durch normalverteilte Zufallsvariablen mit Erwartungswert null und gleicher, aber unbekannter Varianz. Für die Rekonstruktion des Signals u verwenden wir wieder den Ansatz

$$\hat{u} : \mathbb{R} \to \mathbb{R}, \quad t \mapsto \frac{a_0}{2} + \sum_{j=1}^{8} \left(a_j \cos\left(\frac{2\pi j}{3} t\right) + b_j \sin\left(\frac{2\pi j}{3} t\right) \right).$$

Da wir nun zur Berechnung der Koeffizienten die Varianz des Rauschens zugrunde legen, können wir durch geeignete Tests entscheiden, welche Koeffizienten gleich null sind. Geht man vom empfangenen Signal Abb. 0.4 aus, so können die Koeffizienten

$$b_1, \ldots, b_8 \quad \textit{und} \quad a_6, a_7, a_8$$

gleich null gesetzt werden. Es sind also nur die Koeffizienten $a_0, \ldots, a_5$ durch das entsprechende lineare Ausgleichsproblem zu berechnen. Das unbekannte Signal u wird also in einem sechsdimensionalen Unterraum rekonstruiert. Das so rekonstruierte Signal $\hat{u}$ ist in Abbildung 0.6 dargestellt.

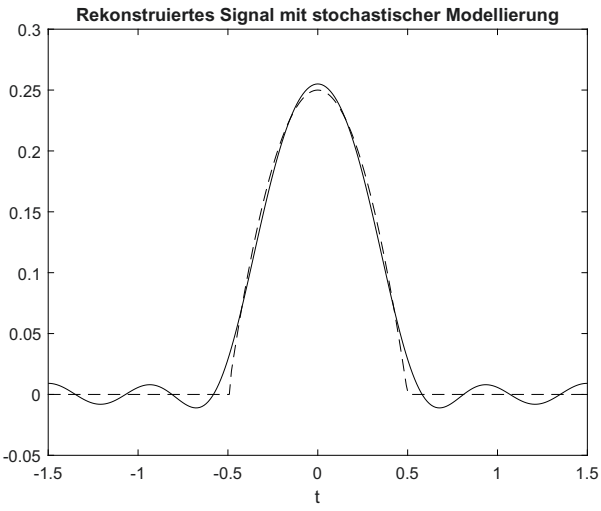

Abb. 0.6 *Rekonstruiertes Signal bei stochastischer Modellierung*

Dieses Beispiel zeigt den enormen Nutzen einer stochastischen Modellierung von Messdaten.

Um ein inverses Problem bearbeiten zu können, sind im Wesentlichen drei Schritte erforderlich. Zunächst muss das direkte Problem, also die Frage, welche Wirkung eine gegebene Ursache nach sich zieht, modelliert werden. Dies geschieht mit Methoden der **Funktionalanalysis** durch die Anwendung entsprechender problemrelevanter Gesetzmäßigkeiten (z. B. Naturgesetze, wirtschaftswissenschaftliche Axiome). Da die beobachtete Wirkung im Allgemeinen nur partiell durch Messungen zugänglich ist und da Messungen stets fehlerbehaftet sind, besteht der zweite Schritt in einer **stochastischen** Modellierung der gegebenen Messungen. Dieser Schritt bildet einen wichtigen Schwerpunkt des vorliegenden Buches, da die stochastische Modellierung von Messdaten im Rahmen der inversen Probleme und die daraus resultierenden Konsequenzen für die Rekonstruktion der Ursache bis jetzt weitestgehend vernachlässigt wird. Der dritte Schritt besteht schließlich darin, geeignete Verfahren der **numerischen Mathematik** zur computergestützten Approximation der gesuchten Ursache zu entwickeln und anzuwenden.

Da einerseits die zu behandelnde Problemstellung äußerst komplex ist und da andererseits das vorliegende Buch auch und gerade für Anwender hilfreich sein soll, ist ein großes einführendes Kapitel über Grundlagen zur linearen Algebra, zur Funktionalanalysis, zur Numerik und zur Stochastik unabdingbar. Das zweite Kapitel ist

der Analyse inverser Probleme gewidmet. Während man die Charakterisierung inverser Probleme, Fragen der Diskretisierung und Fragen der Regularisierung in der ein oder anderen Form auch in anderen Lehrbüchern zu diesem Thema finden wird, werden diese Fragen nun unter dem Gesichtspunkt stochastisch modellierter Messungen untersucht. Das dritte Kapitel dokumentiert in ausgewählten anwendungsrelevanten Beispielen die Notwendigkeit, verschiedenste Methoden der numerischen Mathematik problemadäquat anwenden zu können.

Inhaltsverzeichnis

Kapitel 1
Grundlagen

1.1 Lineare Algebra

Grundkenntnisse der linearen Algebra werden vorausgesetzt. Insbesondere wird vorausgesetzt, dass das Konzept des Vektorraums (linearen Raums) über dem Körper der reellen oder der komplexen Zahlen bekannt ist, ebenso die Begriffe Linearkombination, lineare Abhängigkeit und Unabhängigkeit, Untervektorraum (Teilraum), Basis. Bekannt sein sollten auch die Begriffe der Abbildung und ihrer Injektivität, Surjektivität, Bijektivität und Inversen (Umkehrabbildung) sowie insbesondere der linearen Abbildung, ihres Bildraums und Nullraums (Kerns). Die folgende Zusammenstellung von Begriffen und Resultaten dient der Festlegung von Schreibweisen und dem Überblick über die im Weiteren benötigten Ergebnisse.

Wir benutzen die Bezeichnung $\mathbb{K}$ stellvertretend sowohl für den Körper $\mathbb{R}$ der reellen Zahlen als auch für den Körper $\mathbb{C}$ der komplexen Zahlen, wenn wir uns nicht genauer festlegen wollen.

Der n-dimensionale Euklidische Raum

Für jede natürliche Zahl $n \in \mathbb{N}$ bezeichnen wir mit $\mathbb{R}^n$ den **n-dimensionalen, reellen Euklidischen Raum** und mit $\mathbb{C}^n$ den **n-dimensionalen, komplexen Euklidischen Raum**, $\mathbb{K}^n$ steht stellvertretend für beide. Die Elemente dieser Räume notieren wir wie folgt:

$$x \in \mathbb{K}^n \quad \Longleftrightarrow \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{alle } x_i \in \mathbb{K}.$$

Wir schreiben auch $x = (x_1, \ldots, x_n)^\top$. Der Hochindex $\top$ bedeutet **transponiert** und macht eine Spalte aus einer Zeile und umgekehrt. Die Elemente $x \in \mathbb{R}^n$ (oder $\mathbb{C}^n$)

M. Richter und S. Schäffler, *Inverse Probleme mit stochastisch modellierten Messdaten*, https://doi.org/10.1007/978-3-662-66343-1_1

nennen wir **Spaltenvektoren** oder kurz **Vektoren**. Ihre Addition und Skalarmultiplikation ist komponentenweise definiert. Der von Vektoren $x^1,\dots,x^k \in \mathbb{K}^n$ aufgespannte Untervektorraum wird mit

$$\langle x^1,\dots,x^k\rangle \quad \text{oder} \quad \operatorname{span}\{x^1,\dots,x^k\}$$

bezeichnet.

Jeder n-dimensionale Vektorraum V über dem Körper $\mathbb{K}$ ist isomorph zu $\mathbb{K}^n$, das heißt es gibt eine lineare, bijektive Abbildung $\phi : V \to \mathbb{K}^n$ (einen **Isomorphismus**). In diesem Sinn kann V mit $\mathbb{K}^n$ identifiziert werden.

