Statistische Methoden der Datenanalyse

Vorlesung im Sommersemester 2008

H. Kolanoski
Humboldt-Universität zu Berlin

Inhaltsverzeichnis

Literaturverzeichnis **iv**

1 Grundlagen der Statistik **3**
1.1 Wahrscheinlichkeit . . . 3
1.1.1 Definition über die Häufigkeit . . . 3
1.1.2 Kombinatorische Definition . . . 3
1.1.3 Axiomatische Definition der Wahrscheinlichkeit . . . 4
1.2 Verteilungen von Zufallsvariablen . . . 7
1.2.1 Eigenschaften von Verteilungen . . . 8
1.2.2 Erwartungswerte . . . 11
1.2.3 Wahrscheinlichster Wert und Median . . . 12
1.2.4 Stichproben und Schätzwerte . . . 13
1.3 Simulation von Verteilungen . . . 14
1.3.1 Umkehrung der Verteilungsfunktion . . . 14
1.3.2 'Hit and Miss' Methode . . . 16

2 Spezielle Verteilungen einer Variablen **19**
2.1 Binomial-Verteilung . . . 19
2.2 Multinomial-Verteilung . . . 24
2.3 Poisson-Verteilung . . . 25
2.4 Gleichverteilung . . . 27
2.5 Normalverteilung . . . 30
2.5.1 Vertrauensintervalle: . . . 32
2.6 Zentraler Grenzwertsatz . . . 34

3 Verteilungen mehrerer Variablen **37**
3.1 Eigenschaften von Verteilungen mehrerer Variablen . . . 37
3.1.1 Wahrscheinlichkeitsdichte, Verteilungsfunktion, Randverteilung 37
3.1.2 Bedingte Wahrscheinlichkeitsdichten, Selektionsschnitte . . . 38
3.2 Erwartungswerte . . . 39
3.3 Kovarianzmatrix . . . 40
3.3.1 Definition und Eigenschaften der Kovarianzmatrix . . . 40
3.3.2 Beispiel: Multi-dimensionale Gaussverteilung . . . 40
3.3.3 Kovarianzen von Stichproben . . . 41
3.3.4 Kovarianzmatrix von unabhängigen Variablen . . . 41
3.3.5 Korrelationen . . . 42
3.4 Lineare Funktionen von mehreren Zufallsvariablen . . . 44

3.5 Nicht-lineare Funktionen von Zufallsvariablen 46
3.5.1 Eine Funktion von einem Satz von Zufallsvariablen 46
3.5.2 Mehrere Funktionen von einem Satz von Zufallszahlen 47
3.6 Transformationen von Zufallsvariablen 50

4 Stichproben und Schätzungen **53**
4.1 Stichproben, Verteilungen und Schätzwerte 53
4.2 Eigenschaften von Schätzwerten 54
4.3 Stichproben aus Normalverteilungen;
χ^2-Verteilung 56

5 Monte-Carlo-Methoden **63**
5.1 Einführung 63
5.2 Zufallszahlengeneratoren 65
5.2.1 Multiplikativ kongruentielle Generatoren 65
5.2.2 Mersenne-Twister 66
5.2.3 Quasi-Zufallszahlen 66
5.3 Monte-Carlo-Erzeugung von Ereignissen 67
5.3.1 Inversionsmethode 67
5.3.2 'Hit-and-Miss'-Verfahren 70
5.3.3 Majorantenmethode 70
5.3.4 Wichtung der Ereignisse 71
5.4 Monte-Carlo-Integration 72
5.4.1 Majoranten-Methode mit Hit-or-Miss 72
5.4.2 MC-Integration mit Ereigniswichtung 73
5.4.3 Varianz-reduzierende Verfahren 73
5.4.4 Stratified Sampling ('Geschichtete Stichproben') 74

6 Die Maximum-Likelihood-Methode **77**
6.1 Das Maximum-Likelihood-Prinzip 77
6.2 ML-Methode für Histogramme 80
6.3 Berücksichtigung von Zwangsbedingungen 81
6.3.1 Methode der Lagrange-Multiplikatoren 82
6.3.2 Zwangsbedingungen als Zufallsverteilungen 82
6.3.3 Erweiterte ML-Methode 83
6.3.4 Freiheitsgrade und Zwangsbedingungen 84
6.4 Fehlerbestimmung für ML-Schätzungen 85
6.4.1 Allgemeine Methoden der Varianzabschätzung 85
6.4.2 Varianzabschätzung durch Entwicklung um das Maximum . . 86
6.4.3 Vertrauensintervalle und Likelihood-Kontouren 86
6.5 Eigenschaften von ML-Schätzungen 88

7 Methode der kleinsten Quadrate **91**
7.1 Prinzip der Methode der kleinsten Quadrate 91
7.2 Lineare Anpassung 92
7.2.1 Anpassung der Messwerte an eine Gerade 92
7.2.2 Anpassung einer allgemeinen linearen Funktion der Parameter 95
7.3 Anpassung nicht-linearer Funktionen der Parameter 100

8 Signifikanzanalysen **103**
8.1 Einführung . . . 103
8.2 Prüfung von Hypothesen . . . 104
8.2.1 χ^2-Test . . . 104
8.2.2 Studentsche t-Verteilung . . . 104
8.2.3 F-Verteilung . . . 106
8.2.4 Kolmogorov-Smirnov-Test . . . 108
8.3 Vertrauensintervalle . . . 110
8.3.1 Bayes-Vertrauensintervalle . . . 110
8.3.2 'Klassische' Vertrauensintervalle . . . 110

9 Klassifikation und statistisches Lernen **115**
9.1 Einführung . . . 115
9.2 Schätzung von Wahrscheinlichkeitsdichten . . . 117
9.3 Lineare Diskriminanten . . . 118
9.3.1 Klassentrennung durch Hyperebenen . . . 118
9.3.2 Fisher-Diskriminante . . . 119
9.4 Neuronale Netze zur Datenklassifikation . . . 120
9.4.1 Einleitung: Neuronale Modelle . . . 120
9.4.2 Natürliche neuronale Netze . . . 122
9.4.3 Künstliche neuronale Netze (KNN) . . . 126
9.4.4 Das einfache Perzeptron . . . 128
9.4.5 Das Mehrlagen-Perzeptron . . . 132
9.4.6 Lernen . . . 136
9.4.7 Typische Anwendungen für Feed-Forward-Netze . . . 141
9.4.8 BP-Lernen und der Bayes-Diskriminator . . . 144
9.5 Entscheidungsbäume . . . 149
9.5.1 Aufwachsen eines Baumes . . . 149
9.5.2 Verstärkte Entscheidungsbäume . . . 151
9.6 Stützvektormaschinen . . . 152
9.6.1 Lineare SVM-Klassifikation . . . 153
9.6.2 Nichtlineare Erweiterung mit Kernelfunktionen . . . 155

Literaturverzeichnis

[1] S. Brandt: 'Datenanalyse', 4. Auflage, 1999, Spektrum Akademischer Verlag.

