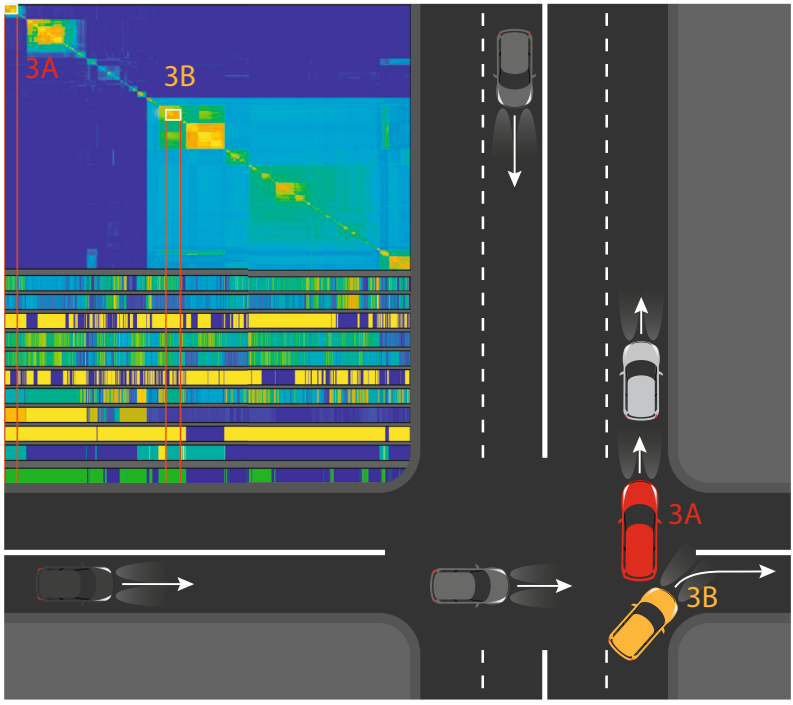
Michael Botsch
Wolfgang Utschick

Fahrzeugsicherheit und automatisiertes Fahren

Methoden der Signalverarbeitung und des maschinellen Lernens



HANSER

Botsch / Utschick

Fahrzeugsicherheit und automatisiertes Fahren

Michael Botsch
Wolfgang Utschick

Fahrzeugsicherheit und automatisiertes Fahren

Methoden der Signalverarbeitung und des maschinellen Lernens

HANSER

Die Autoren:
Prof. Dr.-Ing. Michael Botsch, Technische Hochschule Ingolstadt
Prof. Dr.-Ing. Wolfgang Utschick, Technische Universität München

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

www.hanser-fachbuch.de
Lektorat: Julia Stepp
Coverkonzept: Marc Müller-Bremer, www.rebranding.de, München
Coverrealisation: Max Kostopoulos
Satz: Michael Botsch
Druck und Bindung: CPI buchbücher.de GmbH Birkach
Printed in Germany

Print-ISBN: 978-3-446-45326-5
E-Book-ISBN: 978-3-446-46804-7

Vorwort

Die Inhalte dieses Lehrbuchs basieren auf den Vorlesungen „Signalverarbeitung in der Fahrzeugsicherheit“, „Wissensmodellierung und Maschinelles Lernen“, „Sensor Technology and Signal Processing“, „Integrated Safety and Assistance Systems“, „Mathematische Methoden der Signalverarbeitung“, „Convex Optimization“, „Statistical Signal Processing“ und „Signal Processing and Machine Learning“, die wir an der Technischen Hochschule Ingolstadt und an der Technischen Universität München anbieten.

In Kapitel 1 wird eine kurze Einführung in Aspekte des automatisierten Fahrens und der integralen Fahrzeugsicherheit gegeben. Grundlagen der Signalverarbeitung und Mathematik, die für das Verständnis und die Auslegung von Algorithmen für das automatisierte Fahren und die Fahrzeugsicherheit unerlässlich sind, werden in Kapitel 2 wiederholt und vertieft. In Kapitel 3 werden Fahrzeugmodelle vorgestellt, die sich zur Modellierung des Fahrverhaltens eignen und damit die Grundlage für die Prädiktion und Regelung von Fahrzeugtrajektorien sind. Ausgehend von den Fahrzeugmodellen stellt Kapitel 4 Verfahren der Signalverarbeitung für die Zustandsschätzung von Fahrzeugen und Objekten im Fahrzeugumfeld vor, die einen zentralen Bestandteil von Trackingverfahren und Sensordatenfusionsmethoden darstellen. In Kapitel 5 werden die Methoden der Signalverarbeitung um rein datenbasierte Verfahren erweitert. Es handelt sich um maschinelle Lernverfahren, die angesichts der Komplexität des sicheren automatisierten Fahrens als Schlüsselmethoden zur Realisierung von Innovationen im Automobilbau gelten.

Das Buch enthält am Ende eines jeden Kapitels Übungsaufgaben mit ausführlichen Lösungsvorschlägen. Sie sollen den Leser bei der intensiven Auseinandersetzung mit dem behandelten Stoff unterstützen. Für Aufgaben, bei denen es erforderlich ist, werden Matlab-Skripte auf der Internetseite zum Buch[1] zur Verfügung gestellt.

Das Buch richtet sich an Studierende und Doktoranden der Elektrotechnik, Mechatronik und Informatik an Universitäten und Fachhochschulen sowie an Ingenieure in der Praxis, die ein grundlegendes Verständnis von Methoden der Signalverarbeitung für das sichere automatisierte Fahren erwerben wollen.

Wir danken unserer Lektorin Frau Julia Stepp für die angenehme Zusammenarbeit, Prof. Andreas Gaull und Anja Zupfer für das Korrekturlesen sowie den wissenschaftlichen Mitarbeitern Amit Chaulwar, Friedrich Kruber, Marcus Müller, Parthasarathy Nadarajan, Eduardo Sanchez Morales und Jonas Wurst für die gemeinsamen Forschungsarbeiten an Anwendungen von maschinellen Lernverfahren im Bereich des sicheren automatisierten Fahrens, die kurz in Unterkapitel 5.7 beschrieben werden. Besonders möchten wir uns bei unseren Familien, bei Anja, Hannah, Steffi und Viktoria für ihre tatkräftige und liebevolle Unterstützung bedanken.

Ingolstadt, im Mai 2020 Michael Botsch und Wolfgang Utschick

[1] *http://www.fahrzeugsicherheitundautomatisiertesfahren.de*

Inhalt

1 Einführung in das automatisierte Fahren und die Fahrzeugsicherheit

Das automatisierte Fahren und die dafür erforderliche Fahrzeugsicherheit werden die Zukunft der Mobilität entscheidend prägen. In diesem Kapitel soll eine kurze Einführung in die Herausforderungen dieser Entwicklungen und die Bedeutung der Signalverarbeitung im Kontext des sicheren automatisierten Fahrens gegeben werden.

Lernziele in Kapitel 1

Der Lernende[1] …

- kennt die fünf Stufen des automatisierten Fahrens und kann Fahrzeugfunktionen dem entsprechenden Automatisierungsgrad zuordnen;
- kennt die Vorteile, die man sich vom automatisierten Fahren erhofft;
- kennt grundlegende Aspekte des rechtlichen und ethischen Rahmens beim automatisierten Fahren;
- kennt wichtige Begriffe der Verkehrssicherheit, insbesondere der integralen Fahrzeugsicherheit;
- kann die Notwendigkeit zur Verbesserung der integralen Fahrzeugsicherheitssysteme mit Hilfe von Statistiken zum Unfallgeschehen begründen;
- versteht die technische Komplexität, die sich hinter Fahrzeugfunktionen des automatisierten Fahrens und der integralen Fahrzeugsicherheit verbirgt;
- versteht die Bedeutung der Signalverarbeitung für die Realisierung von Fahrzeugfunktionen des automatisierten Fahrens und der integralen Fahrzeugsicherheit.

1.1 Automatisiertes Fahren

Zu den drei „Megatrends“ der Mobilität werden die **Digitalisierung**, die **Elektrifizierung** und die **Urbanisierung** gezählt. Die drei Bereiche sind miteinander verzahnt und beeinflussen sich gegenseitig. Das Auto spielt für all diese Bereiche eine zentrale Rolle, und es wird prognostiziert, dass das Auto der Zukunft „autonom, vernetzt und elektrisch“ fahren wird. Bereits heute wird in der Automobilindustrie an Lösungen gearbeitet, die es ermöglichen, die Fahraufgaben

[1] Um die Lesbarkeit des Buchs zu erleichtern, wird bei den Personenbezeichnungen die männliche Form verwendet. Es werden jedoch Personen jeglichen Geschlechts (m/w/d) gleichberechtigt angesprochen.

zunehmend ohne Eingreifen des Fahrers zu bewältigen, und das **automatisierte Fahren** gehört ohne Zweifel zu den großen Zukunftsthemen der Mobilität. Das automatisierte Fahren bzw. in seiner letzten Ausbaustufe auch **autonomes Fahren** genannt, ermöglicht es, während der Fahrzeit anderen Tätigkeiten als der Fahrzeugführung nachzugehen, und verhilft Personen mit eingeschränkten Fahrfähigkeiten zu einem mobilen Leben. Doch zusätzlich zu diesen beiden Aspekten birgt das autonome Fahren sowohl im Bereich der Sicherheit als auch im Bereich der Verkehrseffizienz große Potenziale. Weil ca. 90 % der Unfälle mit Toten oder Schwerverletzten auf menschliches Fehlverhalten zurückzuführen sind, ist zu erwarten, dass automatisiertes und vernetztes Fahren die Schäden und das Leid durch Verkehrsunfälle stark reduzieren kann. Im Bereich der Verkehrseffizienz kann durch einen besser aufeinander abgestimmten Verkehr die Kapazitätsauslastung optimiert werden, was angesichts der Prognosen für die Zukunft notwendig sein wird: Laut [VBW17] geht das Bundesverkehrsministeriums für das Jahr 2030 von einem Zuwachs um 13 % beim Personenverkehr und um 38 % beim Güterverkehr aus.

Übung 1.1

Während Aspekte der Signalverarbeitung für das automatisierte Fahren in den anschließenden Kapiteln im Vordergrund stehen, sollen im Folgenden relevante Informationen zu den Stufen des automatisieren Fahrens sowie dem rechtlichen und ethischen Rahmen kurz vorgestellt werden, um eine Einordnung in einen größeren Zusammenhang zu ermöglichen.

Stufen des automatisierten Fahrens

Der Übergang vom manuellen zum autonomen Fahren findet nicht schlagartig statt, sondern im Rahmen einer Entwicklung, bei der die Automatisierung in fünf Stufen untergliedert ist [SAE14, Bun15]. Diese Stufen sind in Abb. 1.1 dargestellt und beschreiben, welche Anforderungen sowohl an die Systeme als auch an den Fahrer gestellt werden.

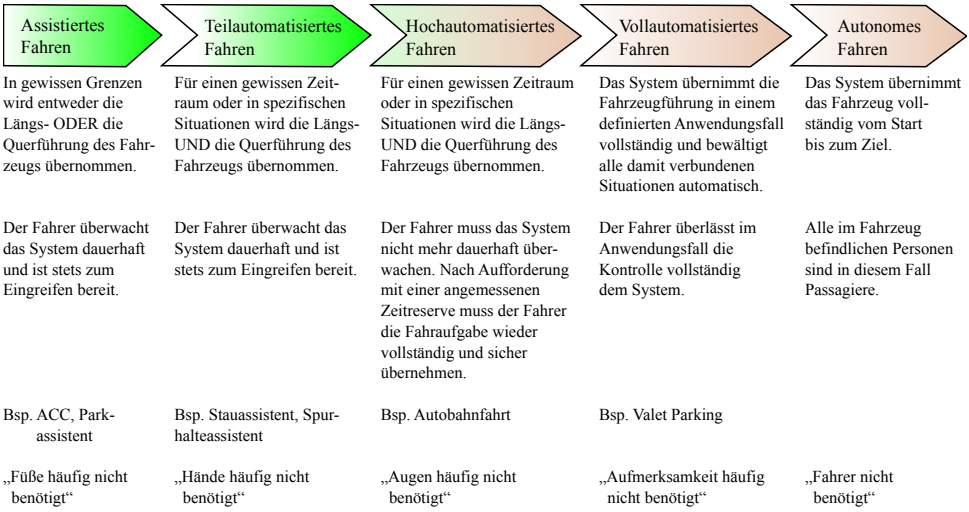


Abbildung 1.1 Stufen des automatisierten Fahrens.

Die meisten Fahrzeuge, die heute auf deutschen Straßen unterwegs sind, gehören der Stufe 1, **Assistiertes Fahren**, an und beinhalten Fahrzeugfunktionen wie Tempomat, Parkassistent oder Adaptive Cruise Control (ACC) und erfordern, dass der Fahrer die Hände am Lenker hat, den Verkehr überwacht und jederzeit die Kontrolle über das Fahrzeug übernehmen kann. Diese Systeme haben häufig Einschränkungen, die dazu führen, dass sie erst ab einer bestimmten Geschwindigkeit aktiv oder bei schlechten Witterungsbedingungen nicht verfügbar sind. Einprägsam für diese Stufe ist der Satz „Füße werden beim Fahren häufig nicht mehr benötigt“.

Bei Stufe 2, **Teilautomatisiertes Fahren**, der einige Fahrzeuge seit ca. dem Jahr 2015 zugeordnet werden können, ist das Fahrzeug in der Lage, in einigen Situationen autonom zu fahren, z. B. auf der Autobahn der Spur zu folgen und gleichzeitig den Abstand zum Vordermann zu regeln oder im Stau vollständig autonom zu fahren. Der Fahrer muss allerdings auch bei dieser Automatisierungsstufe das System dauerhaft überwachen und jederzeit zum Eingreifen bereit sein. Zu dieser Stufe gehören Fahrzeuge wie der Audi Q7 (2015), die BMW 7er Reihe (2015), die Mercedes E-Klasse (2016) oder der Tesla Model S (2014). Einprägsam für diese Stufe ist der Satz „Hände werden beim Fahren häufig nicht mehr benötigt“.

Bei Stufe 3, **Hochautomatisiertes Fahren**, übernimmt das Fahrzeug in einer spezifischen Situation, wie z. B. der Autobahnfahrt, für einen gewissen Zeitraum komplett die Fahraufgaben inklusive komplexer Manöver wie dem Überholen. Der Fahrer muss das Fahrzeug nicht mehr dauerhaft überwachen. Er muss nur in der Lage sein, nach einer Aufforderung mit angemessener Zeitreserve die Kontrolle über das Fahrzeug zu übernehmen, z. B. um eine sehr komplexe Fahraufgabe wie etwa bei Baustellen oder dem Herunterfahren von der Autobahn zu meistern. Rechtliche und technische Hürden haben dafür gesorgt, dass Personenkraftwagen (PKW) der Automatisierungsstufe 3 bisher noch nicht auf dem Markt sind. Allerdings ist zu erwarten, dass dies bald der Fall sein wird. Einprägsam für diese Stufe ist der Satz „Augen werden beim Fahren häufig nicht mehr benötigt“.

Bei Stufe 4, **Vollautomatisiertes Fahren**, bewältigt das Fahrzeug in definierten Anwendungsfällen vollständig die Fahraufgaben, und der Fahrer kann in diesen Anwendungsfällen die Kontrolle komplett übergeben. In dieser Stufe sollen nicht nur Autobahnfahrten, sondern auch Landstraßen- und ein Großteil der Stadtfahrten autonom bewältigt werden können. Auch das Einparken soll komplett ohne Fahrer im Fahrzeug realisiert sein, das sogenannte „Valet Parking“. Der Sprung zu dieser Stufe ist sowohl technisch als auch unter rechtlichen und ethischen Gesichtspunkten sehr groß, ermöglicht es aber dem Fahrer, seine Fahrzeit für ganz andere Tätigkeiten zu nutzen. Eine Prognose darüber, wann diese Automatisierungsstufe zum Alltag auf den Straßen gehört, ist schwierig. Wenn man die Entwicklungen der letzten Jahre extrapolieren darf, erscheint ein Zeithorizont von ca. 10 Jahren als denkbar. Einprägsam für diese Stufe ist der Satz „Aufmerksamkeit wird beim Fahren häufig nicht mehr benötigt“.

Bei Stufe 5, **Autonomes Fahren**, übernimmt das Fahrzeug in allen Situationen alle Fahraufgaben, und der Fahrer ist nur noch Passagier. Dieses Ziel wird man in seiner Gesamtheit wohl nie erreichen können, z. B. weil alle Sensoren, unabhängig davon, ob sie im Fahrzeug oder außerhalb verbaut sind, ihre Grenzen haben. Trotzdem wird man dem Ziel nahe kommen und Fahrzeuge dieser Automatisierungsstufe zuordnen. Entsprechend schwer ist eine Prognose darüber, wann diese Stufe zum Alltag auf den Straßen gehört. Eine besondere Herausforderung für das Erreichen dieser Stufe ist der sogenannte Mischverkehr, bei dem Fahrzeuge mit verschiedenem Automatisierungsgrad auf den Straßen unterwegs sind und eine Kooperation zwischen allen Verkehrsteilnehmern und der Infrastruktur in vielen Verkehrsszenarien nicht möglich ist. Einprägsam für diese Stufe ist der Satz „Fahrer wird nicht mehr benötigt“.

Es sollte erwähnt werden, dass im Bereich des öffentlichen Personenverkehrs dank fester Routen und der Möglichkeit, eine höhere Anzahl an Sensoren zu verbauen als in PKW, der Einsatz von hochautomatisierten Bussen bereits weltweit an verschiedenen Orten in der Öffentlichkeit erprobt wird. Zum Beispiel fahren hochautomatisierte Busse, die Passagiere auf fest vorgegebenen Strecken ohne Fahrer transportieren, im Versuchsbetrieb in den Orten Bad Birnbach seit 2017 und in Berlin seit 2019. Für diese Busse gibt es Ausnahmegenehmigungen, und sie dürfen nur auf vordefinierten Spuren, mit niedrigen Geschwindigkeiten und unter Aufsicht von Begleitpersonen unterwegs sein. Es handelt sich also bei diesen Bussen nicht um die Automatisierungsstufe 5; sie leisten aber einen wertvollen Beitrag, um die Akzeptanz für das fahrerlose Fahren in der Bevölkerung zu erhöhen.