Eine **Matrix** A ist ein rechteckiges Schema reeller oder komplexer Zahlen. Hat sie m Zeilen und n Spalten ($m,n \in \mathbb{N}$), dann schreibt man $A \in \mathbb{K}^{m,n}$ und definiert A durch ihre Komponenten

$$A \in \mathbb{K}^{m,n} \quad \Longleftrightarrow \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \text{alle } a_{ij} \in \mathbb{K}$$

oder gleichermaßen durch ihre Spalten

$$A \in \mathbb{K}^{m,n} \quad \Longleftrightarrow \quad A = \left(\begin{array}{c|c|c|c} & & & \\ a^1 & a^2 & \cdots & a^n \\ & & & \end{array}\right), \quad \text{alle } a^j \in \mathbb{K}^m.$$

Im Sonderfall $m = n$ nennt man eine Matrix **quadratisch** und n heißt ihre **Ordnung**. Die Dimension des **Spaltenraums** $\langle a^1,\dots,a^n\rangle$ heißt **Rang** der Matrix und wird mit $\operatorname{Rg}(A)$ bezeichnet. Die Regeln der Matrizenrechnung (Addition, Skalarmultiplikation und Multiplikation) werden als bekannt vorausgesetzt ebenso wie die Tatsache, dass Matrizen der Beschreibung linearer Abbildungen $f : \mathbb{K}^n \to \mathbb{K}^m$ dienen. Wir identifizieren eine Matrix $A \in \mathbb{K}^{m,n}$ mit der linearen Abbildung $f : \mathbb{K}^m \to \mathbb{K}^n$, $x \mapsto Ax$, und benutzen dann auch die gleiche Bezeichnung für beide: $A = f$. Spezielle Matrizen sind die **Nullmatrix**, deren sämtliche Komponenten null sind und die (für jedes $m,n \in \mathbb{N}$) mit 0 bezeichnet wird. Eine quadratische Matrix heißt **Diagonalmatrix**, wenn alle Außerdiagonalelemente gleich null sind. Eine $n \times n$-Diagonalmatrix, auf deren Diagonale nur Einsen stehen, heißt **Einheitsmatrix** und wird mit I_n bezeichnet, ihre Spalten

$$e^1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, e^n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

heißen **kanonische Einheitsvektoren**. Eine Matrix $T \in \mathbb{K}^{n,n}$ heißt **tridiagonal**, wenn $t_{ij} = 0$ für $|i-j| > 1$. Eine Matrix $L \in \mathbb{K}^{n,n}$ heißt **untere Dreiecksmatrix**, wenn

$$l_{ij} = 0 \quad \text{für} \quad i < j,$$

das heißt wenn oberhalb der Diagonale nur Nullen stehen. Gilt zudem $l_{ii} = 1$ für $i = 1,\ldots,n$, dann heißt L **normierte untere Dreiecksmatrix**. Analog nennt man $R \in \mathbb{K}^{n,n}$ eine **(normierte) obere Dreiecksmatrix**, wenn alle Elemente unterhalb der Diagonalen gleich null sind (und auf der Diagonale nur Einsen stehen).

Die **Determinante** einer quadratischen Matrix A der Ordnung n wird mit $\det(A)$ bezeichnet und durch

$$\det(A) := \sum_P (-)^P a_{1P(1)} \cdots a_{nP(n)}$$

definiert. Die Summe geht über alle $n!$ Permutationen P der Zahlen $1,\ldots,n$ – dies sind gerade die bijektiven Abbildungen der Menge $\{1,\ldots,n\}$ auf sich selbst. Der Ausdruck $(-)^P$ ist ± 1, je nachdem, ob die Permutation durch eine gerade oder eine ungerade Anzahl von paarweisen Vertauschungen der Zahlen $1,\ldots,n$ zustande kommt. Es gilt der **Determinantenmultiplikationssatz**

$$\det(AB) = \det(A) \cdot \det(B) \quad \text{für alle} \quad A, B \in \mathbb{K}^{n,n}.$$

Falls $A = L$ eine untere Dreiecksmatrix ist, dann ist $\det(L) = l_{11} \cdot \ldots \cdot l_{nn}$ und ebenso ist $\det(R) = r_{11} \cdot \ldots \cdot r_{nn}$ für eine obere Dreeicksmatrix R. Falls $\det(A) \neq 0$, heißt die Matrix $A \in \mathbb{K}^{n,n}$ **invertierbar** oder **nichtsingulär**. Es gibt dann eine Matrix $B \in \mathbb{K}^{n,n}$ mit $AB = BA = I_n$. B nennt man dann die **Inverse** von A und schreibt $B =: A^{-1}$.

Für $A \in \mathbb{K}^{m,n}$ ist die **transponierte Matrix** $A^\top \in \mathbb{K}^{n,m}$ durch Umstellen der Zeilen zu Spalten gegeben, also durch

$$(A^\top)_{ij} := a_{ji}, \quad i = 1,\ldots,n \text{ und } j = 1,\ldots,m.$$

Werden die Komponenten von A zusätzlich konjugiert, so erhält man die **adjungierte Matrix** oder **hermitisch konjugierte Matrix** $A^* \in \mathbb{K}^{n,m}$ mit Komponenten

$$(A^*)_{ij} := \overline{a_{ji}}, \quad i = 1,\ldots,n \text{ und } j = 1,\ldots,m,$$

wobei $\bar{z}$ die zu $z \in \mathbb{C}$ konjugiert komplexe Zahl ist. (Für reelle Matrizen stimmt die adjungierte mit der transponierten Matrix überein.) Es gelten die Rechenregeln $\det(A^\top) = \det(A)$, $(AB)^\top = B^\top A^\top$ und $(AB)^* = B^* A^*$, sofern das Matrixprodukt AB definiert ist. Ist A invertiertbar, dann gilt $(A^{-1})^* = (A^*)^{-1} =: A^{-*}$. Eine Matrix mit der Eigenschaft $A = A^\top$ heißt **symmetrisch** und eine Matrix mit der Eigenschaft $A = A^*$ heißt **hermitisch** oder **selbstadjungiert**.

Orthogonalität

Für $x, y \in \mathbb{K}^n$ definieren wird das **Euklidische Skalarprodukt** durch

$$\langle x|y\rangle := \overline{x^* y} = \sum_{i=1}^{n} x_i \overline{y_i}.$$

Hier ist der Zeilenvektor $x^* = (\overline{x_1}, \ldots, \overline{x_n})$ adjungiert zum Spaltenvektor x und $x^* y$ ist ein Matrixprodukt. Im reellen Fall vereinfacht sich das Euklidische Skalarprodukt zu $\langle x|y\rangle = \sum_{i=1}^{n} x_i y_i$ für $x, y \in \mathbb{R}^n$. Vektoren $x, y \in \mathbb{K}^n$ heißen **orthogonal**, falls $\langle x|y\rangle = 0$. In diesem Fall schreiben wir $x \perp y$. Vektoren $b^1, \ldots, b^k \in \mathbb{K}^n$ heißen **orthonormal**, falls $\langle b^i|b^j\rangle = 0$ für $i \neq j$ und $\langle b^i|b^i\rangle = 1$ für alle i. Gilt zusätzlich $k = n$, dann heißt $\{b^1, \ldots, b^n\}$ eine **Orthonormalbasis (ONB)** des Euklidischen Raums $\mathbb{K}^n$. Jeder nicht nur aus dem Nullvektor bestehende Unterraum $U \subseteq \mathbb{K}^n$ besitzt eine Orthonormalbasis. Eine Matrix $V \in \mathbb{K}^{n,n}$ wird **unitär** und im Spezialfall $V \in \mathbb{R}^{n,n}$ auch **orthogonal** genannt, falls ihre Spalten eine ONB des Euklidischen Raums $\mathbb{K}^n$ sind. Dies ist äquivalent zu den Matrixidentitäten