[2] R.J. Barlow, 'Statistics: A Guide to the Use of Statistic al Methods in the Physical Sciences', Wiley, 1989.

[3] V. Blobel und E. Lohrmann, 'Statistische und numerische Methoden der Datenanalyse', Teubner Studienbücher, 1998.

[4] G. Bohm und G. Zech, 'Einführung in Statistik und Messwertanalyse für Physiker', Hamburg, DESY 2005; e-book:<http://www-library.desy.de/preparch/books/vstatmp.pdf>

Zu "Neuronale Netze":

[5] D.E.Rumelhart and J.L.McClelland: *'Parallel Distributed Processing',* MIT Press 1984 (9.Aufl. 1989).

[6] J.Hertz, A.Krogh and R.G.Palmer: *'Introduction to the Theory of Neural Computation',* Addison-Wesley Publishing Company, 1991.

[7] R.Brause: *'Neuronale Netze',* Teubner Verlag 1991.

[8] R.Hecht-Nielsen: *'Neurocomputing',* Addison-Wesley Publishing Company, 1987.

[9] H.Ritter, T.Martinetz und K.Schulten: *'Neuronale Netze. Eine Einführung in die Neuroinformatik selbstorganisierender Netze',* Addison-Wesley Publishing Company, 1991.

[10] G.E.Hinton: *'Wie Neuronale Netze aus Erfahrung lernen',* Spektrum der Wissenschaft, Nov. 1992.

[11] T.Kohonen: *'Self-Organization and Associative Memory',* Springer Verlag, 3.Auflage 1989.

[12] A.Zell: *'Simulation Neuronaler Netze',* Addison-Wesley, 1.Auflage 1994.

[13] Scientific American: *'The Brain',* Vol. 241, Sept. 1979.

[14] Spektrum der Wissenschaft, Nov. 1992.

"PDG":

[15] W.-M. Yao et al. (Particle Data Group), 'Review of Particle Physics', J. Phys. G33, 1 (2006); http://pdg.lbl.gov (Kapitel 31-32, reviews).

Monte-Carlo-Methoden:

[16] F. James,"Monte Carlo Theory and Practice", Rept. Prog. Phys. 43 (1980) 1145.

Einführung

Der Ausgang physikalischer Experimente ist in der Regel mit Unsicherheiten behaftet, das heißt, das Resultat ist unvorhersagbar, zufällig. Diese Unsicherheit kann zwei unterschiedliche Ursachen haben:

- eine **Unsicherheit im Messprozess**, die zu **Messfehlern** führt;
- der grundsätzlich **statistische Charakter von physikalischen Prozessen** (statistisches Verhalten in Vielteilchensystemen, zum Beispiel Molekülbewegung in Gasen, oder quantenmechanische Prozesse, die nur Wahrscheinlichkeitsaussagen zulassen).

Um physikalische Experimente interpretieren zu können, benötigt man deshalb statistische Methoden, die in dieser Vorlesung selektiv und auf einem einführenden Niveau behandelt werden. Beispiele für die Anwendung statistischer Methoden sind:

- Bestimmung von **Wahrscheinlichkeiten** für das Auftreten von Ereignissen, häufig als Funktion einer oder mehrere Variablen, für die man dann Wahrscheinlichkeitsverteilungen erhält.
- Bestimmung der **Unsicherheit einer Messgröße**. Die Angabe eines Messergebnisses ohne einen Messfehler ist sinnlos!

  Beispiel: Die Messung der Lichtgeschwindigkeit zu $2.8 \cdot 10^8\,\mathrm{m/s}$ ist konsistent mit dem festgelegten Wert $2.99792458 \cdot 10^8\,\mathrm{m/s}$, wenn der Fehler der Messung zum Beispiel zu etwa ± 0.2 abgeschätzt wird:

  $$c = (2.8 \pm 0.2) \cdot 10^8\,\mathrm{m/s}$$

  Bei der Angabe

  $$c = (2.8 \pm 0.01) \cdot 10^8\,\mathrm{m/s}$$

  wird man sich andererseits wundern müssen, ob das eine große Entdeckung ist oder ob eher Quellen von Unsicherheit unberücksichtig geblieben sind.

  Es gibt zwei unterschiedliche Quellen von Unsicherheiten in einem Messprozess:

  - **statistische Fehler**, die in der Regel experimentell bestimmt werden können;

  - **systematische Fehler**, zu deren Abschätzung häufig die Erfahrung eines guten Experimentators notwendig ist.

- Beurteilung der **Signifikanz von Messsignalen** basiert auf der Bestimmung der Messfehler (Beispiel: das Signal einer kosmischen Radioquelle über einem Hintergrundrauschen).

  Die zu erwartende Signifikanz eines experimentellen Ergebnisses sollte bereits bei der **Vorbereitung des Experimentes** berücksichtigt werden. So könnte man zum Beispiel mit statistischen Methoden festlegen, welcher Anteil der Messzeit bei dem obigen Beispiel für die Messung des Hintergrundes verwendet werden soll. Solche Planungen sind natürlich besonders wichtig, wenn die Experimente sehr zeitaufwendig und/oder kostspielig sind.

- **Entscheidung über Modellhypothesen**, die die Daten beschreiben: wann kann eine Hypothese akzeptiert werden, wann sollte sie verworfen werden, in welchem Bereich liegen die Parameter eines Modells.

- **Ausgleichsrechnung**: statistisch korrekte Ausgleich von Messwerten, die ein System überbestimmen (mehr Messungen als freie Parameter). Beispiele sind die Anpassung von Modellen an Daten und Bestimmung von Modellparametern oder die Berücksichtigung von Zwangsbedingungen in der Rekonstruktion von Teilchenreaktionen aus gemessenen Viererimpulsen.

- Berechnung komplizierter Prozesse durch **Simulationen**: die sogenannte Monte-Carlo-Methode bedient sich dabei statistischer Methoden. Zum Beispiel bei der Bestimmung der Nachweiswahrscheinlichkeit eines Detektors oder bei der Analyse von Produktionsabläufen, Vorratshaltung, Finanzierungsmodellen usw. in der Wirtschaft.

- **Entfaltung:** Rückrechnung einer "wahren" Verteilung aus einer gemessenen mit Berücksichtigung von Auflösungs- und Effizienz-Effekten.

- **Klassifizierung:** Einteilung von Ereignissen in Klassen auf der Basis der, im allgemeinen multivariaten, Messwerte. Es gibt Klassifikationsalgorithmen, die auf die Erkennung der richtigen Klasse eines Ereignisses trainiert werden können, wie zum Beispiel Neuronale Netze (**'statistisches Lernen'**)

Bei der Analyse von Daten kann man in der Regel auf Statistik- und Datenanalyseprogramme auf Computern zurückgreifen. Die Anwendung solcher Programme setzt aber ein gutes Verständnis der statistischen Methodik und sorgfältige Analysen der jeweils vorliegenden Problematik voraus.