Übung 1.2

Rechtliche Aspekte des automatisierten Fahrens

Entscheidend für das automatisierte Fahren ist der rechtliche Rahmen, an den sich diese Form der Mobilität halten muss. Es gibt in diesem Kontext Fragen des Verkehrs- und Zulassungsrechts, die an den technischen Fortschritt peu à peu angepasst werden. Das **Wiener Übereinkommen** von 1968 [Uni68] bildet die völkerrechtliche Grundlage für die nationalen Verkehrsregeln, und in seiner ursprünglichen Fassung wird gefordert, dass jedes Fahrzeug von einem Fahrer, der jederzeit die Kontrolle über das Fahrzeug hat, geführt werden muss.[2] Um dem technologischen Fortschritt gerecht zu werden, wurde das Wiener Übereinkommen 2014 durch die United Nations angepasst, so dass Systeme, welche die Führung eines Fahrzeuges beeinflussen, zulässig sind, wenn sie jederzeit vom Fahrer überstimmt oder abgeschaltet werden können [Uni14].[3]

Es herrscht eine hohe Dynamik bezüglich der Anpassung des rechtlichen Rahmens an die technischen Möglichkeiten zur Realisierung des automatisierten Fahrens. In Kalifornien (USA) wurde bereits im Jahr 2015 erlaubt, dass automatisierte Fahrfunktionen im Straßenverkehr getestet werden, vorausgesetzt, dass jederzeit ein Fahrer in der Lage ist, die Kontrolle über das autonom fahrende Auto zu übernehmen [Cal15]. Anschließend wurde vom California Department of Motor Vehicles an Regelungen gearbeitet, um Tests mit fahrerlosen Autos zuzulassen [Cal17]. Im Jahr 2018 hat es die California Public Utilities Commission einigen Industrieunternehmen genehmigt, im Rahmen eines Pilotprojekts auch Passagiere in autonomen Fahrzeugen zu transportieren [Cal18]. Dabei ist es erlaubt, dass sich kein Fahrer im Fahrzeug befindet, dafür muss aber eine Kommunikationsverbindung zwischen dem Fahrzeug und einem

[2] Artikel 8, Absatz 1: “Every moving vehicle or combination of vehicles shall have a driver”;
Artikel 8, Absatz 5: “Every driver shall at all times be able to control his vehicle or to guide his animals”;
Artikel 13 Absatz 1: “Every driver of a vehicle shall in all circumstances have his vehicle under control so as to be able to exercise due and proper care and to be at all times in a position to perform all manoeuvres required of him. He shall, when adjusting the speed of his vehicle, pay constant regard to the circumstances, in particular the lie of the land, the state of the road, the condition and load of his vehicle, the weather conditions and the density of traffic, so as to be able to stop his vehicle within his range of forward vision and short of any foreseeable obstruction. He shall slow down and if necessary stop whenever circumstances so require, and particularly when visibility is not good.”

[3] Kapitel V. Teil A: “Vehicle systems which influence the way vehicles are driven ... shall be deemed to be in conformity with paragraph 5 of this Article and with paragraph 1 of Article 13, when such systems can be overridden or switched off by the driver”.

„remote operator“ während des Betriebs gewährleistet sein. Auch in anderen Ländern wie Japan, Großbritannien oder Schweden werden nach und nach Änderungen des rechtlichen Rahmens vorgenommen [MGLW15].

In Deutschland hat der Bundestag Regelungen zur Änderung des Straßenverkehrsgesetzes[4] (StVG) für das Fahren von Autos mit hoch- und vollautomatisierter Fahrfunktion im März 2017 verabschiedet [Deu17]. Dabei wird das StVG dahingehend ergänzt, dass der Fahrzeugführer dem technischen System in bestimmten Situationen die Fahrzeugsteuerung übergeben kann. Auch die Haftungsfrage im Falle eines Unfalls wird hierbei festgelegt: „Die Inanspruchnahme des Halters im Wege der Gefährdungshaftung wird dazu führen, dass die Haftpflichtversicherung des Halters und die Versicherung des Herstellers klären, wer im Ergebnis die Kosten des Unfalls zu tragen hat“, wobei jedes Fahrzeug aufzeichnen muss, „wann das automatisierte System zur Fahrzeugsteuerung eingeschaltet war und wann nicht und wann das automatisierte System den Fahrzeugführer zur Übernahme der Fahrzeugsteuerung aufforderte“, um so sicherzustellen, „dass der Fahrzeugführer sich nicht pauschal auf ein Versagen des automatisierten Systems berufen kann“. Es ist zu erwarten, dass in der Folge entsprechende Anpassungen auch in der Straßenverkehrs-Ordnung (StVO) gemacht werden, um nicht mehr die dauernde Beherrschbarkeit eines Fahrzeugs durch den Fahrer (StVO, §3 Absatz 1) zu fordern. Auch Änderungen in der Fahrzeug-Zulassungsverordnung sind zu erwarten. Gemäß §3 Absatz 1 ist für Kraftfahrzeuge in Deutschland eine Zulassung, bei der geprüft wird, ob das Fahrzeug einem genehmigten Typ entspricht, notwendig. Die Anforderungen an die Typgenehmigung wird in Europa durch die Richtlinie 2007/46/EG des Europäischen Parlaments und des Rates von 2007 geregelt. Die Europäische Union (EU) gehört zu den Unterzeichnern der angeschlossenen Regelungen zum „Übereinkommen über die Annahme einheitlicher technischer Vorschriften für Radfahrzeuge, Ausrüstungsgegenstände und Teile, die in Radfahrzeuge(n) eingebaut und/oder verwendet werden können, und die Bedingungen für die gegenseitige Anerkennung von Genehmigungen, die nach diesen Vorschriften erteilt wurden“[5], und entsprechend spielen die ECE-Regelungen eine wichtige Rolle für die Typgenehmigungen für die Länder der EU. In diesem Zusammenhang ist die ECE-Regel UN-R79 und seine Anpassung aus dem Jahr 2018 zu erwähnen, in der festgehalten ist, welche autonomen Lenkeingriffe unter welchen Voraussetzungen erlaubt sind [Uni18].

Ethische Aspekte des automatisierten Fahrens

Zusätzlich zu den rechtlichen Aspekten kommen beim automatisierten Fahren ethische Fragestellungen auf, die sich bisher im Straßenverkehr nicht gestellt haben, vor allem im Zusammenhang mit Unfällen, in denen automatisierte Fahrzeuge involviert sind. Im Sommer 2017 hat eine vom Bundesministerium für Verkehr und digitale Infrastruktur eingesetzte Ethikkommission zum automatisiertennFahren einen Bericht vorgelegt, aus dem folgende Kernpunkte erwähnt werden sollten [Bun17]:

[4] Das Straßenverkehrsgesetz ist ein Bundesgesetz, das die Grundlagen des Straßenverkehrsrechts regelt. Zusammen mit der Fahrzeug-Zulassungsverordnung (FZV), der Straßenverkehrs-Ordnung (StVO) und der Straßenverkehrs-Zulassungs-Ordnung (StVZO) wird das Straßenverkehrsrecht in Deutschland zum größten Teil festgelegt.

[5] Häufige vereinfachte Bezeichnung für die Regelungen zum Übereinkommen: „ECE-Regelungen“ oder „UN/ECE-Regelungen“.

- Das automatisierte und vernetzte Fahren ist ethisch geboten, wenn die Systeme weniger Unfälle verursachen als menschliche Fahrer (positive Risikobilanz).
- Sachschaden geht vor Personenschaden: In Gefahrensituationen hat der Schutz menschlichen Lebens immer höchste Priorität.
- Bei unausweichlichen Unfallsituationen ist jede Qualifizierung von Menschen nach persönlichen Merkmalen (Alter, Geschlecht, körperliche oder geistige Konstitution) unzulässig.
- In jeder Fahrsituation muss klar geregelt und erkennbar sein, wer für die Fahraufgabe zuständig ist: der Mensch oder der Computer. Wer fährt, muss dokumentiert und gespeichert werden (u. a. zur Klärung möglicher Haftungsfragen).
- Der Fahrer muss grundsätzlich selbst über Weitergabe und Verwendung seiner Fahrzeugdaten entscheiden können (Datensouveränität).

Übung 1.3

1.2 Integrale Fahrzeugsicherheit und Unfallstatistiken

Ziel dieses Unterkapitels ist es, wichtige Begriffe der Fahrzeugsicherheit einzuführen und mit Hilfe von Unfallstatistiken die Motivation für die Entwicklung und Erforschung von Fahrzeugsicherheitssystemen zu begründen.

Die Norm IEC 61508 definiert **Sicherheit** als „Freiheit von unvertretbaren Risiken“. Das **Risiko** $R(\mathscr{H}_i)$ eines Ereignisses $\mathscr{H}_i$ wird von den Sicherheitsnormen als Produkt seiner Eintrittswahrscheinlichkeit $\mathrm{P}(\mathscr{H}_i)$ und dessen Schadensausmaß $K(\mathscr{H}_i)$ definiert:

$$R(\mathscr{H}_i) = \mathrm{P}(\mathscr{H}_i) K(\mathscr{H}_i). \tag{1.1}$$

Das Gesamtrisiko R_{Sys} eines Systems ergibt sich aus den Risiken der einzelnen Ereignisse

$$R_{\mathrm{Sys}} = \sum_i R(\mathscr{H}_i) \tag{1.2}$$

und ist somit der Erwartungswert des möglichen Schadensausmaßes. Auch bei etablierten technischen Systemen ist immer ein Restrisiko vorhanden, das jedoch von der Gesellschaft akzeptiert werden kann, wenn es kleiner als bereits vorhandene Risiken ist [LPP10]. Das tolerierbare Risiko R_{tol}, das die Sicherheit definiert, lässt sich im Allgemeinen nicht quantifizieren und hängt sehr stark von dem betrachteten System ab. Versteht man die Sicherheit eines Systems komplementär zu dem Begriff **Gefahr** und zwar so, dass hohe Sicherheit ein geringes Risiko und hohe Gefahr ein hohes Risiko bedeuten, so lässt sich die Sicherheit eines Systems erhöhen, indem entweder die Eintrittswahrscheinlichkeiten der Schadensereignisse oder die zugehörigen Schadensausmaße verringert werden. Abb. 1.2 stellt die Begriffe Sicherheit und Gefahr eines Systems in Abhängigkeit von dem Risiko R_{Sys} dar.

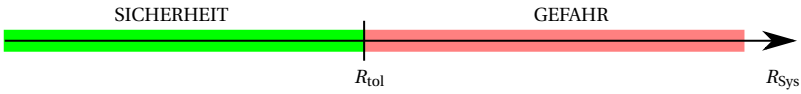


Abbildung 1.2 Sicherheit und Gefahr in Abhängigkeit von dem Risiko R_{Sys}.

Übung 1.4

Überträgt man diese Begriffe auf die **Verkehrssicherheit**, so kann man das Ziel, Unfälle zu vermeiden, als „Verringerung der Eintrittswahrscheinlichkeiten der Schadensereignisse“ und das Ziel der Unfallfolgenminderung, als „Verringerung der Schadensausmaße“ einführen. Um die Verkehrssicherheit zu erhöhen, können Maßnahmen in den verschiedenen Bereichen der Verkehrssicherheit umgesetzt werden. Diese Bereiche sind in Abb. 1.3 dargestellt. Im Folgenden soll insbesondere der Begriff der **Fahrzeugsicherheit** genauer erläutert werden.

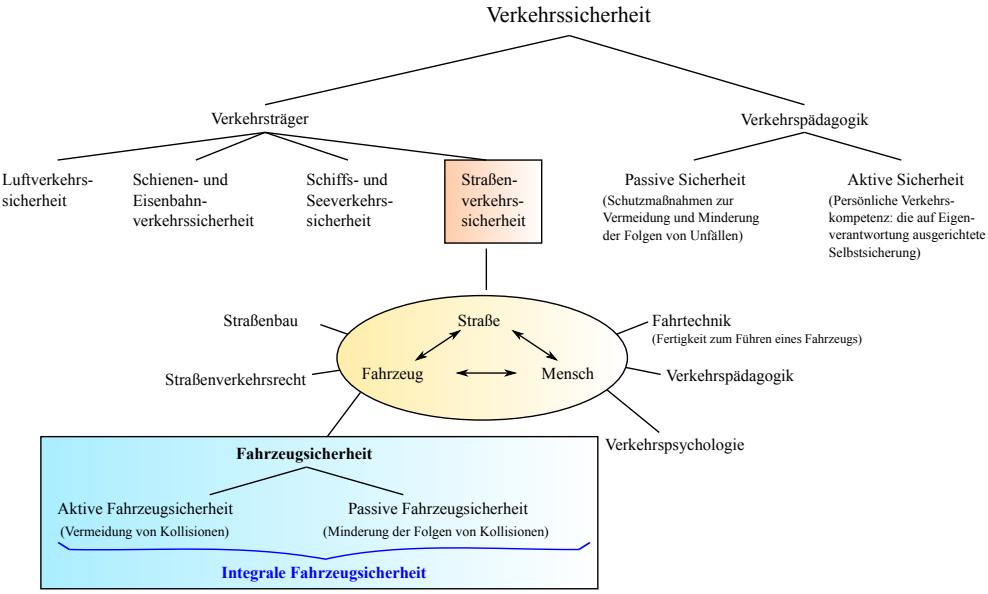


Abbildung 1.3 Übersicht Verkehrssicherheit.

Die Fahrzeugsicherheit gehört zur Straßenverkehrssicherheit und ist ein Teilaspekt des Systems Umwelt-Fahrzeug-Mensch. Sie lässt sich untergliedern in die **„passive“** und die **„aktive“ Fahrzeugsicherheit**. Die passive Fahrzeugsicherheit umfasst alle Maßnahmen am Fahrzeug, die bei einem nicht vermeidbaren Unfall die Folgen so gering wie möglich machen, z. B. Gurte, Airbags, Kindersitze, Struktur der Fahrzeugfront und der Fahrgastzelle, energieabsorbierende Lenksäulen, energieabsorbierende Elemente im Fahrzeuginnenraum, Stabilität des Dachs bei Überschlägen, Überrollbügel, verbesserte Sitze mit Kopfstützen, usw. Die aktive Fahrzeugsicherheit umfasst alle Maßnahmen am Fahrzeug, die darauf abzielen eine Unfall zu verhindern, z. B. können Fahrerassistenzsysteme wie Forward Collision Warning (FCW), Blind Spot Detection (BSD), Anti-Lock Braking System (ABS) oder Electronic Stability Control (ESC) dazu gezählt werden, aber vor allem autonome Eingriffe in die Längsdynamik wie die Notfallbrem-

sung (Autonomous Emergency Braking, kurz AEB) oder zukünftig autonome Eingriffe in die Querführung des Fahrzeugs. Im Gegensatz zu passiven Fahrzeugsicherheitssystemen und vollständig autonomen Eingriffen in die Fahrzeugführung, handelt es sich beim Großteil der heute verfügbaren aktiven Fahrzeugsicherheitssysteme um einen Regelkreis, bei dem der Mensch als Regler eine wichtige Rolle spielt. Alle in kritischen Verkehrssituationen warnenden Systeme zielen darauf ab, dass der Fahrer die Gefahr rechtzeitig erkennt und selber entsprechende Maßnahmen zur Unfallvermeidung einleitet. Dabei sind die objektive Sicherheit – gegeben durch physikalische Grenzen, z. B. die Kraftübertragungsmöglichkeiten zwischen Reifen und Fahrbahn – und die subjektive Sicherheit des Fahrers zu berücksichtigen. Die Maßnahmen der aktiven Fahrzeugsicherheit müssen so ausgelegt werden, dass das Fahrzeug objektiv mehr Sicherheit bietet als subjektiv wahrgenommen wird. Ist dies nicht der Fall, so kann es zur Risikokompensation kommen, d. h. die Wirkung der Maßnahmen kann dadurch kompensiert oder sogar überkompensiert werden, wenn im Wissen um diese zusätzliche objektive Sicherheit das Fahrverhalten risikoreicher wird [WMN02].

Für die gesamtheitliche Betrachtung von passiver und aktiver Fahrzeugsicherheit wurde der Begriff **integrale Fahrzeugsicherheit** eingeführt. Abb. 1.4 visualisiert den Begriff als gesamt-

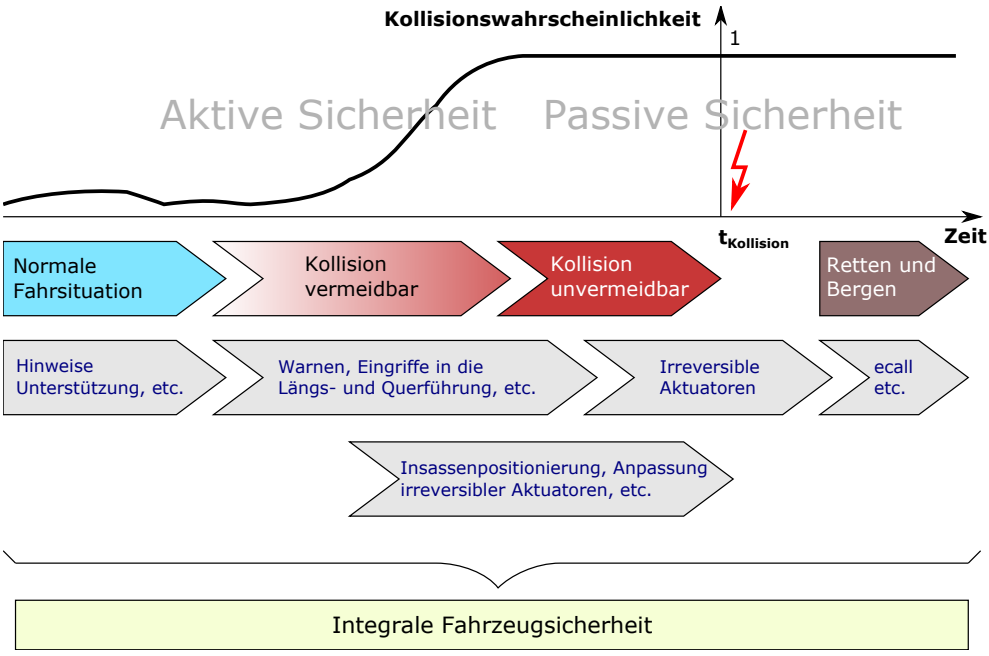


Abbildung 1.4 Integraler Ansatz: Fahrzeugsicherheit gesamthaft betrachten.

heitliche Betrachtung von aktiven und passiven Maßnahmen sowie Maßnahmen des Rettungswesens, mit dem Ziel, den Schutz aller Verkehrsteilnehmer zu steigern. Für die Zeitintervalle vor einer Kollision, während einer Kollision und nach einer Kollision werden häufig auch die Begriffe **„Pre-Crash“**, **„In-Crash“** und **„Post-Crash“** verwendet. Abb. 1.4 zeigt auch, dass die Wahrscheinlichkeit für den Eintritt eines Schadenfalls, d. h. die **Kollisionswahrscheinlichkeit**, sich in jeder Verkehrssituation mit der Zeit ändert, d. h. zu jedem Zeitpunkt t_i gibt es, entsprechend Gl. (1.1), eine neue Wahrscheinlichkeit $P(\mathscr{H}_i)$ und damit ein neues Risiko $R(\mathscr{H}_i)$. Man sieht auch, dass die Kollisionswahrscheinlichkeit bereits vor Eintritt der Kollision den

Wert 1 erreicht, d. h. es ist fahrdynamisch nicht mehr möglich, einen Unfall zu vermeiden. Tritt dies ein, so sind Maßnahmen der passiven Sicherheit notwendig, um das Schadenmaß, also die **Unfallschwere**, zu reduzieren und damit die Fahrzeugsicherheit zu erhöhen.