$$V^* V = I_n \quad \text{bzw.} \quad V^{-1} = V^*.$$

Falls $A \in \mathbb{K}^{m,n}$ mit $m \geq n$ und $\mathrm{Rg}(A) = n$, dann sind alle Spalten $a^1, \ldots, a^n$ von A linear unabhängig und überdies haben die linearen Räume

$$\langle a^1, \ldots, a^i\rangle \subseteq \mathbb{K}^m, \quad i = 1, \ldots, n,$$

die Dimension i und die Basis $\{a^1, \ldots, a^i\}$. Es können (konstruktiv mit dem sogenannten **Orthonormalisierungsverfahren von Gram-Schmidt**) n orthonormale Vektoren $q^1, \ldots, q^n \in \mathbb{K}^m$ so gefunden werden, dass

$$\langle a^1, \ldots, a^i\rangle = \langle q^1, \ldots, q^i\rangle, \quad i = 1, \ldots, n,$$

insbesondere kann jeder Vektor a^i als Linearkombination

$$a^i = r_{1i} \cdot q^1 + \ldots + r_{ii} \cdot q^i, \quad i = 1, \ldots, n,$$

geschrieben werden. Diese n Vektorgleichungen lassen sich zu einer einzigen Matrixgleichung zusammenfassen:

$$\left(\begin{array}{c|c|c|c} a^1 & a^2 & \cdots & a^n \end{array}\right) = \underbrace{\left(\begin{array}{c|c|c|c} q^1 & q^2 & \cdots & q^n \end{array}\right)}_{=: \hat{Q}} \underbrace{\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & & \\ & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix}}_{=: \hat{R}}, \tag{1.1}$$

wobei die Diagonalelemente r_{ii} ungleich null sind und $\hat{Q}$ orthonormale Spalten hat. Die Identität (1.1) heißt **reduzierte QR-Zerlegung** der Matrix A. Stets kann man $q^1,\ldots,q^n$ so um $m-n$ Vektoren $q^{n+1},\ldots,q^m \in \mathbb{K}^m$ ergänzen, dass eine Orthonormalbasis des $\mathbb{K}^m$ entsteht (die $q^{n+1},\ldots,q^m$ dürfen ansonsten beliebig sein). Die Vektoren $q^1,\ldots,q^m$ sind Spalten der unitären Matrix $Q = (q^1|\cdots|q^n|q^{n+1}|\cdots|q^m) \in \mathbb{K}^{m,m}$. Ergänzt man weiterhin $\hat{R} \in \mathbb{K}^{n,n}$ um $m-n$ Nullzeilen zu einer Matrix $R \in \mathbb{K}^{m,n}$, dann erhält man die Faktorisierung

$$A = QR,$$

die sogenannte **QR-Zerlegung** von A.

Eigenwerte

Eine Matrix $A \in \mathbb{K}^{n,n}$ hat einen **Eigenwert** $\lambda \in \mathbb{C}$ und zugehörigen **Eigenvektor** $v \in \mathbb{C}^n$, falls

$$Av = \lambda v \quad \text{und} \quad v \neq 0. \tag{1.2}$$

Die Gleichung $Av = \lambda v$ besitzt genau dann eine Lösung $v \neq 0$, wenn die Matrix $A - \lambda I_n$ linear abhängige Spalten hat, also singulär ist. Ohne Kenntnis der Eigenvektoren lassen sich deswegen die Eigenwerte bestimmen als diejenigen komplexen Zahlen λ, für welche $\chi_A(\lambda) := \det(A - \lambda I_n) = 0$ gilt. Die (auf ganz $\mathbb{C}$ definierte) Funktion χ_A nennt man das **charakteristische Polynom** von A. Dieses hat den Grad n und deswegen genau n Nullstellen in $\mathbb{C}$. Somit besitzt jede quadratische Matrix A der Ordnung n genau n Eigenwerte. Dabei gelten Mehrfachnullstellen von χ_A als Mehrfacheigenwerte von A. Die zugehörigen Eigenvektoren sind – mit Ausnahme des Nullvektors – die Elemente des Lösungsraums des linearen Gleichungssystems $(A - \lambda I_n)x = 0$, den wir mit $\mathcal{N}_{A-\lambda I_n}$ bezeichnen[1]. Aus dem Gesagten ergibt sich, dass Eigenwerte und Eigenvektoren im Allgemeinen komplexwertig sein können selbst dann, wenn A eine reelle Matrix ist. Ferner kann für einen k-fachen Eigenwert λ die Dimension des Vektorraums $\mathcal{N}_{A-\lambda I_n}$ kleiner als k sein, so dass eine quadratische Matrix der Ordnung n weniger als n linear unabhängige Eigenvektoren besitzen kann. Ist jedoch $A \in \mathbb{K}^{n,n}$ eine selbstadjungierte Matrix, dann lässt sich zeigen, dass alle Eigenwerte reell sind. Außerdem existiert in diesem Fall eine Orthonormalbasis $\{v^1,\ldots,v^n\} \subset \mathbb{C}^n$ aus Eigenvektoren. Somit gilt

$$Av^i = \lambda_i v^i,\ i = 1,\ldots,n \quad \Longleftrightarrow \quad AV = V\Lambda \quad \Longleftrightarrow \quad V^*AV = \Lambda, \tag{1.3}$$

wobei $V = (v^1|\cdots|v^n) \in \mathbb{C}^{n,n}$ (Spalten sind Eigenvektoren) und $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ (Diagonalelemente sind Eigenwerte). Falls A reellwertig ist, dann ist auch V reellwertig. Die Matrixidentität (1.3) nennt man **unitäre Diagonalisierung** der Matrix A. Eine quadratische Matrix A lässt sich genau dann unitär diagonalisieren, wenn

[1] Es handelt sich um den Nullraum der Matrix $A - \lambda I_n$

$AA^* = A^*A$ gilt. Matrizen mit dieser Eigenschaft nennt man **normal**.

Eine Matrix $A \in \mathbb{K}^{n,n}$ heißt **positiv definit**, wenn sie selbstadjungiert ist *und* wenn $x^*Ax > 0$ für alle $x \in \mathbb{K}^n \setminus \{0\}$ gilt. Sie heißt **positiv semidefinit**, wenn sie selbstadjungiert ist *und* wenn $x^*Ax \geq 0$ für alle $x \in \mathbb{K}^n$ gilt. Eine Matrix $A \in \mathbb{K}^{n,n}$ ist genau dann positiv definit, wenn sie selbstadjungiert ist und alle ihre Eigenwerte positiv sind. Sie ist genau dann positiv semidefinit, wenn sie selbstadjungiert ist und keinen negativen Eigenwert hat. Die Anzahl positiver Eigenwerte einer positiv semidefiniten Matrix ist gleich dem Rang der Matrix. Weiterhin ist A genau dann positiv definit, wenn es eine invertierbare obere Dreiecksmatrix $R \in \mathbb{K}^{n,n}$ so gibt, dass

$$A = R^*R. \tag{1.4}$$

Dies nennt man die **Cholesky-Zerlegung** von A. Die Matrix R kann reellwertig gewählt werden, falls A reellwertig ist.