Kapitel 1

Grundlagen der Statistik

1.1 Wahrscheinlichkeit

Grundlegend für statistische Analysen, das heißt der Behandlung von Vorgängen mit zufälligem, unvorhersagbarem Ausgang, ist der Begriff der Wahrscheinlichkeit. Obwohl so grundlegend, wird über die Definition der Wahrscheinlichkeit immer noch, zum Teil sehr emotional, gestritten. Es gibt eine, nicht umstrittene, axiomatische Definition, die die Rechenregeln festlegt, aber offen lässt, wie man tatsächlich Wahrscheinlichkeiten bestimmt. In der Praxis benutzt man meistens eine Definition über die relative Häufigkeit von Ereignissen.

1.1.1 Definition über die Häufigkeit

Wenn man N Versuche macht, bei denen das Ereignis e auftreten kann, und dabei n mal das Ereignis e tatsächlich auftritt, ordnet man dem Ereignis e die Wahrscheinlichkeit $p(e)$ durch die relative Häufigkeit des Auftretens des Ereignisses zu:

$$p(e) = \lim_{N\to\infty} \frac{n}{N} \tag{1.1}$$

In der Praxis wird der Grenzübergang zu unendlich vielen Versuchen erschlossen oder aus endlichen 'Stichproben' abgeschätzt.

1.1.2 Kombinatorische Definition

Wahrscheinlichkeiten können erschlossen werden, wenn man zum Beispiel aus Symmetriebetrachtungen argumentieren kann, dass alle möglichen Ereignisse gleich wahrscheinlich sind, zum Beispiel welche Zahl beim Würfeln erscheint. Dann ist die Wahrscheinlichkeit für jedes einzelne Ereignis durch die Anzahl der mögliche Ereignisse N gegeben:

$$p(e) = \frac{1}{N} \tag{1.2}$$

Zum Beispiel ist die Wahrscheinlichkeit für das Würfeln einer 6 gerade 1/6 und das Werfen von 'Zahl' bei einer Münze 1/2. Beim Werfen von zwei Würfeln ist jede Kombination von Zahlen gleich wahrscheinlich, also 1/36 (weil es $6 \cdot 6 = 36$ Kombinationen gibt). Was ist die Wahrscheinlichkeit, dass mindestens eine 6 auftritt?

Dazu muss man die Anzahl der Kombinationen mit mindestens einer 6 abzählen: 1) der erste Würfel hat eine 6 und der andere hat die Zahlen 1 bis 5; 2) dasselbe für die ausgetauschten Würfel; 3) beide haben eine 6. Das sind also $2 \cdot 5 + 1 = 11$ Kombinationen und damit ist die Wahrscheinlichkeit 11/36.

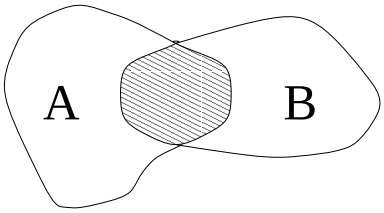
Der Fall, das alle Möglichkeiten gleich wahrscheinlich sind, hat in der Physik eine besondere Bedeutung: in der Quantentheorie kann ein physikalisches System verschiedene Zustände einnehmen, die alle mit gleicher Wahrscheinlichkeit auftreten.

1.1.3 Axiomatische Definition der Wahrscheinlichkeit

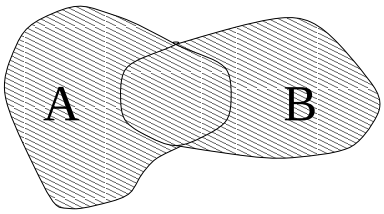
Ereignismenge: Es sei

$$\Omega = \{e_i\} \tag{1.3}$$

die Menge aller möglichen Ereignisse, zum Beispiel die möglichen Resultate eines Experimentes. Für Untermengen A, B, $C \subseteq \Omega$ werden die üblichen Verknüpfungen, Durchschnitt und Vereinigung, definiert:



$$A \cdot B = C; \quad A \text{ UND } B = C; \quad A \cap B = C \tag{1.4}$$



$$A + B = C; \quad A \text{ ODER } B = C; \quad A \cup B = C \tag{1.5}$$

Durchschnitt $\cap$ und Vereinigung $\cup$ entsprechen den logischen Operationen UND $(\cdot)$ und ODER $(+)$.

Weiterhin wird ein **elementares Ereignis**, das **Komplement** $\bar{A}$ von A und das **sichere Ereignis** E definiert ($\emptyset$ ist die leere Menge):

$$A \text{ elementar} \iff A \cdot B = \emptyset \text{ oder } A \cdot B = A \quad \forall\, B \,\epsilon\, \Omega \tag{1.6}$$

Das Nichteintreten von A ist $\bar{A}$ und damit sind

$$A + \bar{A} = E, \qquad A \cdot \bar{A} = \emptyset \tag{1.7}$$

das sichere und das unmögliche Ereignis.

Wahrscheinlichkeitsaxiome: Jedem Ereignis $A \epsilon \Omega$ wird eine Zahl $p(A)$ mit folgenden Eigenschaften zugeordnet:

(1) $0 \leq p(A) \leq 1$

(2) p(E)=1

(3) $A \cdot B = \emptyset \implies p(A + B) = p(A) + p(B)$

Offensichtlich erfüllen die beiden oben angegebenen Definitionen für die Wahrscheinlichkeit diese Axiome. Andererseits legen die Axiome nicht fest, wie man tatsächlich Wahrscheinlichkeiten bestimmen soll.

Aus den Axiomen ergibt sich:

- Eine Untermenge A von B hat eine kleinere Wahrscheinlichkeit als B:

$$A \subset B \implies p(A) \leq p(B) \tag{1.8}$$

- Im allgemeinen, falls (3) nicht zutrifft, also $A \cdot B \neq \emptyset$ ist, gilt das **Additionstheorem**:

$$p(A + B) = p(A) + p(B) - p(A \cdot B) \tag{1.9}$$

Bedingte Wahrscheinlichkeiten: Die Wahrscheinlichkeit von A, wenn B gegeben ist, wird mit $p(A|B)$ bezeichnet:

$$p(A|B) = p(A) \text{ gegeben } B \tag{1.10}$$

Zum Beispiel ändert sich die Wahrscheinlichkeit schwarzhaarig zu sein, wenn man die beiden Bedingung betrachtet, dass die Person eine Deutsche oder dass die Person eine Griechin ist. Die bedingte Wahrscheinlichkeit ergibt sich zu:

$$p(A|B) = \frac{p(A \cdot B)}{p(B)} \tag{1.11}$$

Das ist also zum Beispiel die Wahrscheinlichkeit, schwarzhaarig und Grieche zu sein, normiert auf die Wahrscheinlichkeit Grieche zu sein. Mit der Häufigkeitsdefinition würde man also die Anzahl der schwarzhaarigen Griechen durch die Zahl aller Griechen dividieren.