Übung 1.5

Maßzahlen für die aktive und passive Sicherheit werden in [Kra88] hergeleitet und in [KFL+13] für die Jahre 1953 – 2011 vorgestellt. Dabei wird als Maßzahl für die aktive Sicherheit der Quotient aus Fahrzeugbestand und Anzahl der Unfall-Fahrzeuge, mit der Einheit KFZ/Unfall-KFZ, gewählt und als Maßzahl für die passive Sicherheit der Quotient aus der Anzahl polizeilich gemeldeter Unfälle und den Verletzungsfolgekosten, mit der Einheit Unfälle/Mio. €. Dieses Maß für die passive Sicherheit berücksichtigt die Tatsache, dass sich die volkswirtschaftlichen Verletzungsfolgekosten für tödlich Verletzte bei ca. 18 %, für Schwerverletzte bei ca. 52 % und für Leichtverletzte bei 30 % eingependelt haben [KFL+13]. Die Höhe der volkswirtschaftlichen Kosten in Deutschland durch Verkehrsunfälle für das Jahr 2017 wurde von der Bundesanstalt für Straßenwesen auf ca. 34 Milliarden Euro geschätzt, d. h. mehr als 1% des Bruttoinlandprodukts [Bun19].

In Deutschland erhebt, sammelt und analysiert das Statistische Bundesamt statistische Informationen zu Wirtschaft, Gesellschaft und Umwelt. Unter anderem veröffentlicht es monatlich und jährlich umfangreiche Daten zur Beurteilung der aktuellen Unfallentwicklung. Es gibt Zeitreihen heraus, die langfristige Vergleiche und Analysen ermöglichen. Die folgenden Statistiken ergeben sich aus den Daten des Deutschen Statistischen Bundesamts[6].

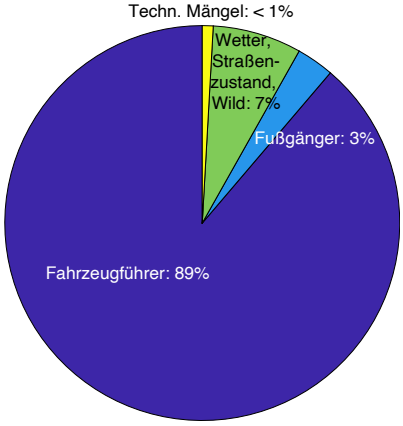


Abbildung 1.5 Ursachen für Unfälle mit Personenschaden 2018.

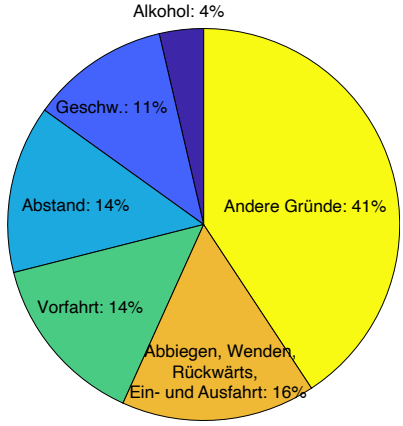


Abbildung 1.6 Fehlverhalten des Fahrzeugführers 2018.

Die häufigste Ursache für Unfälle mit Personenschaden war im Jahr 2018 mit 88,7% das Fehlverhalten des Fahrzeugführers, gefolgt mit 7,3% allgemeine Ursachen (Witterung, Straßenverhältnisse, Wild auf der Fahrbahn), 3,1% Fehlverhalten von Fußgängern und 0,9% technische Mängel [Deu19a]. Diese Aufteilung ist in Abb. 1.5 zu sehen.

Das Fehlverhalten des Fahrzeugführers lässt sich in folgende Gruppen zusammenfassen: Alkoholeinfluss (2018: 3,6%), nicht angepasste Geschwindigkeit (2018: 11,4%), nicht angepasster

[6] Seit dem 1. Januar 2006 ist die Verwendung der Statistiken des Bundesamtes lizenz-/gebührenfrei.

Abstand (2018: 13,9%), Missachtung der Vorfahrt/Vorranges (2018: 14,3%), Abbiegen, Wenden, Rückwärtsfahren, Ein- und Anfahren (2018: 16,0%), Sonstige Ursachen (2018: 40,8%). Für das Jahr 2018 findet man die prozentuale Aufteilung dieser Fehler in Abb. 1.6.

Übung 1.6

Betrachtet man allerdings das Schadensausmaß, so zählen erhöhte Geschwindigkeit, die falsche Straßenbenutzung und Alkoholeinfluss zu den häufigsten Ursachen der Unfälle mit Getöteten [Deu19b]. Im Jahr 2018 sind in Deutschland 3275 Getötete (ca. 9 Getötete pro Tag), 67967 Schwerverletzte und 2,6 Mio. polizeilich erfasste Straßenverkehrsunfälle [Deu19b] zu verzeichnen. In Europa waren es 25100 Getötete (ca. 69 Getötete pro Tag) [Kom19] und weltweit ca. 1,35 Mio. Getötete (ca. 3700 Getötete pro Tag) [WHO18] im Straßenverkehr. In Abb. 1.7 wird die Anzahl der Getöteten in Deutschland seit 1950 dargestellt sowie die Anzahl der gemeldeten PKW in Deutschland [DES16, Sta17]. Man sieht in der Abbildung, dass die Anzahl

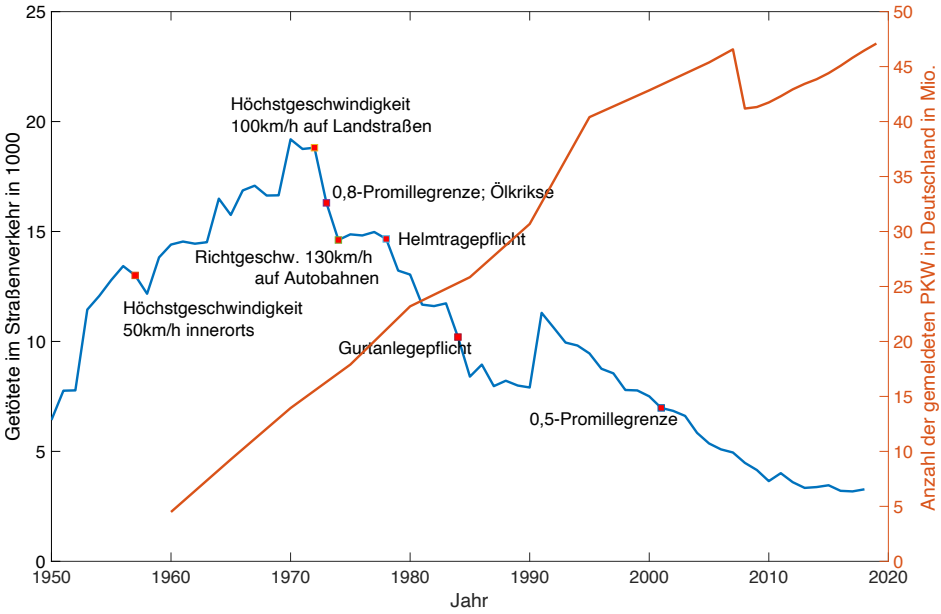


Abbildung 1.7 Anzahl Getöteter und Anzahl gemeldeter PKW in Deutschland.

Getöteter im Straßenverkehr sich seit 1970 auf ein Sechstel verringert hat, obwohl die Anzahl der angemeldeten PKW sich seitdem mehr als verdreifacht hat. Die Gründe für diese positive Entwicklung sind vor allem:

- Verkehrsrechtliche Regelungen wie z. B. die Einführung der Gurtanlege- und Helmtragepflicht oder die Senkung der Höchstgrenze für den Blutalkoholkonzentrationswert;
- Verbesserung der technischen Ausstattung von Fahrzeugen;
- Maßnahmen in der Infrastruktur, z. B. bessere Straßengestaltung, Trennung von geschützten und ungeschützten Verkehrsteilnehmern (Fußgängerzonen, Radwege);
- Bessere Verkehrserziehung;
- Verbesserte medizinische Erstversorgung.

Übung 1.7

Trotz dieser positiven Tendenz gibt es immer noch täglich ca. 9 Getötete im Straßenverkehr in Deutschland, und das menschliche Leid, das dadurch verursacht wird, lässt sich nicht in Zahlen fassen. Durch die Vermeidung von Fehlern des Fahrzeugführers kann die Sicherheit im Straßenverkehr durch Fahrzeugsicherheitssysteme noch stark erhöht werden. Allein schon die Analyse des Bremsverhaltens bei Unfällen zeigt, dass der größte Teil der Fahrer vor einer Kollision nicht ausreichend bremst [PS16]. Nur ca. 1 % der Fahrer bremst mit einer Verzögerung von $8-10\,\mathrm{m/s^2}$. Etwa 22 % bremsen mit $6-8\,\mathrm{m/s^2}$, etwa 16 % mit $4-6\,\mathrm{m/s^2}$, etwa 10 % mit $2-4\,\mathrm{m/s^2}$ und der Großteil von 51 % mit weniger als $2\,\mathrm{m/s^2}$. Dies wird in Abb. 1.8 visualisiert.

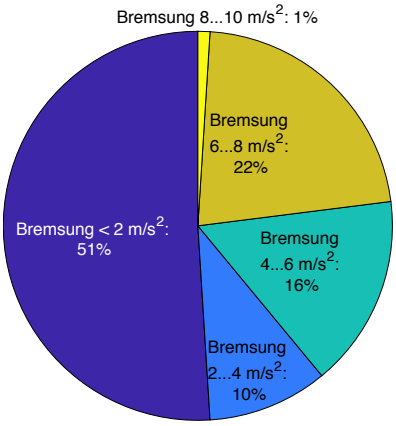


Abbildung 1.8 Bremsverhalten bei Unfällen.

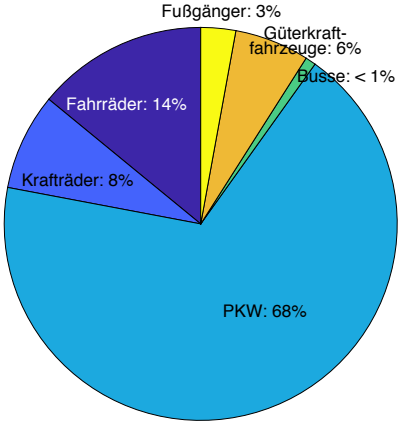


Abbildung 1.9 Hauptverursacher von Unfällen mit Personenschaden 2018.

Das Potenzial in der aktiven Sicherheit durch die Umsetzung unfallvermeidender Maßnahmen in PKW lässt sich zusätzlich auch daran erkennen, dass die Hauptverursacher von Unfällen mit Personenschaden PKW (68%) sind. Eine Visualisierung der Statistik zu den Hauptverursachern von Unfällen mit Personenschaden in Deutschland für das Jahr 2018 ist in Abb. 1.9 dargestellt. In [FKB+15] wird vorgestellt, dass allein durch die autonome Notbremsung für geringe Geschwindigkeiten die Anzahl der Heckaufprallunfälle um 38 % verringert wird. Zu einer ähnlichen Abschätzung kommt auch das Insurance Institute for Highway Safety (IIHS) in den USA und zwar zu einer Reduktion der Heckaufprall-Unfälle um 40 % durch die Einführung der automatischen Notbremsung [Nat17]. Auch das Potenzial, das sich bei autonomen Eingriffen in die Querdynamik ergibt, lässt sich in der Veröffentlichung der National Highway Traffic Safety Administration (NHTSA) ablesen: Nach Einführung der „Autosteer"-Funktionalität in Tesla-Fahrzeugen ist die Anzahl der Unfälle, bei der es zu einer Auslösung von Airbags in diesen Fahrzeugen kam, um fast 40 % gesunken [Nat17].

In der sogenannten **Vision Zero** [BTV12], zu der sich auch die Europäische Union im EU-Weißbuch Verkehr [Eur11] bekannt hat, wird das Ziel formuliert die Straßen und Verkehrsmittel so sicher zu gestalten, dass trotz menschlicher Fehler, diese nicht zu lebensbedrohlichen Verletzungen führen dürfen. Die technischen Systeme der aktiven und passiven Sicherheit leisten damit auch einen wesentlichen Beitrag zur Erreichung der Vision Zero. Wichtige Entwicklungen der letzten Jahrzehnte in der Fahrzeugsicherheit sind in Tab. 1.1 zu sehen [Aut08, Wik19].

Tabelle 1.1 Wichtige Entwicklungen in der Fahrzeugsicherheit.

Jahr	Entwicklung in der Fahrzeugsicherheit
1903	Erfindung des Sicherheitsgurts
1904	Einführung von Luftreifen mit Profil
1908	Einführung der ersten Fahrzeugbeleuchtung
1922	Einführung der selbsttragenden Karosserie
1949	Entwicklung der ersten Dummies
1952	Einführung der Knautschzonen
1959	Einführung des Dreipunktegurts
1957	Geschwindigkeitsbegrenzung auf 50 km/h innerorts
1972	Geschwindigkeitsbegrenzung auf 100 km/h auf Landstraßen
1973	Festlegung der 0,8-Promille-Grenze (bis 1998)
	Einführung der Dreipunktegurte auf den Vordersitzen aller neuen PKW
	Einführung der Anschnallpflicht – wenn Gurte vorhanden sind
1974	Festlegung der Richtgeschwindigkeit von 130 km/h auf Bundesautobahnen
1976	Einführung der Helmpflicht für Motorradfahrer
1978	Pflicht von Sicherheitsgurten für Rücksitze in Neuwagen
1980	Vermehrter Einbau von Fahrer-Airbags
1985	Einführung des elektronischen Antiblockiersystems (ABS)
	Vermehrter Einbau von Beifahrer-Airbags
1991	Einführung der ABS-Pflicht für schwere Nutzfahrzeuge
1994	Einführung von Seitenairbag für Fahrer und Beifahrer
1995	Einführung des elektronischen Stabilitätsprogramms (ESP)
1996	Einführung des ersten Bremsassistenten (BAS)
	Gründung von Euro NCAP für die herstellerunabhängige Bewertung von Crashergebnissen
1998	Einführung der 0,5-Promille-Grenze
	Einführung des Abstandsregeltempomaten
2005	Einführung von passiven Fußgängerschutzmaßnahmen
2009	Pflicht von Bremsassistenten für alle neuen PKW-Typen und leichten Nutzfahrzeuge (Bremsunterstützung beim ruckartigen Tritt auf die Bremse)
2011	Verpflichtende Einführung von ESP bei PKW
2014	Vergabe von Punkten durch Euro NCAP für die automatische Notbremsung (PKW)
2015	Neue schwere Nutzfahrzeuge (NFZ) müssen mit vorausschauenden Notbrems- und Spurhalteassistentsystemen ausgerüstet werden.
2016	Vergabe von Punkten durch Euro NCAP für die automatische Notbremsung (Fußgänger)
2018	Alle neuen Automodelle mit EU-Typgenehmigung müssen mit dem automatischen Notrufsystem eCall ausgestattet sein.
	Alle neuen Nutzfahrzeuge ab 3,5 t zulässiges Gesamtgewicht müssen mit einem Notbremssystem ausgerüstet sein.

Jahr	Entwicklung in der Fahrzeugsicherheit
2019	Der europäische Rat beschließt neue Vorschriften zur Fahrzeugsicherheit. Diese sehen vor, dass ab 2022 alle neu konstruierten Kraftfahrzeuge (inklusive Lastkraftwagen, Busse und Lieferwagen) mit den folgenden Sicherheitsmerkmalen ausgestattet sein müssen: intelligenter Geschwindigkeitsassistent; Vorrichtung zum Einbau einer alkoholempfindlichen Wegfahrsperre; Fahrer-Müdigkeitserkennung und-, Aufmerksamkeitswarnsystem; fortgeschrittene Ablenkungserkennung; Notbremslichter; Systeme für die Erkennung beim Rückwärtsfahren; Unfalldatenspeicher; präzise Reifendrucküberwachung. Zusätzlich sind für für PKW und Lieferwagen vorgesehen: Notbremsassistenzsysteme; Spurhalteassistenzsysteme; erweiterte Kopfaufprallschutzbereiche, um potenzielle Verletzungen von ungeschützten Verkehrsteilnehmern wie Fußgängern und Radfahrern zu mindern.

1.3 Schlüssel zur Wertschöpfung: Elektronikkomponenten und Signalverarbeitung

Dieses Unterkapitel soll zum einen hervorheben, dass sich die automobilen Wertschöpfungsketten durch die in Unterkapitel 1.1 genannten „Megatrends“ der Mobilität im Wandel befinden und dabei den Elektronikkomponenten und der Signalverarbeitung in diesen Komponenten eine Schlüsselrolle zukommt. Das Unterkapitel soll zum anderen einen Überblick zur aktuellen Komplexität der Elektronikkomponenten und der Signalverarbeitungsalgorithmen, die zur Umsetzung von Fahrzeugfunktionen der Fahrzeugsicherheit und des automatisierten Fahrens notwendig sind, geben.