Zwei quadratische Matrizen $A, B \in \mathbb{K}^{n,n}$ besitzen den **verallgemeinerten Eigenwert** $\lambda \in \mathbb{C}$ und dazu den **verallgemeinerten Eigenvektor** $v \in \mathbb{C}^n$, falls

$$Av = \lambda Bv \quad \text{und} \quad v \neq 0.$$

Sei nun insbesondere A positiv semidefinit und B positiv definit. Unter Benutzung der Cholesky-Zerlegung $B = R^*R$ und der Transformation $Rv = w$ kann das Problem der Bestimmung verallgemeinerter Eigenwerte und Eigenvektoren in ein äquivalentes gewöhnliches Eigenwertproblem umformuliert werden:

$$R^{-*}AR^{-1}w = \lambda w, \quad w \neq 0.$$

Hier ist $R^{-*}AR^{-1}$ eine positiv semidefinite Matrix und folglich existiert eine ONB $\{w^1, \ldots, w^n\}$ von Eigenvektoren zu Eigenwerten $\lambda_1, \ldots, \lambda_n \geq 0$ von $R^{-*}AR^{-1}$. Mit der orthogonalen Matrix $W := (w^1 | \cdots | w^n)$ und der nicht-singulären Matrix $V := R^{-1}W$ erhält man

$$V^*BV = W^*R^{-*}R^*RR^{-1}W = W^*W = I_n$$

und außerdem

$$V^*AV = W^*(R^{-*}AR^{-1})W = W^*W\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n) = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),$$

wobei $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ die Diagonalmatrix der Ordnung n mit Diagonalelementen $\lambda_1, \ldots, \lambda_n$ ist. Zusammenfassend ergibt sich: Falls $A \in \mathbb{K}^{n,n}$ positiv semidefinit und $B \in \mathbb{K}^{n,n}$ is positiv definit ist, dann gibt es eine nicht-singuläre Matrix $V \in \mathbb{K}^{n,n}$ so, dass

$$V^*AV = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),\ \lambda_1, \ldots, \lambda_n \geq 0, \quad \text{und} \quad V^*BV = I_n. \tag{1.5}$$

Normen für Vektoren und Matrizen

Für Vektoren $x \in \mathbb{K}^n$ definiert man die **Summennorm**

$$\|x\|_1 := \sum_{j=1}^{n} |x_j|,$$

die **Euklidische Norm**

$$\|x\|_2 := \sqrt{\sum_{j=1}^{n} |x_j|^2}$$

und die **Maximumsnorm**

$$\|x\|_\infty := \max\left\{|x_j|;\ j = 1,\ldots,n\right\}.$$

Allgemein nennt man **Norm** auf $\mathbb{K}^n$ eine Abbildung $\|\bullet\| : \mathbb{K}^n \to [0,\infty)$, welche die drei Eigenschaften

(1) **Definitheit**, das heißt $x \neq 0 \Longrightarrow \|x\| > 0$,
(2) **Homogenität**, das heißt $\|\lambda x\| = |\lambda|\|x\|$ für alle $\lambda \in \mathbb{K}$ und
(3) **Subadditivität**, das heißt $\|x+y\| \leq \|x\| + \|y\|$ für alle $x, y \in \mathbb{K}^n$

besitzt. Die Schreibweise $\|\bullet\|$ bedeutet, dass es sich um eine Funktion handelt, in welche an Stelle des Symbols $\bullet$ das Argument einzusetzen ist: $\|x\| = \|\bullet\|(x)$. Die bei der Subadditivität angegebene Ungleichung heißt **Dreiecksungleichung**. Die **Cauchy-Schwarzsche Ungleichung**

$$|\langle x|y\rangle| \leq \|x\|_2 \|y\|_2$$

gilt für alle $x, y \in \mathbb{K}^n$. In der Ungleichung von Cauchy-Schwarz gilt Gleichheit genau dann, wenn x und y linear abhängig sind, wenn also $x = 0$ oder $y = \lambda x$ mit einem $\lambda \in \mathbb{K}$. Der **Satz des Pythagoras** besagt

$$\|b^1 + \ldots + b^k\|_2^2 = \|b^1\|_2^2 + \ldots + \|b^k\|_2^2,$$

wenn $b^1, \ldots, b^k \in \mathbb{K}^n$ paarweise orthogonale Vektoren sind.

Alle möglichen Normen auf $\mathbb{K}^n$ sind **äquivalent**, das heißt zu jeder Norm $\|\bullet\|$ auf $\mathbb{K}^n$ gibt es zwei positive Konstanten α und β so, dass

$$\alpha\|x\|_\infty \leq \|x\| \leq \beta\|x\|_\infty \quad \text{für alle} \quad x \in \mathbb{K}^n. \tag{1.6}$$

Eine Möglichkeit, Normen für Matrizen einzuführen besteht darin, die Matrix-Elemente als Komponenten eines Vektors aufzufassen. Entsprechend der Euklidischen Norm von Vektoren erhält man so die **Frobenius-Norm**:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}, \quad A \in \mathbb{K}^{m,n}.$$

Passender zur Auffassung von Matrizen als lineare Abbildungen ist jedoch das folgende Konzept. Die **Operatornorm** $\|\bullet\| : \mathbb{K}^{m,n} \to [0,\infty)$ einer Matrix $A \in \mathbb{K}^{m,n}$ wird definiert durch

$$\|A\| := \max\left\{ \frac{\|Ax\|}{\|x\|};\, x \in \mathbb{K}^n \setminus \{0\} \right\} = \max\{\|Ax\|;\, x \in \mathbb{K}^n,\ \|x\| = 1\}$$

wobei für $\|Ax\|$ und $\|x\|$ Vektornormen auf $\mathbb{K}^m$ beziehungsweise $\mathbb{K}^n$ benutzt werden – dies könnten grundsätzlich unterschiedliche Normen sein. Verwendet man jeweils die gleiche Vektornorm $\|\bullet\|_p$, $p \in \{1,2,\infty\}$, dann bezeichnet man die entsprechende Operatornorm ebenfalls mit $\|\bullet\|_p$. In diesen Fällen gelten folgende Berechnungsformeln:

$$\begin{aligned}
\|\bullet\|_1 \rightsquigarrow \|A\|_1 &= \max_j \sum_i |a_{ij}|, \text{ größte Spaltenbetragssumme,} \\
\|\bullet\|_\infty \rightsquigarrow \|A\|_\infty &= \max_i \sum_j |a_{ij}|, \text{ größte Zeilenbetragssumme,} \\
\|\bullet\|_2 \rightsquigarrow \|A\|_2 &= \sqrt{\lambda_1}, \\
&\quad \text{wobei } \lambda_1 \geq \lambda_2 \geq \ldots \geq 0 \text{ die absteigend geordneten} \\
&\quad \text{Eigenwerte der positiv semidefiniten Matrix } A^\top A \\
&\quad \text{sind. } \|\cdot\|_2 \text{ heißt } \textbf{Spektralnorm}.
\end{aligned}$$

Operatornormen zeichnen sich durch die folgenden fünf Eigenschaften aus, von denen die ersten drei gelten müssen, damit die Bezeichnung als Norm überhaupt gerechtfertigt ist:

Definitheit:	$A \neq 0 \quad \Longrightarrow \quad \|A\| > 0,$
Homogenität:	$\|\lambda A\| = \|\lambda\|\|A\|, \quad \forall \lambda \in \mathbb{R},$
Subadditivität:	$\|A+B\| \leq \|A\| + \|B\|,$
Submultiplikativität:	$\|AB\| \leq \|A\| \cdot \|B\|$ und
Konsistenz:	$\|Ax\| \leq \|A\| \cdot \|x\|.$

Die Singulärwertzerlegung (SVD)

Es sei $m \geq n$ und $A \in \mathbb{K}^{m,n}$ mit $\mathrm{Rg}(A) = r$. Dann ist die Matrix $A^*A \in \mathbb{K}^{n,n}$ positiv semidefinit und hat ebenfalls Rang r (wie auch die Matrizen A^* und AA^*). Ihre Eigenwerte seien $\sigma_1^2 \geq \ldots \geq \sigma_r^2 > 0$ und $\sigma_{r+1}^2 = \ldots = \sigma_n^2 = 0$ mit zugehörigen orthonormalen Eigenvektoren $v_1, \ldots, v_n \in \mathbb{K}^n$, so dass

$$A^*Av_k = \sigma_k^2 v_k, \quad k = 1, \ldots, n,$$

entsprechend (1.3). Dann sind $u_k := Av_k/\sigma_k \in \mathbb{K}^m$, $k = 1,\dots,r$, Eigenvektoren von AA^*, da $AA^*u_k = AA^*Av_k/\sigma_k = A\sigma_k v_k = \sigma_k^2 u_k$. Die Vektoren u_k sind ebenfalls orthonormal:

$$u_i^* u_k = v_i^* A^* A v_k/(\sigma_i\sigma_k) = v_i^* v_k \sigma_k/\sigma_i = \delta_{i,k}.$$

Hier wurde das sogenannte **Kronecker-Symbol** benutzt, welches durch $\delta_{i,k} := 0$ für $i \neq k$ und $\delta_{i,i} := 1$ definiert ist. Die Menge $\{u_1,\dots,u_r\}$ wird durch $m-r$ orthonormale Vektoren $u_{r+1},\dots,u_m \in \mathbb{K}^m$ ergänzt, welche den $(m-r)$-dimensionalen[2] Nullraum $\mathcal{N}_{A^*}$ aufspannen:

$$A^* u_k = 0, \quad k = r+1,\dots,m,$$

und welche die verbleibenden Eigenvektoren von AA^* sind. Für $i \leq r < k$ erhalten wir $u_i^* u_k = v_i^* A^* u_k/\sigma_i = v_i^* 0/\sigma_i = 0$, so dass $U := (u_1|\cdots|u_m) \in \mathbb{K}^{m,m}$ ebenso eine unitäre Matrix ist wie $V := (v_1|\cdots|v_n) \in \mathbb{K}^{n,n}$. Aus den Definitionen von u_k und v_k bekommen wir $Av_k = \sigma_k u_k$ für $k = 1,\dots,r$ und $Av_k = 0$ für $k = r+1,\dots,n$. Zusammen ergibt das

$$AV = U\Sigma \quad \Longleftrightarrow \quad A = U\Sigma V^* \quad \text{mit} \quad \Sigma_{i,j} = \sigma_i \delta_{i,j}. \tag{1.7}$$

Lässt man die letzten $m-n$ Spalten der Matrix U weg und die letzten $m-n$ Zeilen von Σ, dann ergibt sich

$$A = \hat{U}\hat{\Sigma}V^*, \quad \hat{U} := (u_1|\cdots|u_n) \in \mathbb{C}^{m,n}, \quad \hat{\Sigma} = \operatorname{diag}(\sigma_1,\dots,\sigma_n) \in \mathbb{R}^{n,n} \tag{1.8}$$

anstelle von (1.7). Im Fall $m < n$ kann man eine Faktorisierung (1.7) von A^* wie oben berechnen. Danach geht man auf beiden Seiten von (1.7) beziehungsweise (1.8) zu den hermitisch konjugierten Matrizen über. Damit ist der folgende Satz gezeigt.

Satz und Definition 1.1 (Singulärwertzerlegung (SVD)). *Es habe $A \in \mathbb{K}^{m,n}$ den Rang r. Dann gibt es unitäre Matrizen $U \in \mathbb{K}^{m,m}$ und $V \in \mathbb{K}^{n,n}$ sowie eine Matrix $\Sigma \in \mathbb{R}^{m,n}$ mit Komponenten $\Sigma_{i,j} = \sigma_i \delta_{i,j}$ und*

$$\sigma_1 \geq \ldots \geq \sigma_r > 0, \quad \sigma_{r+1} = \ldots = \sigma_{\min\{m,n\}} = 0$$

derart dass

$$A = U\Sigma V^*.$$

*Diese Faktorisierung heißt **Singulärwertzerlegung** (singular value decomposition — SVD) und die Zahlen $\sigma_1 \geq \ldots \geq \sigma_{\min\{m,n\}} \geq 0$ heißen **Singulärwerte** von A. Im Detail ist*

[2] Für jede Matrix $A \in \mathbb{K}^{m,n}$ mit Nullraum $\mathcal{N}_A$ gilt $\dim(\mathcal{N}_A) + \operatorname{Rg}(A) = n$.

$$A = U\Sigma V^* = \underbrace{\begin{pmatrix} U_1 & U_2 \end{pmatrix}}_{\substack{r \quad m-r}} \left.\begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}\right\} \begin{matrix} r \\ m\text{-}r \end{matrix} \underbrace{\begin{pmatrix} V_1 & V_2 \end{pmatrix}^*}_{\substack{r \quad n-r}} \tag{1.9}$$
$$\underbrace{\qquad\qquad}_{\substack{r \quad n-r}}$$
$$= U_1\Sigma_1 V_1^*$$

*und die Faktorisierung $A = U_1\Sigma_1 {V_1}^*$ wird **reduzierte SVD** genannt.* ◁

Aus der reduzierten SVD leitet sich unmittelbar folgende Interpretation der Matrizen U und V ab:

- Die Spalten von U_1 sind eine ONB des Spaltenraums $\mathcal{R}_A$ der Matrix A. Die Spalten von U_2 sind eine ONB des orthogonalen Komplements $\mathcal{R}_A^\perp = \{y \in \mathbb{K}^m;\ \langle y|z\rangle = 0 \text{ für alle } z \in \mathcal{R}_A\}$ von $\mathcal{R}_A$ in $\mathbb{K}^m$.
- Die Spalten von V_2 sind eine ONB des Kerns $\mathcal{N}_A$ der Matrix A. Die Spalten von V_1 sind eine ONB des orthogonalen Komplements $\mathcal{N}_A^\perp = \{x \in \mathbb{K}^n;\ \langle x|z\rangle = 0 \text{ für alle } z \in \mathcal{N}_A\}$ von $\mathcal{N}_A$ in $\mathbb{K}^n$.

Weiterhin lässt sich feststellen, dass die Matrix $V_1 {V_1}^*$ (interpretiert als Abbildung von $\mathbb{K}^n$ nach $\mathbb{K}^n$) alle Spalten von V_1 auf sich selbst und alle Spalten von V_2 auf 0 abbildet, diese Matrix ist deswegen gleich dem orthogonalen Projektor $P_{\mathcal{N}_A^\perp}$ von $\mathbb{K}^n$ auf $\mathcal{N}_A^\perp$. Analog ist $U_1 {U_1}^*$ der orthogonale Projektor $P_{\mathcal{R}_A}$ von $\mathbb{K}^m$ auf $\mathcal{R}_A$.