Die Gleichung (1.11) lässt sich nach $p(A \cdot B)$ auflösen:

$$p(A \cdot B) = p(A|B) \cdot p(B) = p(B|A) \cdot p(A) \tag{1.12}$$

Daraus folgt das **Bayes-Theorem**:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)} \tag{1.13}$$

Beispiel: Eine Krankheit K trete in der gesamten Bevölkerung mit der Häufigkeit $p(K) = 10^{-4}$ auf. Auf diese Krankheit reagiert ein zu derem Nachweis entwickelter Test mit einer Wahrscheinlichkeit von 98% positiv (+), also

$p(+|K) = 0.98$. Allerdings spricht die Gesamtbevölkerung mit einer Wahrscheinlichkeit von 3% ebenfalls positiv an, also $p(+) = 0.03$. Was ist die Wahrscheinlichkeit, die Krankheit zu haben, wenn das Testresultat positiv ist? Die Rechnung ergibt:

$$p(K|+) = \frac{p(+|K) \cdot p(K)}{p(+)} = \frac{0.98 \cdot 10^{-4}}{0.03} \approx 0.003 \tag{1.14}$$

Diese geringe Wahrscheinlichkeit von nur 3 Promille würde zum Beispiel einen schwereren Eingriff, der im Krankheitsfall notwendig würde, nicht rechtfertigen. Obwohl die Effizienz des Tests, die Krankheit nachzuweisen, recht gut ist, ist die Fehlerrate bei Gesunden relativ hoch. Das liegt daran, dass die 'a priori' Wahrscheinlichkeit für das Auftreten der Krankheit sehr klein ist. Das gleiche Problem tritt auf, wenn man in Experimenten sehr seltene Ereignisse identifizieren will, die Identifikation aber auch auf die anderen Ereignisse mit einer zwar kleinen aber endlichen Wahrscheinlichkeit anspricht. Abhilfe schaffen hier nur weitere unabhängige Tests, so dass sich die Ansprechwahrscheinlichkeiten multiplizieren.

Unabhängige Ereignisse: Man nennt zwei Ereignisse unabhängig, wenn gilt:

$$A, B \text{ unabhängig} \iff p(A|B) = p(A) \iff p(A \cdot B) = p(A) \cdot p(B) \tag{1.15}$$

Beispiel: Wenn man zwei Würfel wirft, sind die Ergebnisse beider Würfel unabhängig voneinander. Die Wahrscheinlichkeit zweimal 6 zu würfeln ist demnach

$$\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36},$$

wie man auch mit dem kombinatorischen Ansatz durch Abzählen der möglichen Fälle findet.

Allgemeine Form des Bayes-Theorems: Wenn die Gesamtheit der Ereignisse E sich vollständig in unabhängige Ereignisse oder Klassen A_i zerlegen läßt,

$$E = \sum_{i=1}^{n} A_i, \tag{1.16}$$

dann läßt sich B als Summe der möglichen Klassenzugehörigkeiten darstellen:

$$p(B) = \sum_{i=1}^{n} p(B|A_i)p(A_i) \tag{1.17}$$

Eingesetzt in (1.13) ergibt sich das Bayes-Theorem in allgemeinerer Form:

$$p(A_j|B) = \frac{p(B|A_j) \cdot p(A_j)}{\sum_{i=1}^{n} p(B|A_i)p(A_i)} \tag{1.18}$$

Beispiel: In dem obigen Beispiel mit dem Test zum Nachweis einer Krankheit hatten wir $p(+) = 0.03$ als die Wahrscheinlichkeit, mit der die Gesamtbevölkerung auf den Test anspricht, angesetzt. Zerlegen wir die Gesamtheit in Kranke und Nichtkranke, K und $\bar{K}$, dann ist diese Wahrscheinlichkeit:

$$p(+) = p(+|K)p(K) + p(+|\bar{K})p(\bar{K}) \tag{1.19}$$

und Gleichung (1.14) wird:

$$p(K|+) = \frac{p(+|K) \cdot p(K)}{p(+|K)p(K) + p(+|\bar{K})p(\bar{K})} \tag{1.20}$$

Eine solche Darstellung ist sinnvoll, wenn die Testergebnisse für beide Klassen getrennt vorliegen.

1.2 Verteilungen von Zufallsvariablen

Das Ergebnis eines Experimentes wird durch eine Zufallsvariable x oder einen Satz von Zufallsvariablen $\vec{x} = (x_1, x_2, \ldots)$ beschrieben. Diese Variablen können diskrete oder kontinuierliche Werte haben.

Bei **diskreten Variablen** n können wir eine Wahrscheinlichkeit $p(n)$ für das Auftreten eines bestimmten Wertes von n angeben. Ein Beispiel ist die Wahrscheinlichkeit für das Auftreten von n Zerfällen eines radioaktiven Präparates in einem festen Zeitintervall Δt. Üblicherweise werden solche Verteilungen diskreter Variablen wie in Abb. 1.1 als Treppenfunktion dargestellt.

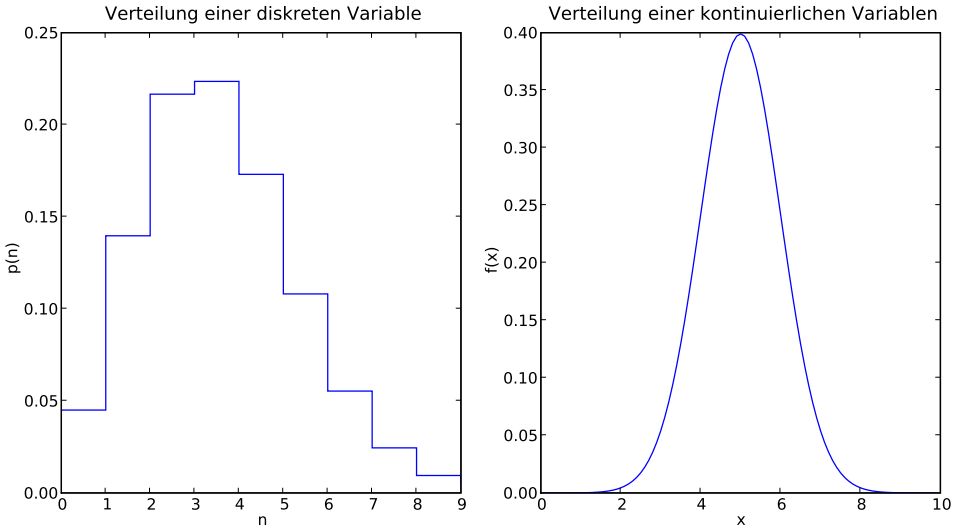


Abbildung 1.1: Beispiele von Wahrscheinlichkeitsverteilungen: diskrete Variable (links); kontinuierliche Variable (rechts).