Während der Automobilbau bis in die 1980-er Jahre hauptsächlich durch mechanische Systeme geprägt war, hat die Elektronik im Fahrzeug seither einen Aufschwung erlebt. Ein großer Teil der Wertschöpfung im Fahrzeug ist ohne Elektronikkomponenten und der darin implementierten Signalverarbeitung nicht realisierbar. Während sich in den 1950-er Jahren die Elektronikkomponenten auf Lichtmaschine, Batterie, Lampen, Blinker, Zündung und Autoradio beschränkten und 40 Kupferkabel für ein gesamtes Fahrzeug ausreichend waren, hatten Anfang der 2000-er Jahre Oberklassefahrzeuge bereits über 80 **Steuergeräte** und benötigten zur Vernetzung dieser ca. 4 km Kupferkabel. Ein großer Teil der Fahrzeugfunktionen, die den Komfort und die Sicherheit beim Fahren gewährleisten, z. B. Regelung der Aggregate, ESP, Fahrerassistenzsysteme, Schutz durch Airbags etc., sind ohne den Einzug der Steuergeräte ins Automobil nicht denkbar. Steuergeräte sind **eingebettete Systeme**, die im Allgemeinen die Steuerung oder Regelung eines Fahrzeugsubsystems übernehmen. Sie arbeiten nach dem **Eingabe-Verarbeitung-Ausgabe** (EVA)-Prinzip und bestehen aus mindestens einem Microcontroller, Speicher (SRAM, EEPROM, Flash), Eingängen (Stromversorgung, Analog- und Digitaleingänge) und Ausgängen (Treiber für Aktoren, Analog- und Digitalausgänge). Steuergeräte sind im Fahrzeug durch Busse (CAN, FlexRay, LIN, LVDS, MOST etc.) miteinander verbunden, um den Signalfluss zwischen den einzelnen Subsystemen zu gewährleisten. Die Eingabesignale kommen im Allgemeinen entweder von Sensoren oder von anderen Steuergeräten, die Verarbeitung wird auf den Microcontrollern durchgeführt, und die Ausgangssignale gehen entweder direkt an Aktoren, um diese zu steuern, oder zu anderen Steuergeräten. Mit dem Einzug der

Steuergeräte in die Fahrzeuge steigt der Anteil an Software kontinuierlich und damit die Anforderungen an Methoden und Prozesse, die für die Wertschöpfung im Bereich der Mobilität erforderlich sind. Prognosen, die aufzeigen, welche Komplexität im Forschungs- und Entwicklungsaufwand (FuE-Auwand) für softwarebasierte Fahrzeugfunktionen steckt [VDI04, Ric09], bewahrheiten sich, und es ist zu erwarten, dass beim Übergang von der **Intra-** zur **Inter-Kommunikation** dieser FuE-Auwand erneut zunimmt. Zur Verdeutlichung dieser Komplexität sollen im Folgenden anhand von zwei Beispielen für Oberklassenfahrzeuge elektronische Komponenten vorgestellt werden, die notwendig sind, um Fahrarassistenz- und aktive Sicherheitssysteme umzusetzen.

Die im Jahr 2013 auf den Markt gekommene Mercedes S-Klasse hat zur Realisierung von Fahrzeugfunktionen wie *Adaptive Cruise Control with Steering Assist*, *Adaptive Headlights*, *Self-Parking*, *Night Vision*, *Bird's Eye View*, *Pedestrian Detection*, *Collision Avoidance*, *Collision Preparation* und *Lane Keeping and Traffic Monitoring* folgende vorausschauende Sensoren integriert [How13, TP13]: 2 Nahbereichsradare vorne (Reichweite 80 m, Öffnungswinkel 80°), 1 Fernbereichsradar vorne (Reichweite 200 m, Öffnungswinkel 18°) mit Mittelbereichserfassung (Reichweite 60 m, Öffnungswinkel 60°), 2 Nahbereichsradare seitlich hinten (Reichweite 30 m, Öffnungswinkel 80°), 1 Multi-Mode-Radar hinten (Reichweite 30 m, Öffnungswinkel 80° bzw. 80 m und 16°), Stereokamera hinter der Frontscheibe (Reichweite 500 m, davon 50 m 3D-fähig, Öffnungswinkel 45°), 12 Ultraschall-Sensoren (4 vorne, 4 hinten, je 1 links und rechts vorne und hinten) und 4 Kameras als 360°-Kamerasystem (1 vorne in der Kühlermaske, 1 hinten in der Griffmulde, je 1 links und rechts im Seitenspiegelgehäuse; Öffnungswinkel vertikal ca. 130° und horizontal über 180°).

Berücksichtigt man, dass eine einfache Monovideo-Kamera mit 640 × 480 Pixeln und 8 Bit pro Pixel bei einer Frame-Rate von 25 Hz zu einer Datenrate von 60 MBit/s führt, so wird ersichtlich, dass die Verarbeitung der vorausschauenden Sensordaten eine große Herausforderung darstellt, sowohl was die Hardware betrifft als auch die erforderlichen Algorithmen.

Übung 1.8

Aus diesem Grund wurde im Audi A8, der im Jahr 2017 auf den Markt gekommen ist, zur Vereinfachung der Systemarchitektur für automatisierte Fahrfunktionen ein zentrales Fahrerassistenz-Steuergerät eingeführt. Dieses errechnet aus den Sensordaten ständig ein umfassendes Abbild der Umgebung für eine Vielzahl an Assistenzfunktionen. Es besteht aus vier leistungsfähigen Prozessoren [Aud17b]: Nvidia Tegra K1, Mobilye EyeQ3, Altera Cyclone V und Infineon Aurix. Während der Nvidia Tegra K1 Prozessor hauptsächlich für die Verarbeitung der Kameradaten für eine 360°-Sicht zuständig ist, finden die Bildverarbeitung für die Frontkamera im Mobilye EyeQ3-Prozessor, die Sensordatenvorverarbeitung und -fusion auf dem Altera Cyclone V-Prozessor und die Entscheidungsfindung einiger Fahrzeugfunktionen des automatisierten Fahrens auf dem Infineon Aurix-Prozessor statt. Folgende vorausschauenden Sensoren werden im zentralen Fahrerassistenz-Steuergerät zu einem zentralen Umfeldmodell zusammengeführt [Aud17c]: 1 Laserscanner (vorne), 1 Frontkamera (hinter der Frontscheibe), 1 Fernbereichsradar (vorne), 1 Infrarot-Kamera (vorne), 4 Mittelbereichsradare (an den Ecken des Fahrzeugs), 12 Ultraschallsensoren (vorne, hinten und auf den Seiten), 4 Kameras als 360°-Kamerasystem (vorne, hinten und auf den Seiten) und 1 Sensor für die Fahrerverfügbarkeit (im Innenraum). Betrachtet man die Fahrerassistenzsysteme im Audi A8, werden zusätzlich zu den Sensoren folgende Steuergeräte vernetzt [Aud17a]: Safety Computer, ESP-Steuergerät,

Lenkungssteuergerät, Motorsteuergerät, Getriebesteuergerät, Bodycomputer, Gateway, elektronische Fahrwerksplattform, Kombiinstrument und MMI.

Es wird ersichtlich, dass die Umsetzung automatisierter Eingriffe in die Fahrdynamik das Fahrzeug zu einem „Roboterauto“ mit komplexer Sensorik, Signalverarbeitung und Aktoransteuerung verwandelt. Die Algorithmen und Methoden, die in den nächsten Kapiteln vorgestellt werden, sind entsprechend dem EVA-Prinzip dem Schritt „Verarbeitung“ zuzuordnen. Es handelt sich dabei um Verfahren zur Aufbereitung der Sensordaten, zur Schätzung und Klassifikation von Größen, die in Funktionsalgorithmen für die Regelung bzw. Steuerung von Aktoren benötigt werden, sowie um die Entscheidungsfindung in den Funktionsalgorithmen. Abb. 1.10 zeigt eine allgemeine Systemübersicht für Fahrzeugfunktionen des automatisierten

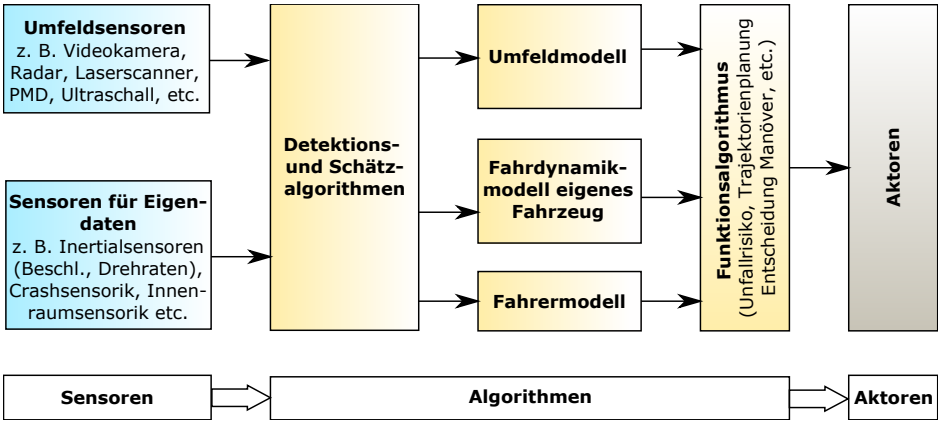


Abbildung 1.10 Systemübersicht.

Fahrens und der integralen Fahrzeugsicherheit. In den einzelnen Modulen des Blockschaltbilds finden unterschiedliche Methoden der Signalverarbeitung Verwendung. Zum einen müssen in der Kategorie „Sensoren“ grundlegende Signalverarbeitungsschritte wie Verstärkung, A/D-Wandlung oder Tiefpassfilterung, zum anderen aber auch komplexe Schritte wie z. B. die Klassifikation von Objekten aus den Rohdaten mittels **maschinellen Lernverfahren** durchgeführt werden. In der Kategorie „Algorithmen“ in Abb. 1.10 finden sich häufig Schätzverfahren, um den Zustand von Objekten im Fahrzeugumfeld aus Beobachtungen über der Zeit zu bestimmen (**Tracking**) oder um den Zustand des eigenen Fahrzeugs sowie des Fahrers zu berechnen. Ein großer Teil der Methoden, die hier Anwendung finden, kann unter dem Begriff **statistische Filterung** zusammengefasst werden. In dem Block „Funktionsalgorithmus“ stecken häufig Signalverarbeitungsalgorithmen, in denen Entscheidungen über das Verhalten des Fahrzeugs gefällt werden. Entsprechend komplex sind hier die Prädiktions- und Schätzalgorithmen, die eine Interpretation der Verkehrssituation erfordern, z. B. um Ausweichmanöver in kritischen Verkehrssituationen zu planen und sich für das richtige Manöver zu entscheiden. Insbesondere bei Fahrzeugfunktionen, die auf vorausschauenden Sensoren basieren und bei denen eine vollständige Spezifikation aufgrund der Komplexität möglicher Objekte und Ereignisse im Fahrzeugumfeld nicht mehr möglich ist, nehmen datenbasierte Verfahren, also das maschinelle Lernen, verstärkt an Bedeutung zu. Das maschinelle Lernen wird nicht nur für die Klassifikation von Objekten aus den Sensorrohdaten, sondern auch für den Schritt der Entscheidungsfindung in Fahrzeugfunktionsalgorithmen als Schlüsseltechnologie gesehen, um es Fahrzeugen zu ermöglichen, in hochkomplexen Situationen angemessen zu reagieren.

Übung 1.9

1.4 Übungen und Lösungen zu Kapitel 1

Übungen

Übung 1.1

Zählen Sie vier Vorteile auf, die man sich vom autonomen Fahren erhofft.

Übung 1.2

Zu welcher Automatisierungsstufe würden Sie folgende Fahrzeugfunktionen zählen: Tempomat, Adaptive Cruise Control (ACC), Automatische Notbremsung (AEB), Valet Parking und Staupilot?

Übung 1.3

Nennen Sie drei ethische Regeln für den automatisierten und vernetzten Fahrzeugverkehr.

Übung 1.4

In dieser Aufgabe wird eine Unfallsituation betrachtet, bei der angenommen wird, dass es keinen Personenschaden, sondern nur Sachschaden gibt. Ca. 700 ms vor einer drohenden Kollision zweier Fahrzeuge soll das Gesamtrisiko berechnet werden. Für eine spezielle Verkehrssituation erhält man folgende Werte: Wenn die Kollisionspartner, ohne zu bremsen, genauso weiterfahren, kommt es zu einer Kollision mit hohem Schaden für beide Fahrzeuge. Die Wahrscheinlichkeit für dieses Ereignis $\mathscr{H}_1$ beträgt $P(\mathscr{H}_1) = 0.75$ und die Kosten liegen bei 60.000 €. Wenn die Kollisionspartner beide bremsen, kommt es zu einem Schaden von 30.000 €. Die Wahrscheinlichkeit für dieses Ereignis $\mathscr{H}_2$ beträgt $P(\mathscr{H}_2) = 0.2$. Wenn allerdings beide Fahrer bremsen und gleichzeitig ausweichend lenken, kann der Unfall vermieden werden, und der Schaden beträgt 0 €. Die Wahrscheinlichkeit für dieses Ereignis $\mathscr{H}_3$ beträgt $P(\mathscr{H}_3) = 0.05$. Berechnen Sie das Gesamtrisiko.

Übung 1.5

Was versteht man unter dem Begriff „Integrale Fahrzeugsicherheit“?

Übung 1.6

Suchen Sie auf der Webseite *https://destatis.de* die aktuellsten Statistiken zu den Ursachen für Unfälle mit Personenschaden in Deutschland sowie für das Fehlverhalten der Fahrzeugführer und erstellen Sie in Matlab Tortendiagramme entsprechend der Darstellungen in den Abb. 1.5 und Abb. 1.6.

Übung 1.7

Suchen Sie auf der Webseite *https://destatis.de* die aktuellsten Statistiken für die Anzahl der Getöteten im Straßenverkehr sowie die Anzahl der gemeldeten PKW in Deutschland und erstellen Sie in Matlab eine aktualisierte Darstellung von Abb. 1.7.

Übung 1.8

Berechnen Sie die Datenrate in MBit/s, die sich bei einer Farbkamera mit 1280 × 800 Pixeln ergibt, wenn diese pro Farbe (rot, grün, blau) 8 Bit pro Pixel und eine Frame-Rate von 30 Hz hat.

Übung 1.9

Zählen Sie fünf Signalverarbeitungsschritte auf, die für Fahrzeugfunktionen des automatisierten Fahrens und der integralen Fahrzeugsicherheit benötigt werden.

Lösungen

Die Matlab-Skripte zu Übungsaufgaben, bei denen eine Implementierung gefordert wird, sind unter *http://www.fahrzeugsicherheitundautomatisiertesfahren.de* (Kennwort: FSuAF_SVuML!) zu finden.

Lösung zu Übung 1.1
Mehr Komfort (anderen Tätigkeiten beim Fahren nachgehen); mehr Sicherheit; höhere Verkehrseffizienz; Mobilität für Menschen mit eingeschränkten Fahrfähigkeiten.

Lösung zu Übung 1.2
Tempomat: Stufe 0; ACC: Stufe 1; AEB: Stufe 1; Valet Parking: Stufe 4; Staupilot: Stufe 3.

Lösung zu Übung 1.3
In Gefahrensituationen hat der Schutz menschlichen Lebens immer höchste Priorität; jede Qualifizierung von Menschen nach persönlichen Merkmalen ist bei unausweichlichen Unfallsituationen unzulässig; in jeder Fahrsituation muss klar geregelt und erkennbar sein, wer für die Fahraufgabe zuständig ist.

Lösung zu Übung 1.4
Das Gesamtrisiko dieser Situation lässt sich mit Gl. (1.2) berechnen

$$R_{\text{Sys}} = \sum_{\forall i} R(\mathscr{H}_i) = 0.75 \cdot 60.000\,€ + 0.2 \cdot 30.000\,€ + 0.05 \cdot 0 = 51.000\,€.$$

Lösung zu Übung 1.5
Integrale Fahrzeugsicherheit: gesamtheitliche Betrachtung von passiver und aktiver Fahrzeugsicherheit mit dem Ziel, den Schutz aller Verkehrsteilnehmer zu steigern.

Lösung zu Übung 1.6
Das Matlab-Skript zu dieser Übung ist unter
http://www.fahrzeugsicherheitundautomatisiertesfahren.de
zu finden.

Lösung zu Übung 1.7
Das Matlab-Skript zu dieser Übung ist unter
http://www.fahrzeugsicherheitundautomatisiertesfahren.de
zu finden.

Lösung zu Übung 1.8
Die Datenrate für diese Kamera beträgt

$$30\,\frac{1}{\text{s}} \cdot 3 \cdot 8\,\text{Bit} \cdot 1280 \cdot 800 = 737{,}28\,\text{MBit/s}.$$

Lösung zu Übung 1.9
Tiefpassfilterung; A/D-Wandlung; Objektklassifikation mittels maschinellem Lernen; Schätzer mittels statistischer Filterung, z. B. Tracking; Entscheidungsfindung mittels maschinellem Lernen.

2 Grundlagen der Signalverarbeitung

Unter **Signalverarbeitung** versteht man das Gebiet der Ingenieurwissenschaften und der angewandten Mathematik, das sich mit der Verarbeitung und Analyse von Signalen beschäftigt. Signale stellen im Allgemeinen eine physikalische Größe dar, sollen aber in diesem Buch abstrakt als Funktionen von unabhängigen Variablen (z. B. Zeit oder Ort) verstanden werden. Gebiete der Mathematik, die in der Signalverarbeitung besonders häufig Anwendung finden, sind: Lineare Algebra, Integral- und Differentialrechnung, Wahrscheinlichkeitstheorie und stochastische Prozesse, Statistik (Schätz- und Entscheidungstheorie), Optimierung sowie die Numerik.

In diesem Kapitel soll eine kompakte Zusammenfassung einiger „Werkzeuge der Signalverarbeitung“ gegeben werden. Tatsächlich steht nur die Anwendung der Werkzeuge, soweit sie für die folgenden Kapitel benötigt wird, im Vordergrund. Dies bedeutet, dass sich die im Folgenden behandelten mathematischen Vektorräume ausschließlich auf endlichdimensionale Vektorräume mit einer gegebenen Metrik, Norm und Innenprodukt beschränken und daher keine Hinweise auf die in der Signalverarbeitung im Allgemeinen unerlässlichen Aspekte der Funktionalanalysis erfolgen, dass im Abschnitt über die Optimierung mittels Lagrange-Multiplikatoren auf fortgeschrittene Themen der konvexen Analysis verzichtet und auf Aspekte der Dualitätstheorie nicht eingegangen wird und dass lediglich die wichtigsten Grundlagen der Wahrscheinlichkeitstheorie und stochastischer Prozesse vermittelt werden.

Lernziele in Kapitel 2

Der Lernende …

- frischt sein Grundlagenwissen zur linearen Algebra auf;
- versteht das Lösen von Optimierungsaufgaben mit Nebenbedingungen mittels Lagrange-Multiplikatoren und kann dieses Wissen anwenden;
- frischt sein Grundlagenwissen zur Wahrscheinlichkeitstheorie auf;
- frischt sein Grundlagenwissen zu linearen Systemen auf;
- kann die zeitkontinuierliche Zustandsdarstellung von linearen zeitinvarianten Systemen diskretisieren;
- frischt sein Grundlagenwissen zur Filterung von Signalen im Frequenzbereich auf;
- kann das aufgefrischte und erworbene Grundlagenwissen zum Lösen von Aufgaben nutzen.

2.1 Lineare Algebra

In diesem Unterkapitel werden einige Grundlagen der linearen Algebra kurz wiederholt. Eine detaillierte Einführung in das Gebiet der linearen Algebra kann beispielsweise in [TB97] gefunden werden.