Definition 1.2 (Pseudoinverse Matrix). *$A \in \mathbb{K}^{m,n}$ habe den Rang r und die reduzierte SVD $A = U_1\Sigma_1 V_1^*$ wie in Satz und Definition 1.1. Dann wird*

$$A^+ := V_1 {\Sigma_1}^{-1} U_1^* \in \mathbb{K}^{n,m}$$

die zu A pseudoinverse Matrix genannt. ◁

Der Name „Pseudoinverse“ erklärt sich folgendermaßen. Die durch die Matrix A repräsentierte lineare Abbildung $A : \mathcal{N}_A^\perp \to \mathcal{R}_A$, $x \mapsto Ax$, ist bijektiv mit inverser Abbildung $B : \mathcal{R}_A \to \mathcal{N}_A^\perp$. Dann kann $B \circ P_{\mathcal{R}_A} : \mathbb{K}^m \to \mathbb{K}^n$ als eine zu $A : \mathbb{K}^n \to \mathbb{K}^m$ pseudoinverse Abbildung aufgefasst werden und es gilt

$$A^+ = B \circ P_{\mathcal{R}_A}. \tag{1.10}$$

Um diese Identität einzusehen, schreiben wir ein beliebiges $y \in \mathbb{K}^m$ in der Form $y = U_1 z_1 + U_2 z_2$ mit $z_1 \in \mathbb{K}^r$ und $z_2 \in \mathbb{K}^{m-r}$. Damit erhält man einerseits $A^+ y = V_1 {\Sigma_1}^{-1} z_1$. Andererseits erhält man

$$B \circ P_{\mathcal{R}_A} y = BU_1 z_1 = V_1 {\Sigma_1}^{-1} z_1,$$

denn $V_1 {\Sigma_1}^{-1} z_1 \in \mathcal{N}_A^\perp$ und $AV_1 {\Sigma_1}^{-1} z_1 = U_1\Sigma_1 {V_1}^* V_1 {\Sigma_1}^{-1} z_1 = U_1 z_1$. Folglich ist $A^+ y = B \circ P_{\mathcal{R}_A} y$ für alle $y \in \mathbb{K}^m$, also stimmt (1.10).

Unter Benutzung der SVD ist leicht nachzuweisen, dass

$$A \text{ invertierbar} \quad \Longrightarrow \quad A^+ = A^{-1} \text{ und}$$
$$\operatorname{Rg}(A) = n \quad \Longrightarrow \quad A^+ = (A^*A)^{-1}A^*$$

gilt.

Mittels SVD und pseudoinverser Matrix lässt sich eine Lösung des **linearen Ausgleichsproblems** angeben. Dieses lautet für $A \in \mathbb{K}^{m,n}$ und $b \in \mathbb{K}^m$:

$$\text{Finde } \hat{x} \text{ so, dass } \|A\hat{x} - b\|_2 \leq \|Ax - b\|_2 \text{ für alle } x \in \mathbb{K}^n. \tag{1.11}$$

Bezeichnet man mit

$$M := \arg\min\{\|Ax - b\|_2\}$$

die Menge der Lösungen von (1.11) dann heißt

$$\hat{x} \in M \text{ mit } \|\hat{x}\|_2 \leq \|x\|_2 \text{ für alle } x \in M \tag{1.12}$$

eine **Minimum-Norm-Lösung** des linearen Ausgleichsproblems (1.11). Falls (1.11) eine eindeutige Lösung besitzt, dann ist diese trivialerweise auch die Minimum-Norm-Lösung. Falls die Lösung von (1.11) nicht eindeutig ist, dann ist die Minimum-Norm-Lösung dennoch eindeutig:

Satz 1.3 (Lösung des linearen Ausgleichsproblems). *$A \in \mathbb{K}^{m,n}$ habe den Rang r und die reduzierte SVD (1.9).*

(a) Jede Lösung x von (1.11) hat die Form

$$x = V_1 {\Sigma_1}^{-1} U_1^* b + V_2 z, \quad z \in \mathbb{K}^{n-r}. \tag{1.13}$$

(b) Stets existiert eine eindeutige Lösung von (1.12), also eine eindeutige Minimum-Norm-Lösung. Diese ist gegeben durch

$$x = V_1 {\Sigma_1}^{-1} U_1^* b = A^+ b.$$

Ihre Norm ist durch $\|x\|_2 \leq \|b\|_2/\sigma_r$ beschränkt.

(c) Ersetzt man b durch einen Vektor $b + \delta b \in \mathbb{K}^m$, dann erhält man eine eindeutige Minimum-Norm-Lösung $x + \delta x$. Die Differenz der beiden Lösungen ist durch

$$\|\delta x\|_2 \leq \frac{\|\delta b\|_2}{\sigma_r}$$

beschränkt.

Insbesondere ist die Lösung des linearen Ausgleichsproblems (1.11) genau dann eindeutig bestimmt, wenn A den Rang n hat und in diesem Fall durch

$$\hat{x} = A^+ b = (A^*A)^{-1}A^*b$$

gegeben – entsprechend der Lösung der Normalengleichungen des Ausgleichsproblems. ◁

Beweis. Teil (a):

$$\begin{aligned}\|b-Ax\|_2^2 &= \left\| U^*b - \begin{pmatrix} U_1^* \\ U_2^* \end{pmatrix} U_1\Sigma_1 V_1^* x \right\|_2^2 = \left\| \begin{pmatrix} U_1^*b - \Sigma_1 V_1^* x \\ U_2^* b \end{pmatrix} \right\|_2^2 \\ &= \|U_1^*b - \Sigma_1 V_1^* x\|_2^2 + \|U_2^* b\|_2^2\end{aligned}$$

wird genau dann minimal, wenn

$$\Sigma_1 V_1^* x = U_1^* b \quad \Longleftrightarrow \quad x = V_1 \Sigma_1^{-1} U_1^* b + V_2 z$$

für beliebiges $z \in \mathbb{R}^{n-r}$, denn

$$\mathcal{N}_{V_1^*} = \mathcal{R}_{V_1}^{\perp} = \{V_2 z;\, z \in \mathbb{K}^{n-r}\}.$$

Teil (b): Da die Spalten von V_1 und V_2 orthogonal sind, erhält man aus (1.13) und dem Satz des Pythagoras

$$\|x\|_2^2 = \|V_1 \Sigma_1^{-1} U_1^* b\|_2^2 + \|V_2 z\|_2^2.$$

Die rechte Seite wird genau dann minimal, wenn $V_2 z = 0$, das heißt genau dann, wenn $z = 0$. Für $z = 0$ erhalten wir

$$\|x\|_2^2 = \|V_1 \Sigma_1^{-1} U_1^* b\|_2^2 = \left\| \begin{pmatrix} u_1^* b/\sigma_1 \\ \vdots \\ u_r^* b/\sigma_r \end{pmatrix} \right\|_2^2 \le \frac{1}{\sigma_r^2} \sum_{j=1}^{r} |u_j^* b|^2 \le \frac{\|b\|_2^2}{\sigma_r^2}.$$

Teil (c): Ersetzt man in Teil (b) den Vektor b durch $b + \delta b$, dann ergibt sich die Minimum-Norm-Lösung $x + \delta x = V_1 \Sigma_1^{-1} U_1^* (b + \delta b)$. Für die Differenz der Lösungen erhält man $\delta x = V_1 {\Sigma_1}^{-1} U_1^* \delta b$ und die Normabschätzung folgt dann direkt aus Teil (b). □

Es gibt eine enge Verbindung zwischen den singulären Werten einer Matrix und ihrer Spektralnorm. Der nachfolgende Satz wird beispielsweise in Lecture 5 des Lehrbuchs [TB97] behandelt.