Bei **kontinuierlichen Variablen** gibt man eine Wahrscheinlichkeit für das Auftreten von x in einem Intervall Δx an:

$$\Delta p(x) = \frac{\Delta p(x)}{\Delta x}\Delta x \qquad \stackrel{\Delta x \to 0}{\longrightarrow} \qquad dp(x) = \frac{dp(x)}{dx}dx = f(x)dx, \tag{1.21}$$

wobei $f(x)$ **Wahrscheinlichkeitsdichte** genannt wird (mit der Dimension von x^{-1}).

1.2.1 Eigenschaften von Verteilungen

Normierung: Die Wahrscheinlichkeit, irgendeinen möglichen Wert von x bzw. n zu erhalten, muss 1 sein:

$$\begin{aligned} &\text{kontinuierliche Variable :} && \int_{-\infty}^{+\infty} f(x)dx = 1 \\ &\text{diskrete Variable :} && \sum_{n=0}^{+\infty} p(n) = 1 \end{aligned} \tag{1.22}$$

Die Integrations- oder Summationsgrenzen können auch allgemeiner gewählt werden (x_{min}, x_{max} bzw. n_{min}, n_{max}), zur Vereinfachung benutzten wir im Folgenden aber meistens die Grenzen wie in (1.22).

Beispiel: In der Physik treten häufig Exponentialfunktionen auf, die Wachstum oder Abnahme proportional dem jeweils Vorhandenen und der Intervallänge dx der Variablen beschreiben. Die physikalische Annahme ist, dass die Wahrscheinlichkeit pro Zeitintervall gleich und unabhängig von der bereits verstrichenen Zeit ist. Für einen Absorptions- oder Zerfallsprozess ergibt sich zum Beispiel:

$$df(x) = -f(x)\,\lambda\,dx \tag{1.23}$$

Bekanntlich ergibt sich daraus:

$$f(x) = f_0\,e^{-\lambda x} \tag{1.24}$$

Diese Wahrscheinlichkeitsdichte soll im x-Intervall $[0, \infty]$ normiert werden:

$$1 = \int_0^\infty f_0\,e^{-\lambda x} = f_0\,\frac{1}{\lambda} \tag{1.25}$$

Daraus folgt:

$$f(x) = \lambda\,e^{-\lambda x} \tag{1.26}$$

Verteilungsfunktion: Häufig möchte man die Wahrscheinlichkeit, dass x in einem Intervall $[x_1, x_2]$ liegt, bestimmen (Abb. 1.2). Dazu muss man das entsprechende Integral der Wahrscheinlichkeitsdichte auswerten:

$$p(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)\,dx = \int_{-\infty}^{x_2} f(x)\,dx - \int_{-\infty}^{x_1} f(x)\,dx = F(x_2) - F(x_1) \tag{1.27}$$

Unter anderem kann man hier auch sehen, dass die Wahrscheinlichkeit, einen ganz bestimmten Wert von x zu erhalten, Null ist, weil die Fläche über einem Punkt Null ist. Das bestimmte Integral

$$F(x) = \int_{-\infty}^{x} f(\xi)\,d\xi \tag{1.28}$$

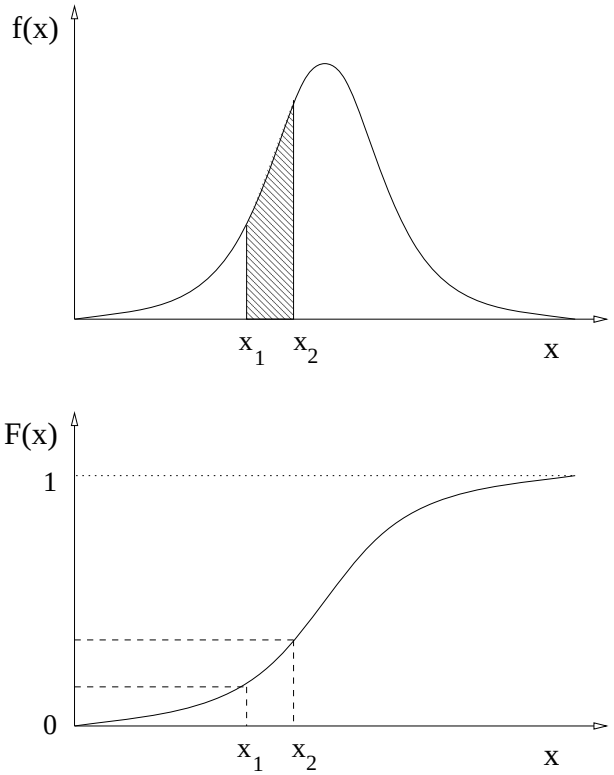


Abbildung 1.2: Wahrscheinlichkeitsdichte (oben) und dazugehörige Verteilungsfunktion (unten).

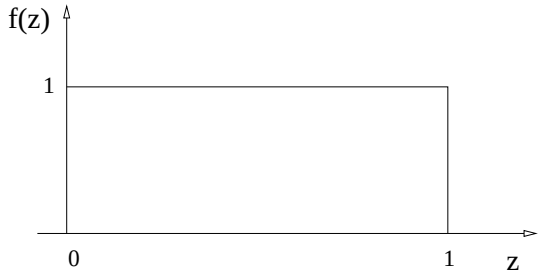


Abbildung 1.3: Wahrscheinlichkeitsdichte einer zwischen 0 und 1 gleichverteilten Variablen.

nennt man die (kumulative) Verteilungsfunktion zu $f(x)$. Der Funktionswert $F(x_0)$ entspricht der Wahrscheinlichkeit, dass x kleiner als x_0 ist:

$$F(x_0) = p(x < x_0). \tag{1.29}$$

Bei diskreten Variablen ergibt sich die Verteilungsfunktion entsprechend:

$$P(n) = \sum_{k=0}^{n} p(k) \tag{1.30}$$

Für wichtige Verteilungen sind Wahrscheinlichkeitsdichte und Verteilungsfunktion in Statistikbüchern tabelliert zu finden.

Die Zuordnung

$$x \to F(x) \tag{1.31}$$

bildet die Zufallsvariable x auf eine gleichverteilte Variable $z = F(x)$ zwischen 0 und 1 ab (Abb. 1.3). Das sieht man wie folgt: Wenn z eine gleichverteilte Variable ist, die aber die gleiche Wahrscheinlichkeit um den Punkt z wie um x beschreibt, muss gelten:

$$dp(x) = f(x)dx = dz = dp(z) \tag{1.32}$$

Der Bezug zu der Verteilungsfunktion ergibt sich dann durch Integration beider Seiten in (1.32):

$$F(x) = \int_{-\infty}^{x} f(\xi)d\xi = \int_{0}^{z} d\zeta = z \tag{1.33}$$

Die Normierung von $f(x)$ stellt sicher, dass z im Intervall [0,1] liegt.