2.1.1 Definitionen und Notation

Die Menge der reellen Zahlen wird mit $\mathbb{R}$, die Menge der natürlichen Zahlen mit $\mathbb{N}$ und die Menge der komplexen Zahlen mit $\mathbb{C}$ bezeichnet.

Eine $M \times N$-**Matrix** $\boldsymbol{A}$ ist ein zweidimensionales Feld

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix} \tag{2.1}$$

und ein N-dimensionaler **Vektor** $\boldsymbol{x}$ ein eindimensionales Feld

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}, \tag{2.2}$$

wobei im Folgenden angenommen wird, dass die Elemente der Matrix $\boldsymbol{A}$ und des Vektors $\boldsymbol{x}$ zur Menge der reellen Zahlen gehören. In diesem Buch werden nur Spaltenvektoren verwendet und Zeilenvektoren mit Hilfe von Spaltenvektoren ausgedrückt. Matrizen werden im gesamten Buch mit fetten Großbuchstaben, Vektoren mit fetten Kleinbuchstaben geschrieben.

Die **Transponierte** der $M \times N$-Matrix $\boldsymbol{A}$ ist die $N \times M$-Matrix $\boldsymbol{A}^{\mathrm{T}}$

$$\boldsymbol{A}^{\mathrm{T}} = \begin{bmatrix} a_{11} & \cdots & a_{M1} \\ \vdots & \ddots & \vdots \\ a_{1N} & \cdots & a_{MN} \end{bmatrix}. \tag{2.3}$$

Eine quadratische Matrix heißt **symmetrisch** bzw. **schiefsymmetrisch**, wenn gilt

$$\boldsymbol{A} = \boldsymbol{A}^{\mathrm{T}} \qquad \text{bzw.} \qquad \boldsymbol{A} = -\boldsymbol{A}^{\mathrm{T}}. \tag{2.4}$$

Die **Determinante** ist eine spezielle Funktion, die einer quadratischen Matrix einen Skalar zuordnet. Häufig wird für die Determinante der Matrix $\boldsymbol{A}$ neben $\det\{\boldsymbol{A}\}$ auch die Notation $|\boldsymbol{A}|$ verwendet. Für eine 2×2-Matrix ist die Determinante

$$\det\{\boldsymbol{A}\} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}. \tag{2.5}$$

Mit dem Entwicklungssatz von Laplace lässt sich die Determinante einer $N \times N$-Matrix nach einer Spalte oder einer Zeile entwickeln:

$$\det\{\boldsymbol{A}\} = \sum_{i=1}^{N} (-1)^{i+j} a_{ij} \det\{\boldsymbol{A}^{(ij)}\} \quad \text{bzw.} \quad \det\{\boldsymbol{A}\} = \sum_{j=1}^{N} (-1)^{i+j} a_{ij} \det\{\boldsymbol{A}^{(ij)}\}, \tag{2.6}$$

wobei $\boldsymbol{A}^{(ij)}$ die $(N-1)\times(N-1)$ Untermatrix von $\boldsymbol{A}$ ist, die durch Streichen der i-ten Zeile und j-ten Spalte entsteht. Für quadratische Matrizen mit reellen Koeffizienten lässt sich auch eine geometrische Interpretation für die Determinante finden. Der Betrag der Determinanten einer Matrix $\boldsymbol{A}$ gibt den Skalierungsfaktor an, mit dem die Fläche oder das Volumen durch den Einfluss der linearen Transformation $\boldsymbol{A}$ multipliziert wird, und das Vorzeichen der Determinanten gibt an, ob die Orientierung beibehalten wird. Zum Beispiel bildet eine zweidimensionale Matrix mit der Determinanten -2 einen Ursprungsraum mit endlicher Fläche auf einen Raum ab, der doppelt so groß ist, und ändert dessen Orientierungsrichtung.

Die **Inverse** einer quadratischen $N\times N$-Matrix $\boldsymbol{A}$ ist die $N\times N$-Matrix $\boldsymbol{A}^{-1}$, so dass gilt

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}_N, \tag{2.7}$$

wobei $\boldsymbol{I}_N$ die $N\times N$-dimensionale Einheitsmatrix (Einsen in der Diagonalen, sonst Nullen) bezeichnet. Die Inverse einer quadratischen Matrix existiert nur, wenn ihre Spalten $\boldsymbol{a}_i$, $i = 1,\ldots,N$ (oder Zeilen) **linear unabhängig** sind, d. h. aus

$$\sum_{i=1}^{N} \alpha_i \boldsymbol{a}_i = \boldsymbol{0} \quad \text{folgt} \quad \alpha_i = 0, \text{ mit } i = 1,\ldots,N. \tag{2.8}$$

Die Inverse einer quadratischen Matrix $\boldsymbol{A}$ erhält man z. B. mittels „Gaußschem Eliminationsverfahren“ bzw. ganz allgemein mit der Formel

$$\boldsymbol{A}^{-1} = \frac{1}{\det\{\boldsymbol{A}\}}\operatorname{adj}\{\boldsymbol{A}\}, \tag{2.9}$$

wobei die Matrix $\operatorname{adj}\{\boldsymbol{A}\}$ aus Kofaktoren von $\boldsymbol{A}$ besteht und $\det\{\boldsymbol{A}\}$ die Determinante von $\boldsymbol{A}$ ist. Der (i,j)-te Kofaktor von $\boldsymbol{A}$, d. h. der Eintrag in der i-ten Zeile und j-ten Spalte von $\operatorname{adj}\{\boldsymbol{A}\}$, ist die Determinante von $\boldsymbol{A}^{(ij)}$, also der Matrix $\boldsymbol{A}$ ohne ihre i-te Zeile und j-te Spalte, multipliziert mit $(-1)^{i+j}$.

Lineare Unabhängigkeit, wie in Gl. (2.8) beschrieben, ist äquivalent zu der Aussage, dass die Determinante von $\boldsymbol{A}$ nicht null ist. Eine invertierbare Matrix wird auch **nichtsinguläre** oder **reguläre** Matrix genannt. Hat eine quadratische Matrix keine Inverse, so heißt sie **singulär**.

Der **Rang** einer Matrix ist die Anzahl an linear unabhängigen Zeilen bzw. Spalten.

Weil eine $M\times N$-Matrix $\boldsymbol{A}$ eine lineare Abbildung aus $\mathbb{R}^N$ nach $\mathbb{R}^M$ darstellt, bezeichnet man mit dem **Bildbereich** der Matrix $\boldsymbol{A}$ den Unterraum von $\mathbb{R}^M$, der durch die linear unabhängigen Spalten von $\boldsymbol{A}$ aufgespannt wird. Der Untervektorraum von $\mathbb{R}^N$, der auf den Nullvektor in $\mathbb{R}^M$ abgebildet wird, heißt **Kern** der Matrix $\boldsymbol{A}$.

Das **Innenprodukt** oder **Skalarprodukt** zweier N-dimensionaler Vektoren $\boldsymbol{x}$ und $\boldsymbol{y}$ in einem Euklidischen Raum ist

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \sum_{i=1}^{N} x_i y_i. \tag{2.10}$$

Die beiden Vektoren $\boldsymbol{x}$ und $\boldsymbol{y}$ heißen **orthogonal**, wenn ihr Innenprodukt den Wert 0 hat, d. h.

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = 0. \tag{2.11}$$

Die quadrierte 2-**Norm** oder $\mathscr{L}_2$-**Norm** des Vektors $\boldsymbol{x}$ ist das Innenprodukt von $\boldsymbol{x}$ mit sich selber

$$\langle \boldsymbol{x}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|_2^2. \tag{2.12}$$

Die $\mathscr{L}_2$-Norm $\|\boldsymbol{x}\|_2$ stellt die Länge des Vektors $\boldsymbol{x}$ dar und wird auch **Euklidische Norm** des Vektors $\boldsymbol{x}$ genannt. Ein Vektor wird als **normiert** bezeichnet, wenn er die Länge 1 hat, d. h. $\|\boldsymbol{x}\|_2 = 1$.

Übung 2.1

Die beiden Vektoren $\boldsymbol{x}$ und $\boldsymbol{y}$ heißen **orthonormal**, wenn sie orthogonal und jeweils normiert sind.

Die **gewichtete Euklidische Norm** des Vektors $\boldsymbol{x}$ erhält man mittels einer symmetrischen positiv-definiten Matrix $\boldsymbol{W}$

$$\|\boldsymbol{x}\|_{\boldsymbol{W}} = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x}}. \tag{2.13}$$

Der **Winkel** $0 \leq \theta \leq \pi$ zwischen zwei N-dimensionalen Vektoren $\boldsymbol{x}$ und $\boldsymbol{y}$ ist definiert durch

$$\cos(\theta) = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|_2 \left\|\boldsymbol{y}\right\|_2}. \tag{2.14}$$

Man sieht mit Hilfe von Gl. (2.11), dass für orthogonale Vektoren $\cos(\theta) = 0$ gilt.

Das **dyadische Produkt** zweier Vektoren $\boldsymbol{x}, \boldsymbol{y}$ ist die Matrix

$$\boldsymbol{C} = \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \tag{2.15}$$

und das **Kreuzprodukt** der Vektor

$$\boldsymbol{c} = \boldsymbol{x} \times \boldsymbol{y} = [x_2 y_3 - x_3 y_2,\, x_3 y_1 - x_1 y_3,\, x_1 y_2 - x_2 y_1]^{\mathrm{T}}. \tag{2.16}$$

Die **Spur** (engl. trace) einer $N \times N$-Matrix ist die Summe ihrer Diagonalelemente

$$\mathrm{tr}\{\boldsymbol{A}\} = \sum_{i=1}^{N} a_{ii} = \mathrm{tr}\left\{\boldsymbol{A}^{\mathrm{T}}\right\}. \tag{2.17}$$

Eine quadratische Matrix $\boldsymbol{A}$ heißt **nilpotent**, wenn eine ihrer Potenzen die Nullmatrix ergibt

$$\boldsymbol{A}^n = \boldsymbol{0} \quad \text{für ein } n \in \mathbb{N}. \tag{2.18}$$

Die **quadratische Wurzel** einer quadratischen Matrix $\boldsymbol{A}$ ist die im Allgemeinen nicht eindeutige Matrix $\boldsymbol{A}^{\frac{1}{2}}$, so dass gilt

$$\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{A}. \tag{2.19}$$

Eine **quadratische Form** q des Vektors $\boldsymbol{x}$ ist definiert als die skalare Funktion

$$q = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} x_i x_j, \tag{2.20}$$

wobei gelten muss, dass $a_{ij} = a_{ji}$. Die quadratische Form kann auch als $q = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$ ausgedrückt werden, wobei $\boldsymbol{A}$ eine symmetrische $N \times N$-Matrix ist und $\boldsymbol{x} = [x_1, \ldots, x_N]^{\mathrm{T}}$. Die Matrix $\boldsymbol{A}$ ist **positiv semidefinit**, wenn

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} \geq 0, \tag{2.21}$$

für alle $\boldsymbol{x} \neq \boldsymbol{0}$. Gilt das Größer-Zeichen statt des Größer-Gleich-Zeichens, dann ist $\boldsymbol{A}$ positiv definit.

Eine quadratische Marix $\boldsymbol{A}$ ist genau dann positiv definit, wenn ihr symmetrischer Teil

$$\boldsymbol{A}_{\mathrm{S}} = \frac{1}{2}\left(\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}\right) \tag{2.22}$$

positiv definit ist. Entsprechendes gilt für die Begriffe „positiv semidefinit“, „negativ definit“ und „negativ semidefinit“. Eine quadratische, symmetrische Matrix $\boldsymbol{A}$ ist genau dann positiv definit, wenn alle Eigenwerte größer als null sind. Sie ist genau dann positiv semidefinit, wenn ihre Eigenwerte größer oder gleich null sind; negativ definit, wenn alle Eigenwerte kleiner als null sind; und negativ semidefinit, wenn alle Eigenwerte kleiner oder gleich null sind. Sind die Eigenwerte sowohl positiv als auch negativ, ist die Matrix **indefinit**.

Eine **Diagonalmatrix** ist eine quadratische Matrix mit $a_{ij} = 0$, für $i \neq j$

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{NN} \end{bmatrix}. \tag{2.23}$$

Die Inverse einer Diagonalmatrix erhält man, indem man die einzelnen Elemente der Diagonalen invertiert.

Eine Verallgemeinerung der Diagonalmatrix ist die **Blockdiagonalmatrix**

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{A}_{kk} \end{bmatrix}, \tag{2.24}$$

in der alle $\boldsymbol{A}_{ii}$ quadratisch sind, ihre Dimension aber nicht gleich sein muss. Wenn alle $\boldsymbol{A}_{ii}$ nichtsingulär sind, dann ist die Inverse einer Blockdiagonalmatrix

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11}^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{A}_{kk}^{-1} \end{bmatrix} \tag{2.25}$$

und die Determinante

$$\det\{\boldsymbol{A}\} = \prod_{i=1}^{k} \det\{\boldsymbol{A}_{ii}\}. \tag{2.26}$$

Eine quadratische Matrix $\boldsymbol{A}$ heißt **orthogonal**, wenn

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^{\mathrm{T}}. \tag{2.27}$$

Wenn eine Matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]$ orthogonal ist, so sind ihre Spalten (und Zeilen) **orthonormal**, d. h.

$$\left\langle \boldsymbol{a}_i, \boldsymbol{a}_j \right\rangle = \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{a}_j = \begin{cases} 0 & \text{wenn } i \neq j \\ 1 & \text{wenn } i = j. \end{cases} \tag{2.28}$$

2.1.2 Einige Rechenregeln der linearen Algebra

Die Addition zweier Matrizen oder die Multiplikation mit einem Skalar wird elementweise durchgeführt, d. h. in der Gleichung

$$\boldsymbol{C} = \alpha \boldsymbol{A} + \beta \boldsymbol{B} \tag{2.29}$$

mit den drei $M \times N$-Matrizen $\boldsymbol{A}$, $\boldsymbol{B}$ und $\boldsymbol{C}$ berechnet sich der Eintrag in der i-ten Zeile und j-ten Spalte in $\boldsymbol{C}$ als

$$c_{ij} = \alpha a_{ij} + \beta b_{ij}. \tag{2.30}$$

Das Produkt der $M \times N$-Matrix $\boldsymbol{A}$ mit der $N \times P$-Matrix $\boldsymbol{B}$ ergibt die $M \times P$-Matrix $\boldsymbol{C}$, dessen Eintrag in der i-ten Zeile und j-ten Spalte sich berechnen lässt als

$$c_{ij} = \sum_{k=1}^{N} a_{ik} b_{kj}. \tag{2.31}$$

Es ist zu beachten, dass die **Matrixmultiplikation nicht kommutativ** ist.

Das **Kronecker-Produkt** zwischen der $M \times N$-Matrix $\boldsymbol{A}$ mit der $Q \times P$-Matrix $\boldsymbol{B}$ ergibt die $MQ \times NP$-Matrix $\boldsymbol{C}$

$$\boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1N}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2N}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1}\boldsymbol{B} & a_{M2}\boldsymbol{B} & \cdots & a_{MN}\boldsymbol{B} \end{bmatrix}. \tag{2.32}$$

Einige nützliche Rechenregeln im Umgang mit den $N \times N$-Matrizen $\boldsymbol{A}$ und $\boldsymbol{B}$ sind

$$(\boldsymbol{AB})^{\mathrm{T}} = \boldsymbol{B}^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \tag{2.33}$$

$$\left(\boldsymbol{A}^{\mathrm{T}}\right)^{-1} = \left(\boldsymbol{A}^{-1}\right)^{\mathrm{T}} \tag{2.34}$$

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1} \boldsymbol{A}^{-1} \tag{2.35}$$

$$\det\left\{\boldsymbol{A}^{\mathrm{T}}\right\} = \det\{\boldsymbol{A}\} \tag{2.36}$$

$$\det\{\alpha \boldsymbol{A}\} = \alpha^{N} \det\{\boldsymbol{A}\} \tag{2.37}$$

$$\det\{\boldsymbol{AB}\} = \det\{\boldsymbol{A}\}\det\{\boldsymbol{B}\} \tag{2.38}$$

$$\det\left\{\boldsymbol{A}^{-1}\right\} = \frac{1}{\det\{\boldsymbol{A}\}} \tag{2.39}$$

$$\mathrm{tr}\{\boldsymbol{AB}\} = \mathrm{tr}\{\boldsymbol{BA}\}. \tag{2.40}$$

Mit Matrizen passender Dimension gilt

$$\mathrm{tr}\{\boldsymbol{ABC}\} = \mathrm{tr}\{\boldsymbol{BCA}\} = \mathrm{tr}\{\boldsymbol{CAB}\}, \tag{2.41}$$

d. h. der Spuroperator ist invariant gegenüber zyklischen Permutationen seiner Argumente.

Die Inverse einer $N \times N$-**partitionierten Matrix** lässt sich durch die Inversen der einzelnen Partitionen ausdrücken. Wenn $\boldsymbol{P}_{11}$ eine $N_1 \times N_1$-, $\boldsymbol{P}_{12}$ eine $N_1 \times N_2$-, $\boldsymbol{P}_{21}$ eine $N_2 \times N_1$- und $\boldsymbol{P}_{22}$ eine $N_2 \times N_2$-Matrix sind, mit $N_1 + N_2 = N$, so erhält man

$$\begin{bmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}, \tag{2.42}$$

mit

$$\boldsymbol{V}_{11} = \boldsymbol{P}_{11}^{-1} + \boldsymbol{P}_{11}^{-1}\boldsymbol{P}_{12}\boldsymbol{V}_{22}\boldsymbol{P}_{21}\boldsymbol{P}_{11}^{-1} = (\boldsymbol{P}_{11} - \boldsymbol{P}_{12}\boldsymbol{P}_{22}^{-1}\boldsymbol{P}_{21})^{-1} \tag{2.43}$$

$$\boldsymbol{V}_{12} = -\boldsymbol{P}_{11}^{-1}\boldsymbol{P}_{12}\boldsymbol{V}_{22} = -\boldsymbol{V}_{11}\boldsymbol{P}_{12}\boldsymbol{P}_{22}^{-1} \tag{2.44}$$

$$\boldsymbol{V}_{21} = -\boldsymbol{V}_{22}\boldsymbol{P}_{21}\boldsymbol{P}_{11}^{-1} = -\boldsymbol{P}_{22}^{-1}\boldsymbol{P}_{21}\boldsymbol{V}_{11} \tag{2.45}$$

$$\boldsymbol{V}_{22} = \boldsymbol{P}_{22}^{-1} + \boldsymbol{P}_{22}^{-1}\boldsymbol{P}_{21}\boldsymbol{V}_{11}\boldsymbol{P}_{12}\boldsymbol{P}_{22}^{-1} = (\boldsymbol{P}_{22} - \boldsymbol{P}_{21}\boldsymbol{P}_{11}^{-1})^{-1}\boldsymbol{P}_{12}. \tag{2.46}$$

Das **Matrix-Inversionslemma** besagt, dass

$$\left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}\boldsymbol{B}^{\mathrm{T}}\right)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{B} + \boldsymbol{C}^{-1}\right)\boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{-1}. \tag{2.47}$$

Es kann leicht gezeigt werden, dass sowohl das Matrix-Inversionslemma (2.47) als auch Gl. (2.43) bis Gl. (2.46) gelten, indem man durch Multiplikation die Einheitsmatrix erhält. Gl. (2.43) bis Gl. (2.46) werden bei der Herleitung der Wahrscheinlichkeitsdichten von bedingten Gaußschen Zufallsvariablen in Unterkapitel 2.3.4 und das Matrix-Inversionslemma in Unterkapitel 4.1 bei der Herleitung des Kalman-Filters benötigt.