Satz 1.4. *Die Matrix $A \in \mathbb{K}^{m,n}$ habe Singulärwerte $\sigma_1 \ge \ldots \ge \sigma_{\min\{m,n\}} \ge 0$. Dann ist*

$$\|A\|_2 = \sigma_1.$$

Im Fall $m = n$ ist A genau dann invertierbar, wenn $\sigma_n > 0$. In diesem Fall ist

$$\|A^{-1}\|_2 = \frac{1}{\sigma_n}.$$

Es sei $\mathbb{M}_k$ die Menge aller Matrizen in $\mathbb{K}^{m,n}$, deren Rang kleiner als k ist (insbesondere enthält $\mathbb{M}_1$ nur die Nullmatrix). Dann gilt für $k = 1, \ldots, \min\{m,n\}$

$$\min\{\|A - X\|_2;\, X \in \mathbb{M}_k\} \quad = \quad \sigma_k. \tag{1.14}$$

A hat also den Abstand σ_k von den Matrizen vom Rang kleiner als k. ◁

Aus der Gleichung (1.14) lässt sich folgern: Im Fall $\sigma_{\min\{m,n\}} \leq \varepsilon$ liegt in einer Entfernung ε von A eine Matrix, deren Rang kleiner als $\min\{m,n\}$ ist. Da der Rang einer Matrix eine unstetige Funktion der Matrixkomponenten ist, kann er zumindest dann numerisch nicht zuverlässig berechnet werden, wenn er kleiner als $\min\{m,n\}$ ist und wenn die Komponenten der Matrix mit Unsicherheiten behaftet sind. Letzteres ist in der Praxis der Regelfall, siehe hierzu die Ausführungen in Abschnitt 1.4. Im Gegensatz dazu können singuläre Werte prinzipiell zuverlässig numerisch berechnet werden. Dies wird durch den nachfolgenden Satz 1.5 ausgesagt, der die „gute Kondition“ der singulären Werte einer Matrix formuliert – siehe den nachfolgenden Abschnitt 1.4. Die Berechnung des kleinsten singulären Wertes einer Matrix beantwortet deswegen am besten die Frage nach ihrem Rang.

Satz 1.5 (Sensitivität singulärer Werte). *Es sei $A, \delta A \in \mathbb{K}^{m,n}$. Es seien $\sigma_1 \geq \ldots \geq \sigma_{\min\{m,n\}} \geq 0$ die singulären Werte von A und es seien $\tilde{\sigma}_1 \geq \ldots \geq \tilde{\sigma}_{\min\{m,n\}} \geq 0$ die singulären Werte der Matrix $A + \delta A$. Dann gilt*

$$|\sigma_i - \tilde{\sigma}_i| \leq \|\delta A\|_2, \quad i = 1, \ldots, \min\{m,n\},$$

und diese obere Schranke ist scharf, das heißt es lässt sich jeweils ein $\delta A \in \mathbb{K}^{m,n}$ so finden, dass sie erreicht wird. ◁

Dieser Satz wird beispielsweise auf Seite 198 des Lehrbuchs [Dem97] bewiesen.

1.2 Funktionalanalysis

Inverse Probleme treten in Form von Gleichungen – etwa Integralgleichungen – auf, bei denen die gesuchte Lösung eine Funktion ist. Eine prägnante Beschreibung inverser Probleme und der mit ihrer Lösung verbundenen Schwierigkeiten erfordert deswegen die Betrachtung von Funktionenräumen, wie dies in der Funktionalanalysis geschieht. Der folgende Abschnitt ersetzt kein Lehrbuch der Funktionalanalysis, sondern gibt einen Überblick über die benötigten Definitionen und Sätze. Für eine Einführung in die Funktionalanalysis empfehlen wir das Lehrbuch [Wer10]. Wir stützen uns auch auf [Alt11], [Kir11] und [Ric20]. Auch in diesem Abschnitt benutzen wir das Symbol $\mathbb{K}$ wahlweise für $\mathbb{R}$ und $\mathbb{C}$, wenn wir uns nicht genauer festlegen wollen.

Vektorräume und Operatoren

Von besonderer Bedeutung für inverse Probleme sind Vektorräume, deren Elemente Funktionen sind.

Beispiel 1.6 (Allgemeiner Funktionenraum). Es seien $\varnothing \neq \Omega \subseteq \mathbb{R}^s$ und

$$\mathscr{F}(\Omega,\mathbb{K}) := \{f : \Omega \to \mathbb{K}\}$$

die Menge aller auf Ω definierten $\mathbb{K}$-wertigen Funktionen. Eine Addition („Superposition“) $f+g$ zweier Funktionen $f,g \in \mathscr{F}(\Omega,\mathbb{K})$ lässt sich punktweise definieren durch $(f+g)(t) := f(t)+g(t)$ für alle $t \in \Omega$. Zu unterscheiden ist hier zwischen der Addition $f+g$ von Funktionen und der Addition $f(t)+g(t)$ der beiden Zahlen $f(t), g(t) \in \mathbb{K}$. Gleichermaßen lässt sich eine Skalarmultiplikation λf für $f \in \mathscr{F}(\Omega,\mathbb{K})$ und $\lambda \in \mathbb{K}$ punktweise durch $(\lambda f)(t) := \lambda f(t)$ definieren. Der Nullvektor ist gegeben durch die Nullfunktion

$$0 : \Omega \to \mathbb{K}, \quad t \mapsto 0(t) := 0,$$

die notationell nicht von der Zahl $0 \in \mathbb{K}$ unterschieden wird. Der negative Vektor zu einer Funktion $f \in \mathscr{F}(\Omega,\mathbb{K})$ ist durch die Funktion $-f$ gegeben, welche durch die Skalarmultiplikation $-f := (-1)f$ erklärt wird. Das Kommutativgesetz $f+g = g+f$ gilt für alle $f,g \in \mathscr{F}(\Omega,\mathbb{K})$, da $f(t)+g(t) = g(t)+f(t) \in \mathbb{K}$ für alle $t \in \Omega$ gilt. In gleicher Weise lässt sich die Gültigkeit aller Assoziativ-, Kommutativ- und Distributivgesetze in $\mathscr{F}(\Omega,\mathbb{K})$ überprüfen. Folglich ist $\mathscr{F}(\Omega,\mathbb{K})$ ein linearer Raum, die Vektoren sind Funktionen, die Vektoraddition ist als Überlagerung von Funktionen erklärt und die Skalarmultiplikation als Skalierung von Funktionen. Die Cosinusfunktion $\cos : \mathbb{R} \to \mathbb{R}$ ist ein Element (ein „Punkt“) des Raums $\mathscr{F}(\mathbb{R},\mathbb{R})$. Wir schreiben $\cos \in \mathscr{F}(\mathbb{R},\mathbb{R})$ genau so, wie wir $(1,0)^\top \in \mathbb{R}^2$ schreiben. ◁

Untervektorräume von $\mathscr{F}(\Omega,\mathbb{K})$ sind beispielsweise durch Mengen stetig differenzierbarer Funktionen gegeben. Dazu vorab ein paar Festlegungen. Für $\varepsilon > 0$ und