Erzeugung von Zufallsvariablen: Computerprogramme haben in der Regel Zugang zu Zufallszahlengeneratoren, die Zufallszahlen im Intervall [0,1] liefern. Wenn die zu der Dichte f gehörende Verteilungsfunktion F eine analytisch invertierbare Funktion ist, ist es besonders einfach, die Zufallsvariable x entsprechend der Dichte $f(x)$ zu würfeln: Man erzeugt sich gleichverteilte Zufallszahlen $z_i,\ i = 1, \ldots, n$ und bestimmt daraus die x_i:

$$F(x_i) = z_i \qquad \Rightarrow \qquad x_i = F^{-1}(z_i) \tag{1.34}$$

Beispiel: Wir wollen die Variable t mit der Wahrscheinlichkeitsdichte

$$f(t) = \lambda e^{-\lambda t}, \tag{1.35}$$

erzeugen. Dazu ordnen wir t der gleichverteilten Variablen z zu:

$$z = \int_{0}^{t} f(\tau)d\tau = 1 - e^{-\lambda t}. \tag{1.36}$$

Die Umkehrung ergibt:

$$t = \frac{1}{\lambda} \ln \frac{1}{1-z}. \tag{1.37}$$

Man sieht, dass zum Beispiel $z = 0$ auf $t = 0$ und $z = 1$ auf $t = \infty$ abgebildet wird.

1.2.2 Erwartungswerte

Eine Funktion $g(x)$ von der Zufallsvariablen x mit der Wahrscheinlichkeitsdichte $f(x)$ hat den Erwartungswert:

$$E(g(x)) = \langle g(x) \rangle = \int_{-\infty}^{+\infty} g(x) f(x) dx \tag{1.38}$$

Entsprechend gilt für den Erwartungswert einer Funktion $q(n)$ der diskreten Variablen n mit der Wahrscheinlichkeitsverteilung $p(n)$:

$$E(q(n)) = \langle q(n) \rangle = \sum_{n=0}^{\infty} q(n) p(n) \tag{1.39}$$

Die Bildung des Erwartungswertes ist eine lineare Operation:

$$E(a \cdot g(x) \; + \; b \cdot h(x)) = a \cdot E(g(x)) \; + \; b \cdot E(h(x)) \tag{1.40}$$

Im Folgenden behandeln wir spezielle Erwartungswerte, die für die Beschreibung von Verteilungen wichtig sind.

Mittelwert: Der Erwartungswert der Zufallsvariablen x selbst, heisst der Mittelwert der Verteilung:

$$\mu = E(x) = \int_{-\infty}^{+\infty} x \, f(x) dx \tag{1.41}$$

Zum Beispiel ergibt sich für das Zerfallsgesetz

$$f(t) = \lambda e^{-\lambda t}, \tag{1.42}$$

eine mittlere Lebensdauer $\langle t \rangle = 1/\lambda$.

Varianz: Der Erwartungswert der quadratischen Abweichung vom Mittelwert heisst mittlere quadratische Abweichung oder Varianz:

$$\sigma^2 = E((x-\mu)^2) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx \tag{1.43}$$

Die Wurzel aus der Varianz, σ, heisst Standardabweichung. Für die praktische Berechnung der Varianz ist folgende Relation nützlich:

$$\sigma^2 = E((x-\mu)^2) = E(x^2 - 2\mu x + \mu^2) = E(x^2) - 2\mu E(x) - \mu^2 = E(x^2) - \mu^2 \tag{1.44}$$

Dabei ist die Linearität des Operators E und $\mu = E(x)$ benutzt worden.

Momente einer Verteilung: Allgemein nennt man die Erwartungswerte von Potenzen von x oder $x - \mu$ Momente der Verteilung:

$$\begin{array}{ll} \mu'_n = E(x^n) & \mathrm{n-tes\ algebraisches\ Moment} \\ \mu_n = E((x-\mu)^n) & \mathrm{n-tes\ zentrales\ Moment} \end{array} \tag{1.45}$$

Spezielle Momente:

- μ_1' = Mittelwert,

- μ_2 = Varianz

- $\beta = \mu_3/\sigma^3$ = Schiefe (=0 für symmetrische Verteilungen)

Mittelwert, Varianz und Schiefe werden benutzt, um Verteilungen zu charakterisieren. Häufig sind diese Größen Parameter von speziellen Verteilungen, die experimentell zu bestimmen sind. Zum Beispiel ist die Gaussverteilung durch Mittelwert und Varianz gegeben; die Wahrscheinlichkeitsverteilung für einen Zerfall nach (1.42) ist durch die mittlere Zerfallszeit $\tau = 1/\lambda$ gegeben.

Eine Wahrscheinlichkeitsdichte kann nach Momenten entwickelt werden, entsprechend einer Taylor-Entwicklung.

Charakteristische Funktion Die charakteristische Funktion einer Wahrscheinlichkeitsdichte ist deren Fourier-Transformierte, was dem Erwartungswert einer komplexen Exponentialfunktion entspricht:

$$\phi(t) = E(e^{itx}) = \int_{-\infty}^{+\infty} e^{itx} f(x) dx; \tag{1.46}$$

entsprechend für diskrete Verteilungen:

$$\phi(t) = E(e^{itx}) = \sum_{0}^{+\infty} e^{itx} p(k). \tag{1.47}$$

Die Eigenschaften einer Fourier-Transformation können vorteilhaft für Rechnungen mit Verteilungen genutzt werden (zum Beispiel wird die Berechnung von Momenten dadurch sehr erleichtert). Allerdings wollen wir es hier im wesentlichen bei der Erwähnung charakteristische Funktionen belassen und im Folgenden auf deren Einsatz verzichten.

1.2.3 Wahrscheinlichster Wert und Median

Zur Charakterisierung von Verteilungen werden auch andere Größen herangezogen:

Wahrscheinlichster Wert: Bei diesem Wert der Variablen hat die Wahrscheinlichkeitsdichte ein Maximum.

Median: Bei diesem Wert der Variablen hat die Verteilungsfunktion gerade 0.5 erreicht, $F(x_m) = 0.5$. Eine Verallgemeinerung sind Quantile, bei der die Verteilungsfunktion einen bestimmten Wert erreicht, zum Beipiel 0.9 (benutzt zur Angabe von Vertrauensbereichen).

Bei asymmetrischen Verteilungen fallen Mittelwert, wahrscheinlichster Wert und Median nicht zusammen.

1.2.4 Stichproben und Schätzwerte

Bei einer Messung entnimmt man meistens der Gesamtheit aller möglichen Werte einer oder mehrerer Zufallsvariablen eine endliche Stichprobe (die Gesamtheit kann endlich oder unendlich sein).