Das **Matrix-Determinantenlemma** für die invertierbare $N \times N$-Matrix $\boldsymbol{A}$, die invertierbare $M \times M$-Matrix $\boldsymbol{B}$ und die beiden $N \times M$-Matrizen $\boldsymbol{C}$ und $\boldsymbol{D}$ lautet:

$$\det\left\{\boldsymbol{C}\boldsymbol{B}\boldsymbol{D}^{\mathrm{T}} + \boldsymbol{A}\right\} = \det\left\{\boldsymbol{D}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{C} + \boldsymbol{B}^{-1}\right\}\det\{\boldsymbol{A}\}\det\{\boldsymbol{B}\}. \tag{2.48}$$

Die **Cauchy-Schwarz-Ungleichung** macht eine Aussage über das Innenprodukt zweier Vektoren und deren $\mathscr{L}_2$-Norm

$$\left|\langle \boldsymbol{x}, \boldsymbol{y}\rangle\right| \leq \|\boldsymbol{x}\|_2 \left\|\boldsymbol{y}\right\|_2. \tag{2.49}$$

Die **Dreiecksungleichung für Vektoren** macht eine Aussage zu der Norm der Summe zweier Vektoren

$$\left\|\boldsymbol{x} + \boldsymbol{y}\right\|_2 \leq \|\boldsymbol{x}\|_2 + \left\|\boldsymbol{y}\right\|_2. \tag{2.50}$$

Eine Abbildung $d : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ heißt **Metrik** auf $\mathbb{R}^N$, wenn für beliebige Vektoren $\boldsymbol{x}$, $\boldsymbol{y}$ und $\boldsymbol{z}$ aus $\mathbb{R}^N$ folgende Bedingungen gelten

$$d\left(\boldsymbol{x}, \boldsymbol{y}\right) = d\left(\boldsymbol{y}, \boldsymbol{x}\right); \tag{2.51}$$

$$d\left(\boldsymbol{x}, \boldsymbol{y}\right) \geq 0; \tag{2.52}$$

$$d\left(\boldsymbol{x}, \boldsymbol{y}\right) = 0 \quad \text{nur wenn } \boldsymbol{x} = \boldsymbol{y}; \tag{2.53}$$

$$d\left(\boldsymbol{x}, \boldsymbol{y}\right) + d\left(\boldsymbol{y}, \boldsymbol{z}\right) \geq d\left(\boldsymbol{x}, \boldsymbol{z}\right). \tag{2.54}$$

Die letzte Bedingung ist die Dreiecksungleichung für metrische Räume.

Die **orthogonale Projektion** von $\boldsymbol{x}$ auf $\boldsymbol{y}$ berechnet sich als

$$\Pi_{\boldsymbol{y}}(\boldsymbol{x}) = \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\left\|\boldsymbol{y}\right\|_2^2}\boldsymbol{y}. \tag{2.55}$$

Entsprechend ist der Vektor $\boldsymbol{x} - \Pi_{\boldsymbol{y}}(\boldsymbol{x})$ orthogonal zu $\boldsymbol{y}$

$$\left\langle \left(\boldsymbol{x} - \Pi_{\boldsymbol{y}}(\boldsymbol{x})\right), \boldsymbol{y}\right\rangle = 0. \tag{2.56}$$

Die orthogonale Projektion kann z. B. verwendet werden, um das optimale lineare statistische Filter herzuleiten (s. Übung 5.4).

Ein Operator, der für die Herleitung einiger maschineller Lernverfahren verwendet wird, ist der **vec-Operator**. Es handelt sich um einen Operator, der aus einer $M \times N$-Matrix $\boldsymbol{A}$ einen Vektor der Dimension MN macht, indem die Spalten der Matrix untereinander angeordnet werden

$$\boldsymbol{a} = \text{vec}\{\boldsymbol{A}\} = [a_{11}, a_{21}, \ldots, a_{M1}, a_{12}, a_{22}, \ldots, a_{M2}, \ldots, a_{1N}, a_{2N}, a_{MN}]^{\mathrm{T}}. \tag{2.57}$$

Betrachtet man die Verallgemeinerung von Skalaren, Vektoren und Matrizen auf **Tensoren**, so kann der vec-Operator auch für Tensoren eingeführt werden. Tensoren höherer Ordnung können, im Gegensatz zu Matrizen, mehr als einen Vektor als Argument haben und sie weisen für jeden dieser Vektoren eine lineare Abbildung auf. Betrachtet man zum Beispiel den Tensor $\overset{\circ}{\boldsymbol{T}} \in \mathbb{R}^{K \times M \times N}$, so erhält man durch den vec-Operator den Spaltenvektor

$$\boldsymbol{t} = \text{vec}\left\{\overset{\circ}{\boldsymbol{T}}\right\} = [t_{111}, t_{211}, \ldots, t_{K11}, t_{121}, \ldots, t_{KMN}]^{\mathrm{T}} \in \mathbb{R}^{KMN}. \tag{2.58}$$

2.1.3 Ableiten nach Vektoren und Matrizen

Der **Gradient**- oder **Nabla**-Operator für einen N-dimensionalen Vektor $\boldsymbol{x}$ ist definiert als

$$\nabla_{\boldsymbol{x}} = \left[\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_N}\right]^{\mathrm{T}}. \tag{2.59}$$

Der Gradient einer M-dimensionalen vektorwertigen Funktion $\boldsymbol{f}(\boldsymbol{x})$ ist

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}} = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_N} \end{bmatrix} \left[f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})\right] = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_M(\boldsymbol{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_N} & \cdots & \frac{\partial f_M(\boldsymbol{x})}{\partial x_N} \end{bmatrix}. \tag{2.60}$$

Die **Jacobi**-Matrix ist die Transponierte dieser Matrix, d. h.

$$\boldsymbol{J} = \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}} = \left(\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)^{\mathrm{T}} = \left(\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\right)^{\mathrm{T}} = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_M(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_M(\boldsymbol{x})}{\partial x_N} \end{bmatrix}. \tag{2.61}$$

Die **Hesse**-Matrix einer skalarwertigen Funktion $g(\boldsymbol{x})$ ist

$$\boldsymbol{H} = \frac{\partial^2 g(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} = \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{x}}^{\mathrm{T}} g(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1 \partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g(\boldsymbol{x})}{\partial x_N \partial x_1} & \cdots & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_N \partial x_N} \end{bmatrix}. \tag{2.62}$$

Die Hesse-Matrix ist eine symmetrische $N \times N$-Matrix.

Die **Taylor-Reihenentwicklung** der skalarwertigen Funktion $g(\boldsymbol{x})$ um den Punkt $\boldsymbol{x}_0$ lautet:

$$g(\boldsymbol{x}) = g(\boldsymbol{x}_0) + \underbrace{\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}}\bigg|_{\boldsymbol{x}=\boldsymbol{x}_0}}_{\boldsymbol{J}} (\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2!} (\boldsymbol{x} - \boldsymbol{x}_0)^{\mathrm{T}} \underbrace{\frac{\partial^2 g(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}}\bigg|_{\boldsymbol{x}=\boldsymbol{x}_0}}_{\boldsymbol{H}} (\boldsymbol{x} - \boldsymbol{x}_0) + O\left(\|\boldsymbol{x} - \boldsymbol{x}_0\|^3\right), \tag{2.63}$$

wobei das Landau-Symbol $O(\cdot)$ verwendet wurde, um zu zeigen, dass das Restglied nicht wesentlich schneller wächst als $\|\boldsymbol{x}-\boldsymbol{x}_0\|^3$.

Durch komponentenweises Ableiten erhält man folgende Regeln für das Ableiten nach Vektoren

$$\frac{\partial(\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{A} \tag{2.64}$$

$$\frac{\partial\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}\right)}{\partial \boldsymbol{x}} = \frac{\partial\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}\right)}{\partial \boldsymbol{x}} = \boldsymbol{y} \tag{2.65}$$

$$\frac{\partial\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}} = \left(\boldsymbol{A}+\boldsymbol{A}^{\mathrm{T}}\right)\boldsymbol{x} \quad \text{bzw. falls } \boldsymbol{A} \text{ symmetrisch ist:} \quad \frac{\partial\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}. \tag{2.66}$$

Betrachtet man die M-dimensionalen vektorwertigen Funktionen $\boldsymbol{f}(\boldsymbol{x})$ und $\boldsymbol{h}(\boldsymbol{x})$ sowie eine $M \times M$-symmetrische Matrix $\boldsymbol{A}$, gilt für die Ableitung nach dem N-dimensionalen Vektor $\boldsymbol{x}$

$$\frac{\partial\left(\boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{x})\right)}{\partial \boldsymbol{x}} = \left(\nabla_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\right)\boldsymbol{h}(\boldsymbol{x}) + \left(\nabla_{\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}}\right)\boldsymbol{f}(\boldsymbol{x}) \tag{2.67}$$

$$\frac{\partial\left(\boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{f}(\boldsymbol{x})\right)}{\partial \boldsymbol{x}} = 2\left(\nabla_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\right)\boldsymbol{A}\boldsymbol{f}(\boldsymbol{x}). \tag{2.68}$$

Der Gradient-Operator einer skalarwertigen Funktion der $M \times N$-Matrix $\boldsymbol{A}$ ist die $M \times N$-Matrix

$$\frac{\partial g(\boldsymbol{A})}{\partial \boldsymbol{A}} = \begin{bmatrix} \frac{\partial g(\boldsymbol{A})}{\partial a_{11}} & \cdots & \frac{\partial g(\boldsymbol{A})}{\partial a_{1N}} \\ \vdots & & \vdots \\ \frac{\partial g(\boldsymbol{A})}{\partial a_{M1}} & \cdots & \frac{\partial g(\boldsymbol{A})}{\partial a_{MN}} \end{bmatrix}. \tag{2.69}$$

Leitet man die Spur der Multiplikation der $M \times N$-Matrix $\boldsymbol{A}$ mit der $N \times P$-Matrix $\boldsymbol{B}$ nach $\boldsymbol{A}$ ab erhält man

$$\frac{\partial \operatorname{tr}\{\boldsymbol{A}\boldsymbol{B}\}}{\partial \boldsymbol{A}} = \boldsymbol{B}^{\mathrm{T}}. \tag{2.70}$$

Betrachtet man eine Funktion $\overset{\circ}{\boldsymbol{f}}$, bei der sowohl die Eingänge $\overset{\circ}{\boldsymbol{X}} \in \mathbb{R}^{K\times M\times N}$ als auch die Ausgänge $\overset{\circ}{\boldsymbol{Y}} = \overset{\circ}{\boldsymbol{f}}\left(\overset{\circ}{\boldsymbol{X}}\right) \in \mathbb{R}^{L\times P\times Q}$ Tensoren sind, so kann man die Komponenten der ersten Ableitung dieser Funktion mit Hilfe des vec-Operators und der Jacobi-Matrix ausdrücken

$$\frac{\partial \operatorname{vec}\left\{\overset{\circ}{\boldsymbol{f}}\left(\overset{\circ}{\boldsymbol{X}}\right)\right\}}{\partial\left(\operatorname{vec}\left\{\overset{\circ}{\boldsymbol{X}}\right\}\right)^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial y_{111}}{\partial x_{111}} & \frac{\partial y_{111}}{\partial x_{211}} & \cdots & \frac{\partial y_{111}}{\partial x_{K11}} & \frac{\partial y_{111}}{\partial x_{121}} & \cdots & \frac{\partial y_{111}}{\partial x_{KMN}} \\ \frac{\partial y_{211}}{\partial x_{111}} & \frac{\partial y_{211}}{\partial x_{211}} & \cdots & \frac{\partial y_{211}}{\partial x_{K11}} & \frac{\partial y_{211}}{\partial x_{121}} & \cdots & \frac{\partial y_{211}}{\partial x_{KMN}} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ \frac{\partial y_{L11}}{\partial x_{111}} & \frac{\partial y_{L11}}{\partial x_{211}} & \cdots & \frac{\partial y_{L11}}{\partial x_{K11}} & \frac{\partial y_{L11}}{\partial x_{121}} & \cdots & \frac{\partial y_{L11}}{\partial x_{KMN}} \\ \frac{\partial y_{121}}{\partial x_{111}} & \frac{\partial y_{121}}{\partial x_{211}} & \cdots & \frac{\partial y_{121}}{\partial x_{K11}} & \frac{\partial y_{121}}{\partial x_{121}} & \cdots & \frac{\partial y_{121}}{\partial x_{KMN}} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ \frac{\partial y_{LPQ}}{\partial x_{111}} & \frac{\partial y_{LPQ}}{\partial x_{211}} & \cdots & \frac{\partial y_{LPQ}}{\partial x_{K11}} & \frac{\partial y_{LPQ}}{\partial x_{121}} & \cdots & \frac{\partial y_{LPQ}}{\partial x_{KMN}} \end{bmatrix} \in \mathbb{R}^{LPQ\times KMN}. \tag{2.71}$$

Die Berechnung der ersten Ableitung spielt bei der Lösung von Optimierungsaufgaben und damit bei maschinellen Lernverfahren eine zentrale Rolle. Die Ableitung aus Gl. (2.71) wird z. B. im Lernprozess für neuronale Netzwerke (s. Unterkapitel 5.2) verwendet.

Es soll im Folgenden auch kurz auf eine **geometrische Interpretation** eingegangen werden. Betrachtet man eine Fläche im N-dimensionalen Raum, die **implizit** durch die Gleichung $g(x_1, x_2, \ldots, x_N) = 0$ gegeben ist, so ist der Gradientenvektor $\nabla_{\boldsymbol{x}} g(\boldsymbol{x}) = \left[\frac{\partial g(\boldsymbol{x})}{\partial x_1}, \frac{\partial g(\boldsymbol{x})}{\partial x_2}, \ldots, \frac{\partial g(\boldsymbol{x})}{\partial x_N}\right]^{\mathrm{T}}$ ein Normalenvektor der Fläche im Punkt $[x_1, x_2, \ldots, x_N]^{\mathrm{T}}$. Es handelt sich beim Normalenvektor um den Vektor, der senkrecht zur Tangentialebene der Fläche im Punkt $[x_1, x_2, \ldots, x_N]^{\mathrm{T}}$ steht.

Betrachtet man hingegen die **explizite** Beschreibung der Fläche als $x_N = h(x_1, x_2, \ldots, x_{N-1})$, so ist ein Normalenvektor gegeben durch

$$\boldsymbol{n} = \left[\frac{\partial h(x_1, x_2, \ldots, x_{N-1})}{\partial x_1}, \frac{\partial h(x_1, x_2, \ldots, x_{N-1})}{\partial x_2}, \ldots, \frac{\partial h(x_1, x_2, \ldots, x_{N-1})}{\partial x_{N-1}}, -1\right]^{\mathrm{T}} \in \mathbb{R}^N. \tag{2.72}$$

Dies liegt daran, dass geschrieben werden kann $g(x_1, x_2, \ldots, x_N) = h(x_1, x_2, \ldots, x_{N-1}) - x_N = 0$, was eine implizite Darstellung der Fläche und damit der Gradientenvektor ein Normalenvektor ist. Die Gleichung einer Ebene mit dem Normalenvektor $\boldsymbol{n} = [n_1, n_2, \ldots, n_N]^{\mathrm{T}}$, die durch den Punkt $\boldsymbol{x}_0 = [x_{1,0}, x_{2,0}, \ldots, x_{N,0}]^{\mathrm{T}}$ gehen soll, lautet

$$\boldsymbol{n}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}_0) = n_1\left(x_1 - x_{1,0}\right) + n_2\left(x_2 - x_{2,0}\right) + \ldots + n_N\left(x_N - x_{N,0}\right) = 0. \tag{2.73}$$

Es ist zu beachten, dass im $(N-1)$-dimensionalen Raum der Gradientenvektor $\left[\frac{\partial h(x_1, x_2, \ldots, x_{N-1})}{\partial x_1}, \frac{\partial h(x_1, x_2, \ldots, x_{N-1})}{\partial x_2}, \ldots, \frac{\partial h(x_1, x_2, \ldots, x_{N-1})}{\partial x_{N-1}}\right]^{\mathrm{T}}$ in Richtung des steilsten Anstiegs von $x_N = h(x_1, x_2, \ldots, x_{N-1})$ zeigt.

Übung 2.2

2.1.4 Eigenwert- und Singulärwertzerlegung; Normen von Matrizen

Ein **Eigenvektor** einer $N \times N$ quadratischen Matrix $\boldsymbol{A}$ ist ein $N \times 1$ Vektor $\boldsymbol{v}$, der die Gleichung

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v} \tag{2.74}$$

erfüllt. Dabei ist λ der zum Eigenvektor $\boldsymbol{v}$ dazugehörende **Eigenwert**, der auch komplex sein kann.[1] Es wird angenommen, dass der Eigenvektor auf den Wert 1 normiert ist, d. h. $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1$. Es gelten folgende Ergebnisse:

a) Ist $\boldsymbol{A}$ symmetrisch, so kann man immer N unabhängige Eigenvektoren finden, auch wenn diese im Allgemeinen nicht eindeutig sind. Ein Beispiel ist die Einheitsmatrix, für die jeder Vektor (mit Norm 1) ein Eigenvektor mit dem Eigenwert 1 ist.

b) Ist $\boldsymbol{A}$ reell und symmetrisch, so sind die Eigenvektoren, die zu unterschiedlichen Eigenwerten gehören, orthogonal. Normiert man diese Eigenvektoren, so erhält man orthonormale

[1] Die Verallgemeinerung der Eigenwertzerlegung in der Funktionalanalysis ist das Konzept des Spektrums eines Operators. Jedem linearen Operator $\boldsymbol{A}$ ordnet man in der Funktionalanalysis ein Spektrum zu. Dies besteht aus allen Werten λ, für die der Operator $\boldsymbol{A} - \lambda \boldsymbol{I}$ nicht invertierbar ist. Die Menge aller Eigenwerte wird Spektrum genannt. Die Eigenwertzerlegung wird deswegen gelegentlich auch als Spektralzerlegung bezeichnet.