$x \in \mathbb{R}^s$ sei

$$K(x,\varepsilon) := \{y \in \mathbb{R}^s;\ \|x-y\|_2 < \varepsilon\}.$$

Wir nennen eine Teilmenge

$$\Omega \subseteq \mathbb{R}^s \textbf{ offen } :\Longleftrightarrow \text{ für jedes } x \in \Omega \text{ existiert ein } \varepsilon > 0 \text{ mit } K(x,\varepsilon) \subseteq \Omega. \quad (1.15)$$

Eine Menge $\Omega \subseteq \mathbb{R}^s$ heißt **zusammenhängend**, wenn es *nicht* möglich ist, zwei disjunkte, nicht leere offene Mengen $\Omega_1, \Omega_2 \subseteq \mathbb{R}^s$ so zu finden, dass $\Omega \subseteq \Omega_1 \cup \Omega_2$ gilt. Eine offene, zusammenhängende Menge Ω nennt man **Gebiet**. Mit $\Omega^c := \mathbb{R}^s \setminus \Omega$ wird das **Komplement** von Ω in $\mathbb{R}^s$ bezeichnet. Man nennt

$$\Omega \subseteq \mathbb{R}^s \textbf{ abgeschlossen } :\Longleftrightarrow \Omega^c \text{ offen.}$$

Eine Menge $\varnothing \neq \Omega \subseteq \mathbb{R}^s$ heißt **beschränkt**, wenn es eine Konstante $N > 0$ gibt, so dass für ein $x \in \Omega$

$$\Omega \subseteq K(x,N).$$

Eine Menge $\Omega \subseteq \mathbb{R}^s$ heißt **kompakt**, wenn jede Folge $(x_n)_{n\in\mathbb{N}} \subseteq \Omega$ eine (in Ω) konvergente Teilfolge hat. Für $\Omega \subseteq \mathbb{R}^s$ ist das genau dann der Fall, wenn Ω abgeschlossen und beschränkt ist. Der **Rand** einer Menge $\Omega \subseteq \mathbb{R}^s$ ist definiert durch

$$\partial\Omega := \{x \in \mathbb{R}^s;\ K(x,\varepsilon) \cap \Omega \neq \varnothing \text{ und } K(x,\varepsilon) \cap \Omega^c \neq \varnothing \text{ für alle } \varepsilon > 0\}$$

und jedes Element $x \in \partial\Omega$ heißt **Randpunkt** von Ω. Die Vereinigung

$$\overline{\Omega} := \Omega \cup \partial\Omega, \quad \Omega \subseteq \mathbb{R}^s,$$

nennt man den **Abschluss** von Ω in $\mathbb{R}^s$. Ein Vektor

$$\alpha = (\alpha_1, \ldots, \alpha_s)^\top \in \mathbb{N}_0^s, \quad s \in \mathbb{N},$$

heißt **Multiindex**. Wir definieren $|\alpha| = \sum_{i=1}^s \alpha_i$. Die partiellen Ableitungen der Ordnung $|\alpha|$ einer Funktion $f : \mathbb{R}^s \to \mathbb{R}$ in einem Punkt $x \in \mathbb{R}^s$ werden in der Form

$$D^\alpha f(x) = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_s^{\alpha_s}}(x)$$

geschrieben.

Beispiel 1.7 (Vektorraum stetig differenzierbarer Funktionen). Es sei $\varnothing \neq \Omega \subseteq \mathbb{R}^s$ eine beliebige Teilmenge des Euklidischen Raums $\mathbb{R}^s$. Durch

$$C(\Omega) := \{f : \Omega \to \mathbb{R};\ f \text{ stetig}\}$$

wird die Teilmenge $C(\Omega) \subset \mathscr{F}(\Omega, \mathbb{R})$ aller stetigen, reellwertigen Funktionen auf Ω erklärt. Da die Summe zweier stetiger Funktionen wieder eine stetige Funktion ist und ebenso das skalare Vielfache einer stetigen Funktion wieder eine stetige Funktion ist, handelt es sich bei $C(\Omega)$ um einen Untervektorraum von $\mathscr{F}(\Omega, \mathbb{R})$.

Nun sei $\Omega \subseteq \mathbb{R}^s$ eine *offene* Teilmenge des $\mathbb{R}^s$. Für $k \in \mathbb{N}_0$ sei

$$C^k(\Omega) := \{f : \Omega \to \mathbb{R};\ D^\alpha f \in C(\Omega) \text{ für } \alpha \in \mathbb{N}_0^s,\ |\alpha| \le k\}, \tag{1.16}$$

wobei $C^0(\Omega) := C(\Omega)$. Die Menge $C^k(\Omega)$ heißt **Menge der k-mal stetig differenzierbaren reellwertigen Funktionen auf Ω**. Auch $C^k(\Omega)$ ist ein Untervektorraum von $\mathscr{F}(\Omega,\mathbb{R})$. Da die Menge Ω als offen vorausgesetzt wurde, enthält sie keine Randpunkte und es entstehen keine Schwierigkeiten bei der Definition der Differentialquotienten $D^\alpha f(x)$. Im eindimensionalen Fall jedoch können Ableitungen von auf abgeschlossenen Intervallen $[a,b] \subset \mathbb{R}$, $a < b$, definierten Funktionen auch in den Randpunkten a und b als einseitige Differentialquotienten definiert werden. Es ist dann sinnvoll, von k-mal stetig differenzierbaren Funktionen $f : [a,b] \to \mathbb{R}$ zu sprechen. Wir benutzen die abkürzenden Bezeichnungen

$$C^k(a,b) := C^k((a,b)) \quad \text{und} \quad C^k[a,b] := C^k([a,b]), \qquad k \in \mathbb{N}_0,$$

für die Räume der k-mal stetig differenzierbaren Funktionen $f : (a,b) \to \mathbb{R}$ beziehungsweise $f : [a,b] \to \mathbb{R}$.

Für komplexwertige, k-mal stetig differenzierbare Funktionen benutzen wir die Schreibweise $C^k(\Omega,\mathbb{C})$ anstelle von (1.16). ◁

Definition 1.8 (Operator, Nullraum, Bildraum). *Ein **Operator** ist eine Abbildung $T : D \to Y$, wobei $D \subseteq X$ und X und Y Vektorräume sind. Ein Operator $T : D \to Y$ heißt **linear**, wenn D ein linearer Raum ist und wenn die Identitäten*

$$T(x+y) = T(x) + T(x) \quad \textit{und} \quad T(\lambda x) = \lambda T(x),$$

*genannt **Additivität** und **Homogenität**, für alle $x, y \in D$ und alle $\lambda \in \mathbb{K}$ gelten. Bei linearen Operatoren schreibt man meist abkürzend Tx statt $T(x)$. Die Menge*

$$\mathscr{N}(T) := \{x \in D;\ T(x) = 0\}$$

*ist ein linearer Raum, der sogenannte **Nullraum** von T, wenn T linear ist. Auch*

$$\mathscr{R}(T) := \{y = T(x);\ x \in D\}$$

*ist ein linearer Raum, der sogenannte **Bildraum** von T, wenn T linear ist.* ◁

Beispiel 1.9 (Lineare Operatoren). Der **Integraloperator**

$$I : C[a,b] \to C^1[a,b], \quad x \mapsto y, \quad y(s) = \int_a^s x(t)\,dt, \quad a \le s \le b,$$

der jeder Funktion $x \in C[a,b]$ eine Stammfunktion zuordnet, ist linear.