Beispiel: Eine Länge x wird n-mal gemessen. Die Messwerte $x_1, \ldots, x_n$ sind eine Stichprobe aus den unendlich vielen möglichen Messungen (Abb. 1.4).

Abbildung 1.4:

Eine Stichprobe benutzt man dann, um auf das Verhalten der Zufallsvariablen zurückzuschließen. Dabei reduziert man die Daten auf wesentliche Informationen, die dann Rückschlüsse auf die ursprünglichen Verteilungen, zum Beispiel über die Bestimmung der Parameter der Verteilungen, erlauben. Die aus einer Stichprobe gewonnenen Parameter von Verteilungen nennt man Schätzwerte. Schätzwerte von Erwartungswerten werden häufig durch Mittelung der entsprechenden Größe über die Stichprobe gebildet.

Schätzung der Verteilung: Die Wahrscheinlichkeitsdichte kann nur gemittelt über endliche Intervalle der Zufallsvariablen geschätzt werden. Falls es sich um eine kontinuierliche Variable handelt, wird man Messwerte in endliche Intervalle ('Bins') zusammenfassen, 'histogrammieren'.

Beispiel: Bei der Messung des Zerfalls einer radioaktiven Probe seien N_0 Zerfälle mit jeweils $N(t_i)$ Zerfällen in Zeitintervallen Δt um t_i gemessen worden (Abb. 1.5). Eine Abschätzung der Wahrscheinlichkeitsdichte erhält man aus:

$$\hat{f}(t_i) = \frac{N(t_i)}{N_0} \tag{1.48}$$

Wie man leicht sieht, ist die Normierung

$$\sum_i \hat{f}(t_i) = 1 \tag{1.49}$$

sichergestellt.

Mittelwert: Den Schätzwert für den Mittelwert einer Verteilung erhält man durch Mittelung der Messwerte. Aus n Messwerten $x_1, \ldots, x_n$ erhält man als Schätzwert $\bar{x}$ des Erwartungswertes $\langle x \rangle$:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1.50}$$

Beispiel: In dem vorigen Beispiel würde man die mittlere Zerfallszeit $\tau = 1/\lambda$ (nach Gleichung (1.42)) durch Mittelung über die Messintervalle bestimmen:

$$\hat{\tau} = \frac{1}{N_0} \sum_i t_i \, N(t_i) = \sum_i t_i \, \hat{f}(t_i). \tag{1.51}$$

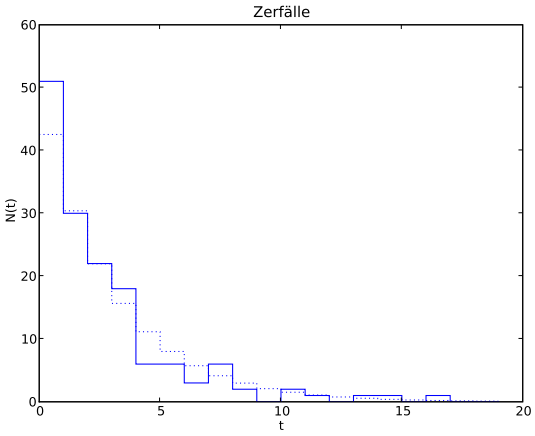


Abbildung 1.5: Histogramm der Anzahl von Zerfällen pro Zeitinterval. Die Messwerte (durchgezogen) und die exakte Verteilung (gepunktet) werden verglichen.

Varianz: Als Schätzwert der Varianz definiert man:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{1.52}$$

Mit der Division durch $n-1$ statt n erhält man eine bessere Abschätzung, wie wir später noch bei der Diskussion der optimalen Eigenschaften von Schätzwerten sehen werden.

1.3 Simulation von Verteilungen

Computer-Simulationen sind ein wichtiges Hilfsmittel in verschiedensten Bereichen geworden, wie zum Beispiel in Wissenschaft, Technik, Wirtschaft. So werden Wetter- und Klimamodelle, Optimierungen von Auto- und Flugzeugformen, Bestimmung von Nachweiswahrscheinlichkeiten von Teilchenreaktionen oder Lösungen von komplizierten Integralen mit Simulationen nach dem Zufallsprinzip (Monte-Carlo-Methode) berechnet. Die Idee ist, repräsentative Stichproben zu erzeugen, die von einem Satz Zufallsvariabler abhängen. Für jedes erzeugte 'Ereignis' werden die Variablen entsprechend ihrer Wahrscheinlichkeitsverteilung 'gewürfelt'.

In der Regel geht man von einem Zufallszahlengenerator aus, der bei jedem Aufruf eine neue Zahl z, die im Intervall $[0, 1]$ gleichverteilt ist, zurückgibt. Die Frage ist dann, wie man eine Variable in einem beliebigen Intervall und mit einer beliebigen Verteilung erzeugt.

1.3.1 Umkehrung der Verteilungsfunktion

Eine Methode haben wir bereits in Abschnitt 1.2.1 kennengelernt: Die Verteilungsfunktion $F(x)$ zu einer Wahrscheinlichkeitsdichte ist gleichverteilt zwischen 0 und 1. Wir können also

$$z = F(x) \tag{1.53}$$

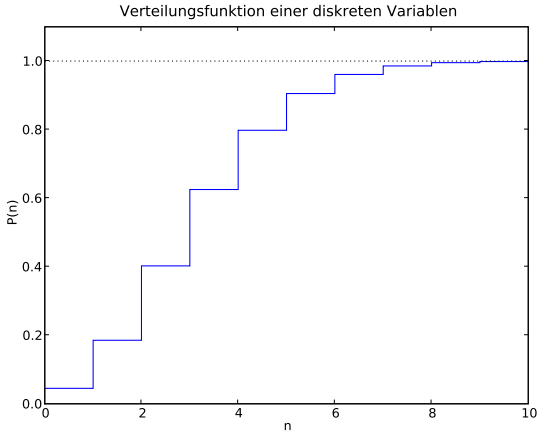


Abbildung 1.6: Verteilungsfunktion einer diskreten Variablen.

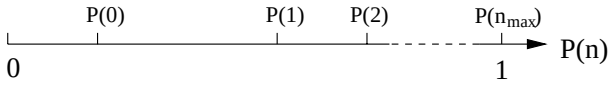


Abbildung 1.7: Abbildung der Verteilungsfunktion einer diskreten Variablen auf das Einheitsintervall.

setzen und erhalten, wenn die Umkehrfunktion F^{-1} existiert, zu jeder gewürfelten Zahl z die entsprechende Zufallszahl x mit der gewünschten Verteilung:

$$x = F^{-1}(z) \tag{1.54}$$

Beispiel: Ein Beispiel ist bereits für die Lebensdauerverteilung gegeben worden (Gleichungen (1.35 - 1.37)).