Eigenvektoren. Man kann dies schreiben als $\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{v}_j = \delta_{ij}$, wobei δ_{ij} das **Kronecker-Delta**-Symbol darstellt und dieses definiert ist als

$$\delta_{ij} = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j. \end{cases} \tag{2.75}$$

c) Ist $\boldsymbol{A}$ symmetrisch und positiv definit (semidefinit), so sind die Eigenwerte positiv (nichtnegativ).

Fasst man die Eigenwerte zur Diagonalmatrix $\boldsymbol{\Lambda}$ und die Eigenvektoren zur Matrix $\boldsymbol{V}$ zusammen, so kann man Gl. (2.74) auch schreiben als

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Lambda}. \tag{2.76}$$

Für die Determinante ergibt sich aus Gl. (2.38), Gl. (2.39) und Gl. (2.76)

$$\det\{\boldsymbol{A}\} = \det\{\boldsymbol{V}\}\det\{\boldsymbol{\Lambda}\}\det\left\{\boldsymbol{V}^{-1}\right\} = \det\{\boldsymbol{\Lambda}\} = \prod_{i=1}^{N}\lambda_i. \tag{2.77}$$

Ist $\boldsymbol{A}$ reell und symmetrisch, so kann $\boldsymbol{V}$ nur orthogonal gewählt werden. Die Inverse von $\boldsymbol{V}$ ist dann die Matrix $\boldsymbol{V}^{\mathrm{T}}$, und man erhält aus Gl. (2.76)

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} = \sum_{i=1}^{N}\lambda_i\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}. \tag{2.78}$$

Die Inverse lässt sich in diesem Fall berechnen als

$$\boldsymbol{A}^{-1} = \boldsymbol{V}^{\mathrm{T},-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^{\mathrm{T}} = \sum_{i=1}^{N}\frac{1}{\lambda_i}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}. \tag{2.79}$$

Die Eigenwerte sind die Nullstellen des **charakteristischen Polynoms**

$$p(\lambda) = \det\{\lambda\boldsymbol{I}_N - \boldsymbol{A}\}. \tag{2.80}$$

Der Grund dafür liegt in der Definition der Eigenwerte durch $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$. Daraus ergibt sich $(\lambda\boldsymbol{I}_N - \boldsymbol{A})\boldsymbol{v} = \boldsymbol{0}$, also die Tatsache, dass eine Linearkombination der Spalten der Matrix $(\lambda\boldsymbol{I}_N - \boldsymbol{A})$, realisiert durch die Multiplikation mit $\boldsymbol{v}$, null ergibt. Dies ist nur der Fall, wenn die Spalten von $(\lambda\boldsymbol{I}_N - \boldsymbol{A})$ linear abhängig sind, oder anders ausgedrückt, wenn $(\lambda\boldsymbol{I}_N - \boldsymbol{A})$ singulär ist. In diesem Fall gilt dann $\det\{(\lambda\boldsymbol{I}_N - \boldsymbol{A})\} = 0$, was der Nullsuche des charakteristischen Polynoms entspricht.

Während die Eigenwertzerlegung nur für quadratische Matrizen anwendbar ist, kann man die $M \times N$-Matrix $\boldsymbol{A}$ in das Produkt dreier spezieller Matrizen zerlegen, die den linearen Operator charakterisieren

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}. \tag{2.81}$$

Diese Zerlegung wird **Singulärwertzerlegung** genannt, wenn $\boldsymbol{U}$ eine orthonormale $M \times M$-Matrix, $\boldsymbol{V}$ eine orthonormale $N \times N$-Matrix und $\boldsymbol{\Sigma}$ eine $M \times N$-Blockdiagonalmatrix ist

$$\boldsymbol{\Sigma} = \left(\begin{array}{ccc|ccc} \sigma_1 & & & & \vdots & \\ & \ddots & & \cdots & 0 & \cdots \\ & & \sigma_r & & \vdots & \\ \hline & \vdots & & & \vdots & \\ \cdots & 0 & \cdots & \cdots & 0 & \cdots \\ & \vdots & & & \vdots & \end{array}\right), \quad \text{wobei} \quad \sigma_i > 0, \quad i = 1, \ldots, r. \tag{2.82}$$

Dabei hat die Matrix $\boldsymbol{A}$ den Rang r, und die positiven Diagonalelemente von $\boldsymbol{\Sigma}$ heißen **Singulärwerte**. Die Singulärwerte werden im oberen linken Block von $\boldsymbol{\Sigma}$ so angeordnet, dass gilt $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$. Die zu nichtverschwindenden Singulärwerten gehörenden Spalten von $\boldsymbol{U}$ nennt man **linke Singulärvektoren** und die entsprechenden Spalten von $\boldsymbol{V}$ **rechte Singulärvektoren**. Durch die Matrix $\boldsymbol{A}$ wird damit der rechte Singulärvektor $\boldsymbol{v}_1$ auf $\sigma_1 \boldsymbol{u}_1$ abgebildet (gestreckter/gestauchter linker Singulärvektor $\boldsymbol{u}_1$), der rechte Singulärvektor $\boldsymbol{v}_2$ auf $\sigma_2 \boldsymbol{u}_2$, usw. Äquivalent gilt auch $\boldsymbol{A}^{\mathrm{T}} \boldsymbol{u}_i = \sigma_i \boldsymbol{v}_i$ für $i = 1, \ldots, r$.

Man kann die Singulärwertzerlegung für folgende Interpretation des linearen Operators $\boldsymbol{A}$, angewandt auf einen Vektor $\boldsymbol{x}$, verwenden: Zunächst wird der Vektor $\boldsymbol{x}$ in der orthonormalen Basis dargestellt, die von den Spalten von $\boldsymbol{V}$ aufgespannt wird (evtl. nur Rotation). Dann werden die Elemente des sich ergebenden Vektors mit den Singulärwerten gewichtet, und zuletzt wird der sich daraus ergebende Vektor in der Basis dargestellt, die von den Spaltenvektoren der Matrix $\boldsymbol{U}$ aufgespannt wird (evtl. nur Rotation).

Durch die Eigenwertzerlegung der Matrix $\boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} = \boldsymbol{V} \boldsymbol{\Sigma}^2 \boldsymbol{V}^{\mathrm{T}}$ erhält man die von null verschiedenen Eigenwerte $\sigma_1^2, \ldots, \sigma_r^2$. Diese sind die quadrierten Singulärwerte von $\boldsymbol{A}$, und die dazugehörigen Eigenvektoren sind die rechten Singulärvektoren $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$. Äquivalent dazu führt die Eigenwertzerlegung von $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ zu den quadrierten Singulärwerten $\sigma_1^2, \ldots, \sigma_r^2$ von $\boldsymbol{A}$, und die dazugehörigen Eigenvektoren sind die linken Singulärvektoren $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$.

Übung 2.3; Übung 2.4

Die **Spektralnorm** der Matrix $\boldsymbol{A}$ entspricht ihrem maximalen Singulärwert

$$\|\boldsymbol{A}\|_2 = \sigma_1. \tag{2.83}$$

Die Singulärwertzerlegung kann verwendet werden, um eine rangreduzierte Matrix $\boldsymbol{A}_\ell$ zu finden, die $\boldsymbol{A}$ am besten approximiert. Da man die Matrix $\boldsymbol{A}$ aus einer Summe von Matrizen mit dem Rang 1 schreiben kann

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} \tag{2.84}$$

und die Singulärwerte in fallender Reihenfolge in $\boldsymbol{\Sigma}$ aufgetragen sind, kann man eine Matrix $\boldsymbol{A}_\ell$, mit $\ell < r$ einführen als

$$\boldsymbol{A}_\ell = \sum_{i=1}^{\ell} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}. \tag{2.85}$$

Diese Matrix ist die beste Rang-ℓ-Approximation von $\boldsymbol{A}$ im Sinne der Minimierung der Spektralnorm von $\boldsymbol{A} - \boldsymbol{A}_\ell$. Dieses Ergebnis (Theorem von Schmidt-Mirsky) ergibt sich aus der Definition der Spektralnorm und durch die Anordnung der Singulärwerte in $\boldsymbol{\Sigma}$.

Die **Konditionszahl** einer Matrix $\boldsymbol{A}$ bezüglich der Spektralnorm spielt eine wichtige Rolle in der Numerik und ist definiert als

$$\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\|_2 \left\|\boldsymbol{A}^{-1}\right\|_2, \tag{2.86}$$

also als Division des größten durch den kleinsten Singulärwert. Große Konditionszahlen deuten darauf hin, dass die Matrix fast singulär ist. Man sagt in diesem Fall, dass die Matrix schlecht konditioniert ist.

Zusätzlich zur Spektralnorm einer Matrix wird häufig die **Frobenius-Norm** für die $M \times N$-Matrix $\boldsymbol{A}$ verwendet

$$\|\boldsymbol{A}\|_\mathrm{F} = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} |a_{ij}|^2}. \tag{2.87}$$

Sie ist die Wurzel aus der Summe der Betragsquadrate aller Matrixelemente a_{ij}. Die Frobenius-Norm ergibt sich auch aus der Spur der Matrix $\boldsymbol{A}^\mathrm{T}\boldsymbol{A}$. Es gilt

$$\mathrm{tr}\left\{\boldsymbol{A}^\mathrm{T}\boldsymbol{A}\right\} = \sum_{i=1}^{M}\sum_{j=1}^{N} |a_{ij}|^2 = \|\boldsymbol{A}\|_\mathrm{F}^2. \tag{2.88}$$

Aus Gl. (2.88) und der Singulärwertzerlegung von $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathrm{T}$ ergibt sich folgender Zusammenhang

$$\|\boldsymbol{A}\|_\mathrm{F}^2 = \mathrm{tr}\left\{\boldsymbol{A}^\mathrm{T}\boldsymbol{A}\right\} = \mathrm{tr}\left\{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\mathrm{T}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\mathrm{T}\right\} = \mathrm{tr}\left\{\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\mathrm{T}\right\} = \mathrm{tr}\left\{\boldsymbol{\Sigma}^2\right\} = \sum_{i=1}^{r}\sigma_i^2, \tag{2.89}$$

wobei r der Rang von $\boldsymbol{A}$ ist.

Die p-**Normen** für Matrizen sind:

- Für $p = 1$: das Maximum der Summe der Beträge in den einzelnen Spalten

$$\|\boldsymbol{A}\|_1 = \max_{1\le j\le N}\left\{\sum_{i=1}^{M} |a_{ij}|\right\}; \tag{2.90}$$

- Für $p = 2$: die Spektralnorm aus Gl. (2.83);
- Für $p \to \infty$: das Maximum der Summe der Beträge in den einzelnen Zeilen

$$\|\boldsymbol{A}\|_\infty = \max_{1\le i\le M}\left\{\sum_{j=1}^{N} |a_{ij}|\right\}. \tag{2.91}$$

Übung 2.5; Übung 2.6

2.2 Optimierung mittels Lagrange-Multiplikatoren

Das Verfahren der Lagrange-Multiplikatoren ist eine Methode zur Lösung von Optimierungsaufgaben mit Nebenbedingungen. Die Aufgabe besteht darin, ein lokales Extremum einer Funktion in mehreren Veränderlichen zu finden und dabei die Nebenbedingungen zu erfüllen. Die Grundlagen dazu werden in diesem Unterkapitel kurz wiederholt. Eine ausführliche Behandlung der Grundlagen zur mathematischen Optimierung kann beispielsweise in [BV04] oder [BSS13] gefunden werden.

2.2.1 Optimierungsaufgaben mit Gleichungsnebenbedingungen

Die Aufgabe, die in diesem Unterkapitel behandelt wird, ist das Finden des Maximums der Funktion $f(\boldsymbol{x})$ unter der Nebenbedingung $g(\boldsymbol{x}) = 0$. Betrachtet man den N-dimensionalen Vektor $\boldsymbol{x}$ und die **Gleichungsnebenbedingung** $g(\boldsymbol{x}) = 0$, so stellt diese Nebenbedingung im N-dimensionalen Raum eine $(N-1)$-dimensionale Hyperfläche dar. Ein Beispiel ist die Nebenbedingung $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + t = 0$. Alle Punkte $\boldsymbol{x}$, für die gilt $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + t = 0$, bilden eine $(N-1)$-dimensionale Hyperebene. Für $N = 2$ definiert $g(\boldsymbol{x}) = 0$ eine Kurve, und in diesem 2-dimensionalen Raum lässt sich die Methode der Lagrange-Multiplikatoren geometrisch mit Hilfe von Abb. 2.1 erklären. Es soll der Punkt $\boldsymbol{x} = [x_1, x_2]^{\mathrm{T}}$ gefunden werden, der auf der Kurve $g(x_1, x_2) = 0$ liegt und

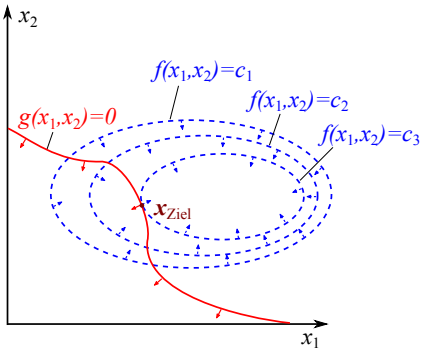


Abbildung 2.1 Geometrische Darstellung des Lagrange-Multiplikatoren-Verfahrens für Gleichungsnebenbedingungen.

bei dem $f(x_1, x_2)$ den größten Wert hat. In Abb. 2.1 sind Höhenlinien $f(x_1, x_2) = c$ für verschiedene Werte von c dargestellt, wobei $c_1 < c_2 < c_3$. Man sieht in der Skizze, dass es sich bei dem gesuchten Punkt um $\boldsymbol{x}_{\text{Ziel}}$ handelt.

Der Gradient $\nabla_{\boldsymbol{x}} g(\boldsymbol{x})$ der Nebenbedingung ist orthogonal zu der $(N-1)$-dimensionalen Hyperfläche, die von $g(\boldsymbol{x}) = 0$ bestimmt wird. In Abb. 2.1 sind Gradienten $\nabla_{\boldsymbol{x}} g(x_1, x_2)$ durch rote Pfeile, die senkrecht zur roten Kurve $g(x_1, x_2) = 0$ stehen, dargestellt. Die Orthogonalität zwischen $\nabla_{\boldsymbol{x}} g(\boldsymbol{x})$ und der Hyperfläche, die durch $g(\boldsymbol{x}) = 0$ bestimmt wird, lässt sich mit der Taylor-Reihenentwicklung der Funktion g um einen Punkt $\boldsymbol{x}_0 + \boldsymbol{\epsilon}$ erklären, wobei sowohl $\boldsymbol{x}_0$ als auch der an $\boldsymbol{x}_0$ nahe liegende Punkt $\boldsymbol{x}_0 + \boldsymbol{\epsilon}$ auf der Hyperfläche, die durch $g(\boldsymbol{x}) = 0$ bestimmt wird,

liegen. Es gilt entsprechend Gl. (2.63)

$$g(\boldsymbol{x}_0+\boldsymbol{\epsilon}) \approx g(\boldsymbol{x}_0) + \left(\nabla_{\boldsymbol{x}} g(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0}\right)^{\mathrm{T}} \boldsymbol{\epsilon} = g(\boldsymbol{x}_0) + \boldsymbol{\epsilon}^{\mathrm{T}} \left(\nabla_{\boldsymbol{x}} g(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0}\right). \tag{2.92}$$

Weil sowohl $g(\boldsymbol{x}_0)$ als auch $g(\boldsymbol{x}_0+\boldsymbol{\epsilon})$ auf der Hyperfläche, die durch $g(\boldsymbol{x})=0$ bestimmt wird, liegen, d. h. $g(\boldsymbol{x}_0)=0$ und $g(\boldsymbol{x}_0+\boldsymbol{\epsilon})=0$, ergibt sich aus Gl. (2.92)

$$\boldsymbol{\epsilon}^{\mathrm{T}} \left(\nabla_{\boldsymbol{x}} g(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0}\right) \approx 0. \tag{2.93}$$

Falls $\|\boldsymbol{\epsilon}\|$ gegen null geht, so gilt $\boldsymbol{\epsilon}^{\mathrm{T}} \left(\nabla_{\boldsymbol{x}} g(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0}\right) = 0$, und weil die Richtung von $\boldsymbol{\epsilon}$ in diesem Fall tangential zu der Hyperfläche, die durch $g(\boldsymbol{x})=0$ bestimmt wird, ist, gilt mit Gl. (2.11), dass $\left(\nabla_{\boldsymbol{x}} g(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}_0}\right)$ senkrecht zu dieser Hyperfläche ist. Deswegen sind die roten Pfeile in Abb. 2.1 senkrecht zu der Kurve $g(x_1,x_2)=0$.

Sucht man nun einen Punkt auf der Hyperfläche der Nebenbedingung $g(\boldsymbol{x})=0$, so dass $f(\boldsymbol{x})$ maximal wird, so muss in diesem Punkt der Gradient $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ auch orthogonal zu der Hyperfläche der Nebenbedingung sein. In Abb. 2.1 heißt dies, dass im gesuchten Punkt $\boldsymbol{x}_{\mathrm{Ziel}}$ der blaue Pfeil auch senkrecht zur roten Kurve $g(x_1,x_2)=0$ ist. Dies liegt daran, dass der Gradient immer in Richtung des steilsten Anstiegs zeigt, und daran, dass man, falls $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ nicht orthogonal zu der Hyperfläche der Nebenbedingung wäre, den Wert von $f(\boldsymbol{x})$ durch Vorwärts- oder Rückwärtsbewegung auf dieser Hyperfläche vergrößern könnte.

Deswegen gilt, dass im gesuchten Punkt $\boldsymbol{x}_{\mathrm{Ziel}}$ die beiden Gradienten $\nabla_{\boldsymbol{x}} g(\boldsymbol{x})$ und $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ parallele oder anti-parallele Vektoren sind. Damit muss es im gesuchten Punkt einen Parameter λ geben, so dass

$$\nabla_{\boldsymbol{x}} f\left(\boldsymbol{x}_{\mathrm{Ziel}}\right) + \lambda \nabla_{\boldsymbol{x}} g\left(\boldsymbol{x}_{\mathrm{Ziel}}\right) = 0, \tag{2.94}$$

wobei λ **Lagrange-Multiplikator** genannt wird. Der Lagrange-Multiplikator kann sowohl positiv als auch negativ sein.

Um den gesuchten Punkt $\boldsymbol{x}_{\mathrm{Ziel}}$ zu finden, der $f(\boldsymbol{x})$ maximiert und die Gleichungsnebenbedingung $g(\boldsymbol{x})=0$ erfüllt, kann man die **Lagrange-Funktion**

$$L(\boldsymbol{x},\lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}) \tag{2.95}$$

einführen. Setzt man die Ableitung $L(\boldsymbol{x},\lambda)$ nach $\boldsymbol{x}$ zu null, d. h. $\nabla_{\boldsymbol{x}} L(\boldsymbol{x},\lambda) = \boldsymbol{0}$, so ergibt sich Gl. (2.94), und wenn man $L(\boldsymbol{x},\lambda)$ nach λ ableitet und zu null setzt, d. h. $\frac{\partial L(\boldsymbol{x},\lambda)}{\partial \lambda} = 0$, so erhält man die Nebenbedingung $g(\boldsymbol{x})=0$. Damit kann man $\boldsymbol{x}_{\mathrm{Ziel}}$ und λ finden, indem man die stationären Punkte von $L(\boldsymbol{x},\lambda)$ berechnet. Aus

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x},\lambda) = \boldsymbol{0} \quad \text{und} \quad \frac{\partial L(\boldsymbol{x},\lambda)}{\partial \lambda} = 0 \tag{2.96}$$

ergeben sich $N+1$ Gleichungen, aus denen man sowohl $\boldsymbol{x}_{\mathrm{Ziel}}$ als auch λ berechnen kann. Falls nur $\boldsymbol{x}_{\mathrm{Ziel}}$ benötigt wird, kann man λ aus den Gleichungen eliminieren, und es ist nicht notwendig, den Wert von λ zu berechnen.

Übung 2.7

Möchte man beispielsweise ein unterbestimmtes Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ lösen, so besteht ein häufig verwendeter Ansatz darin, die 2-Norm des gesuchten Vektors $\boldsymbol{x}$ zu minimieren unter der Nebenbedingung $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, d. h.

$$\text{minimiere } \{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\} \quad \text{u. d. N.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{2.97}$$

Die dazugehörige Lagrange-Funktion lautet

$$L(\boldsymbol{x}, \lambda) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}). \tag{2.98}$$

Mit Gl. (2.96) erhält man daraus

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = 2\boldsymbol{x} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0} \quad \text{und} \tag{2.99}$$

$$\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{x}, \lambda) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0}. \tag{2.100}$$

Setzt man $\boldsymbol{x} = -\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}/2$ aus Gl. (2.99) in Gl. (2.100) ein, so erhält man

$$\boldsymbol{\lambda} = -2\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\right)^{-1}\boldsymbol{b} \tag{2.101}$$

und daraus schließlich die Lösung

$$\boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\right)^{-1}\boldsymbol{b}. \tag{2.102}$$

Die Matrix $\boldsymbol{A}^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\right)^{-1}$ wird **Moore-Penrose-Pseudoinverse** von $\boldsymbol{A}$ genannt.

2.2.2 Optimierungsaufgaben mit Ungleichungsnebenbedingungen

Die Aufgabe, die es zu lösen gilt, ist das Finden des Maximums der Funktion $f(\boldsymbol{x})$, wobei die **Ungleichungsnebenbedingung** $g(\boldsymbol{x}) \geq 0$ erfüllt sein muss. Abb. 2.2 visualisiert die Aufgabe für den Fall, dass die Dimension N des Vektors $\boldsymbol{x}$ den Wert 2 hat. Wie dort dargestellt, muss man

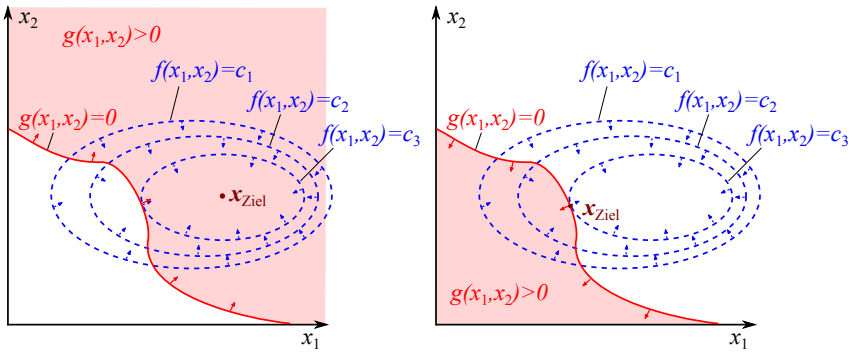


Abbildung 2.2 Geometrische Darstellung des Lagrange-Multiplikatoren-Verfahrens für Ungleichungsnebenbedingungen.

zwischen zwei Fällen unterscheiden, abhängig davon, ob der Punkt $\boldsymbol{x}_{\text{Ziel}}$ innerhalb der Region liegt, in der $g(\boldsymbol{x}) \geq 0$ gilt, oder ob er auf der Hyperfläche $g(\boldsymbol{x}) = 0$ liegt.

Im ersten Fall (in Abb. 2.2 links dargestellt), spricht man auch davon, dass die Nebenbedingung nicht aktiv ist, weil die Nebenbedingung keine Rolle spielt und der Punkt $\boldsymbol{x}_{\text{Ziel}}$ allein durch $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \boldsymbol{0}$ berechnet wird. Dies entspricht der Berechnung eines stationären Punktes der Lagrange-Funktion aus Gl. (2.95), wobei $\lambda = 0$ gilt.

Im zweiten Fall (in Abb. 2.2 rechts dargestellt), bei dem $\boldsymbol{x}_{\text{Ziel}}$ auf der Hyperfläche $g(\boldsymbol{x}) = 0$ liegt, spricht man auch davon, dass die Nebenbedingung aktiv ist. Es handelt sich um einen analogen Fall zu der Optimierung mit Gleichungsnebenbedingungen aus Unterkapitel 2.2.1

und entspricht der Berechnung eines stationären Punkts der Lagrange-Funktion aus Gl. (2.95), wobei $\lambda \neq 0$ gilt. Im Gegensatz zu der Optimierung mit Gleichungsnebenbedingungen spielt das Vorzeichen von λ allerdings eine Rolle. Ein Maximum von $f(\boldsymbol{x})$ wird nur erreicht, wenn der Gradient $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ in die entgegengesetzte Richtung der Region, in der $g(\boldsymbol{x}) \geq 0$ gilt, zeigt, so wie dies in Abb. 2.2 rechts visualisiert ist. Die blauen Pfeile stellen dabei die Gradientenvektoren der Funktion $f(\boldsymbol{x})$ in den jeweiligen Punkten dar. In diesem zweiten Fall gilt also $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = -\lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x})$, mit $\lambda > 0$.

In beiden Fällen gilt $\lambda g(\boldsymbol{x}_{\text{Ziel}}) = 0$. Im ersten Fall, weil $\lambda = 0$, und im zweiten, weil $\boldsymbol{x}_{\text{Ziel}}$ auf der Hyperfläche $g(\boldsymbol{x}) = 0$ liegt. Nutzt man auch hier die Lagrange-Funktion aus Gl. (2.95), $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$, so lässt sich für ein lokales Minimum $\boldsymbol{x}_{\text{Ziel}}$ sagen, dass es ein λ^* gibt, derart, dass

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}_{\text{Ziel}}, \lambda^*) = \mathbf{0} \tag{2.103}$$

$$g(\boldsymbol{x}_{\text{Ziel}}) \geq 0 \tag{2.104}$$

$$\lambda^* \geq 0 \tag{2.105}$$

$$\lambda^* g(\boldsymbol{x}_{\text{Ziel}}) = 0. \tag{2.106}$$

Diese Bedingungen nennt man **Karush-Kuhn-Tucker** (**KKT**)-Bedingungen. Sie sind sehr nützlich, weil damit nach Lösungen für $\boldsymbol{x}$ gesucht wird, für die man λ^* finden kann.

Möchte man die Funktion $f(\boldsymbol{x})$ minimieren (statt maximieren) unter der Ungleichungsnebenbedingung $g(\boldsymbol{x}) \geq 0$, so müsste man zur Visualisierung in Abb. 2.2 annehmen, dass $c_3 < c_2 < c_1$ und damit alle blauen Pfeile in die jeweils entgegengesetzte Richtung einzeichnen. Falls die Nebenbedingungen aktiv sind, gilt dann, dass die Gradienten $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ und $\nabla_{\boldsymbol{x}} g(\boldsymbol{x})$ in die gleiche Richtung zeigen und damit $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x})$ mit $\lambda > 0$, so dass die Lagrange-Funktion lautet

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda g(\boldsymbol{x}). \tag{2.107}$$

Die Aufgabe lautet damit, den stationären Punkt der Lagrange-Funktion aus Gl. (2.107) bezüglich $\boldsymbol{x}$ und λ zu bestimmen, mit $\lambda \geq 0$.

Möchte man die Methode der Lagrange-Multiplikatoren auf Optimierungsaufgaben mit mehreren Gleichheits- und Ungleichungsnebenbedingungen erweitern, so lässt sich dies entsprechend obiger Überlegungen realisieren. Besteht die Aufgabe darin, die Funktion $f(\boldsymbol{x})$ zu maximieren unter den Gleichungsnebenbedingungen $g_k(\boldsymbol{x}) = 0$ mit $k = 1, \ldots, K$ und den Ungleichungsnebenbedingungen $h_m(\boldsymbol{x}) \geq 0$ mit $m = 1, \ldots, M$, muss man die Lagrange-Multiplikatoren $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]^{\mathrm{T}}$ und $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_M]^{\mathrm{T}}$ einführen und die stationären Punkte der Lagrange-Funktion

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{k=1}^{K} \lambda_k g_k(\boldsymbol{x}) + \sum_{m=1}^{M} \mu_m h_m(\boldsymbol{x}) \tag{2.108}$$

bezüglich $\boldsymbol{x}$, $\boldsymbol{\lambda}$ und $\boldsymbol{\mu}$ berechnen mit den Bedingungen $\mu_m \geq 0$ und $\mu_m h_m(\boldsymbol{x}) = 0$, für $m = 1, \ldots, M$.

Übung 2.8

2.3 Wahrscheinlichkeitstheorie

Die **Stochastik**, bestehend aus den zwei Teilgebieten **Wahrscheinlichkeitstheorie** und **Statistik**, beschäftigt sich mit der Untersuchung und Modellierung des Zufalls. Während die Wahrscheinlichkeitstheorie die Formalisierung der Modellierung des Zufalls im Fokus hat, werden in der Statistik Beobachtungen zufälliger Ergebnisse aufbereitet und Aussagen über das zugrunde liegende Modell gemacht.

In diesem Unterkapitel werden einige Grundlagen der Wahrscheinlichkeitstheorie kurz wiederholt. Eine detaillierte Einführung in das Gebiet der Stochastik kann beispielsweise in [PP02] oder [SW11] gefunden werden.

2.3.1 Wahrscheinlichkeitsräume und Zufallsvariablen

Ausgangspunkt für die Wahrscheinlichkeitstheorie ist ein Experiment mit zufälligen Ergebnissen (**Zufallsexperiment**). Alle möglichen Ergebnisse fasst man in der **Ergebnismenge** Ω zusammen. Ein Ereignis ist eine Teilmenge von Ω, und man sagt, dass ein Ereignis eingetreten ist, wenn das Ergebnis des Zufallsexperiments in der entsprechenden Teilmenge liegt.

Um Ereignissen Wahrscheinlichkeiten zuzuordnen, werden sie in einer **σ-Algebra** $\mathbb{F}$ über Ω, auch **Ereignisraum** genannt, aufgeführt. Falls es sich um eine abzählbare Ergebnismenge handelt, ist eine σ-Algebra ein Mengensystem $\mathbb{F}$ mit $\mathbb{F} \subset \mathscr{P}(\Omega)$, wobei $\mathscr{P}(\Omega)$ die Potenzmenge (d. h. die Menge aller Teilmengen von Ω) ist, also eine Menge $\mathbb{F}$ von Teilmengen der Ergebnismenge Ω, die folgende Bedingungen erfüllt:

a) $\Omega \in \mathbb{F}$

b) Wenn für ein Ereignis A gilt $A \in \mathbb{F}$, dann ist auch das Komplementärereignis $\bar{A} \in \mathbb{F}$. Damit ist die leere Menge $\emptyset$ immer ein Element von $\mathbb{F}$.

c) Wenn $A_1, A_2, \ldots \in \mathbb{F}$, dann ist auch $\bigcup_{n \in \mathbb{N}} A_n \in \mathbb{F}$.

Für jede beliebige diskrete Ergebnismenge Ω ist $\{0, \Omega\}$ die kleinste und die Potenzmenge $\mathscr{P}(\Omega)$ die größte mögliche σ-Algebra. Auch für den Fall nicht abzählbarer Ergebnisräume entspricht die σ-Algebra weiterhin einem Teilmengensystem über dem Ergebnisraum Ω.

Die **Wahrscheinlichkeiten** sind die Ergebnisse der Abbildung P der σ-Algebra in das Intervall $[0,1]$. Das Tripel $(\Omega, \mathbb{F}, \mathrm{P})$ heißt **Wahrscheinlichkeitsraum**. Abb. 2.3 veranschaulicht anhand eines Beispiels einen Wahrscheinlichkeitsraum. Um ein Wahrscheinlichkeitsmaß zu sein, muss P die **Axiome von Kolmogorow** erfüllen:

a) Für jedes Ereignis A aus $\mathbb{F}$ gilt $0 \leq \mathrm{P}(A) \leq 1$.

b) $\mathrm{P}(\Omega) = 1$

c) Wenn die Ereignisse A und B disjunkt sind, d. h. $A \cap B = \emptyset$, dann ist die Wahrscheinlichkeit der Vereinigung von A und B die Summe der Wahrscheinlichkeiten der einzelnen Ereignisse: $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$.

Eine **Zufallsvariable** ist eine Funktion, die den Ergebnissen eines Zufallsexperimentes Werte zuordnet. Mit Hilfe von Zufallsvariablen modelliert man die Tatsache, dass die Werte einer Funktion vom Zufall abhängig sind. Die Werte der Zufallsvariablen werden **Realisierungen** genannt. Während die Ergebnisse des Zufallsexperiments sehr unanschaulich sein können, sind

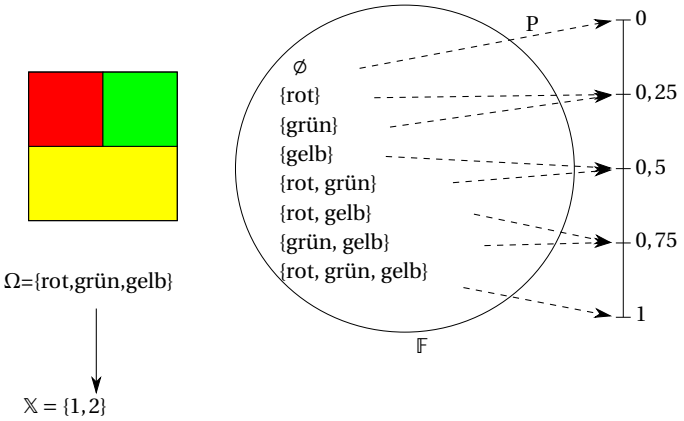


Abbildung 2.3 Wahrscheinlichkeitsraum $(\Omega, \mathbb{F}, \mathrm{P})$ und Zufallsvariable x, die von Ω nach $\mathbb{X}$ abbildet.

die Werte, die die Zufallsvariable annimmt, im Allgemeinen beobachtbar und anschaulich. Aus der Ergebnismenge Ω wird durch die Zufallsvariable der **Beobachtungsraum** $\mathbb{X}$. Zum Beispiel sind für einen Farbenblinden die Ergebnisse des Zufallsexperiments aus Abb. 2.3 unanschaulich. Wenn aber eine Zufallsvariable „rot“ auf 1, „grün“ auf 2 und „gelb“ auf 3 abbildet, so sind diese Werte für den Farbenblinden anschaulich.

Verwendet man in dem Beispiel aus Abb. 2.3 eine Zufallsvariable

$$x: \quad \Omega \to \mathbb{X}, \quad \omega \mapsto x, \quad \text{mit} \quad x = \begin{cases} 1 & \text{falls } \omega = \text{„rot“ oder } \omega = \text{„grün“} \\ 2 & \text{falls } \omega = \text{„gelb“}, \end{cases} \tag{2.109}$$

so erhält man für die Elemente des Beobachtungsraums die Wahrscheinlichkeiten

$$\begin{aligned} &\mathrm{P}(x=1) = \mathrm{P}\left(\omega \in \Omega \text{ mit } x(\omega) = 1\right) = 0{,}5 \quad \text{und} \\ &\mathrm{P}(x=2) = \mathrm{P}\left(\omega \in \Omega \text{ mit } x(\omega) = 1\right) = 0{,}5. \end{aligned} \tag{2.110}$$

In Abb. 2.3 ist auch die Abbildung der Zufallsvariable x dargestellt.

Zufallsvariablen, die Elemente der Ergebnismenge eindeutig den Elementen der natürlichen Zahlen zuordnen, so dass die Realisierungen $x(\omega) \in \{x_1, x_2, \ldots\}$ abzählbar sind, heißen **diskrete Zufallsvariablen**. Diese werden durch ihre Wahrscheinlichkeit $\mathrm{P}(x = x_i)$ beschrieben. Häufig wird vereinfachend statt $\mathrm{P}(x = x_i)$ nur $\mathrm{P}(x_i)$ geschrieben.

Zufallsvariablen, deren Realisierungen $x(\omega)$ nicht abzählbar sind, heißen **kontinuierliche Zufallsvariablen**. Diese werden durch ihre **Wahrscheinlichkeitsdichtefunktion** (**WDF**) beschrieben

$$\mathrm{p}(x = x) = \lim_{\mathrm{d}x \to 0} \frac{\mathrm{P}\left(x - \frac{1}{2}\mathrm{d}x \le x \le x + \frac{1}{2}\mathrm{d}x\right)}{\mathrm{d}x}. \tag{2.111}$$

Unter der WDF einer kontinuierlichen vektorwertigen Zufallsvariablen, bestehend aus N skalaren Zufallsvariablen, versteht man

$$\mathrm{p}(\boldsymbol{x} = \boldsymbol{x}) = \lim_{\mathrm{d}x_1 \to 0, \ldots, \mathrm{d}x_N \to 0} \frac{\mathrm{P}\left(\left\{x_1 - \frac{\mathrm{d}x_1}{2} \le x_1 \le x_1 + \frac{\mathrm{d}x_1}{2}\right\} \cap \ldots \cap \left\{x_N - \frac{\mathrm{d}x_N}{2} \le x_N \le x_N + \frac{\mathrm{d}x_N}{2}\right\}\right)}{\mathrm{d}x_1 \ldots \mathrm{d}x_N}. \tag{2.112}$$