Bei diskreten Verteilungen ist die Verteilungsfunktion eine Stufenfunktion (Abb. 1.6):

$$P(n) = \sum_{k=0}^{n} p(k). \tag{1.55}$$

Wenn man die Werte $P(0)$, $P(1)$, ..., $P(n)$ als Einteilung des Intervalles $[0, 1]$ benutzt (Abb. 1.7) entspricht der Länge jedes Abschnitts gerade eine Wahrscheinlichkeit $p(k)$, beginnend bei $p(0)$ und endend bei $p(n)$. Einer gewürfelten Zufallszahl z ordnet man dann die diskrete Zufallszahl k zu, wenn gilt:

$$\begin{aligned} P(k-1) &< z \leq P(k), & k \neq 0 \\ 0 &\leq z \leq P(0), & k = 0 \end{aligned} \tag{1.56}$$

Wenn man zu der Verteilungsfunktion einer kontinuierlichen Variablen x keine Umkehrfunktion findet, kann man die Variable diskretisieren, zum Beispiel in Intervalle Δx um diskrete Werte x_i aufteilen zu denen Wahrscheinlichkeiten $f(x_i) \cdot \Delta x$ gehören (siehe das Beispiel in Abb. 1.5). Verteilungen, die sich bis $+\infty$ oder $-\infty$ ausdehnen, aber in der Regel mit fallenden Wahrscheinlichkeiten, schneidet man bei geeigneten Grenzen ab. Als Maß benutzt man dafür häufig die Standardabweichung σ (zum Beipiel $\pm 5\sigma$ um den Mittelwert).

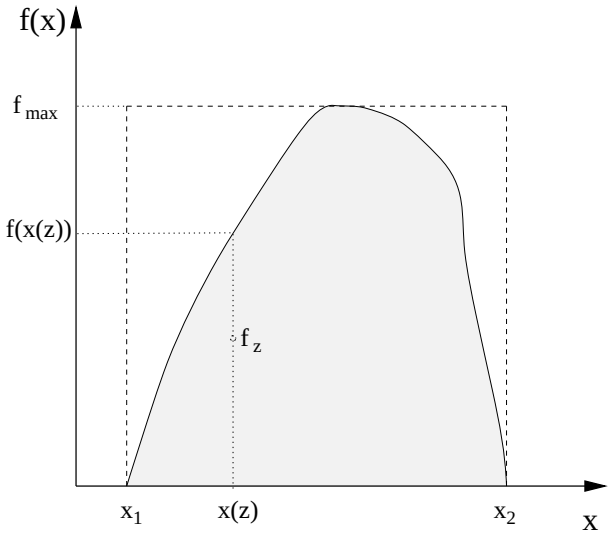


Abbildung 1.8: Zur Erklärung der 'Hit and Miss' Methode.

1.3.2 'Hit and Miss' Methode

Wenn die Wahrscheinlichkeitsdichte sehr unübersichtlich wird, insbesondere bei Abhängigkeit von mehreren Variablen oder wenn man davor zurückschreckt, analytische Berechnungen zu machen, kann man Ereignisse nach der 'Hit and Miss' Methode erzeugen.

Sei x eine Zufallsvariable mit der Wahrscheinlichkeitsdichte $f(x)$ (Abb. 1.8). Sowohl x als auch $f(x)$ sollte in einem endlichen Intervall liegen:

$$\begin{array}{ccccc} x_1 & \leq & x & \leq & x_2 \\ 0 & \leq & f(x) & \leq & f_{max} \end{array} \tag{1.57}$$

Falls das nicht gegeben ist, kann man sich häufig auf relevante Bereiche beschänken, siehe oben. Der 'Hit and Miss' Algorithmus lautet dann:

(i) Erzeuge x gleichverteilt im Intervall $[x_1, x_2]$;

(ii) erzeuge einen Wert f_z gleichverteilt im Intervall $[0, f_{max}]$;

(iii) akzeptiere x falls $f_z \leq f(x)$;

(iv) wiederhole.

Es werden also Punkte $x(z), f(x(z))$ gleichverteilt in der Box (1.57) erzeugt. Ein Punkt wird als Treffer gezählt, wenn er unterhalb der Kurve $f(x)$ liegt. Die so erzeugten Treffer x folgen der Verteilung $f(x)$ normiert auf das eventuell beschränkte Intervall.

Die benötigte Transformation einer Gleichverteilung im Einheitsintervall $[0, 1]$ auf eine beliebige Gleichverteilung zum Beispiel in $[x_1, x_2]$ ergibt sich aus der entsprechenden Umkehrfunktion:

$$z = \frac{\int_{x_1}^{x} dx}{\int_{x_1}^{x_2} dx} = \frac{x - x_1}{x_2 - x_1} \implies x = x_1 + z \cdot (x_2 - x_1) \tag{1.58}$$

Die 'Hit and Miss' Methode ist nicht sehr effizient, wenn sehr große Werte der Wahrscheinlichkeitsdichte $f(x)$ in sehr kleinen x-Intervallen auftreten ($f(x) \to \infty$ ist möglich, solange das Integral über $f(x)$ endlich bleibt). Dann benutzt man andere Verfahren, die wir teilweise in einem späteren Kapitel besprechen werden.

Kapitel 2

Spezielle Verteilungen einer Variablen

In diesem Kapitel werden wir einige häufig benutzte Verteilungen, die von einer Variablen abhängen, vorstellen.

2.1 Binomial-Verteilung

Binomial-Verteilungen treten auf, wenn man die betrachteten Ereignisse in zwei Klassen mit den Eigenschaften A und $\bar{A}$ zerlegen kann, die mit komplementären Wahrscheinlichkeiten auftreten:

Eigenschaft	Wahrscheinlichkeit
A	p
$\bar{A}$	1-p

Wie groß ist die Wahrscheinlichkeit W_k^n, bei n Ereignissen k mit der Eigenschaft A zu erhalten?

Beispiele:

- Aus einer Übungsaufgabe: Die Wahrscheinlich ein Ei zu finden ist p. Wie groß ist die Wahrscheinlichkeit bei n versteckten Eiern k zu finden. Die Kenntnis der entsprechenden Wahrscheinlichkeitsverteilung wird uns helfen, den Fehler in der Abschätzung der Effizienz zu bestimmen.

- Wie groß ist die Wahrscheinlichkeit, dass sich in einem System mit n Spins k in Richtung eines vorgegebenen Magnetfeldes einstellen? Die Wahrscheinlichkeit für jeden einzelnen Spin ist abhängig von Temperatur und Feldstärke: $p = f(T, B)$.

- Es seien n Teilchen in einer Box mit Volumen V. Wie groß ist die Wahrscheinlichkeit, k davon in einem Teilvolumen V_1 zu finden? Die Wahrscheinlichkeit für jedes einzelne Teilchen ist offensichtlich $p = V_1/V$.

- Das Galton-Brett ist eine Anordnung von Nägeln wie in Abb. 2.1 gezeigt. Man setzt eine Kugel auf den obersten Nagel, von dem sie zufällig nach rechts oder

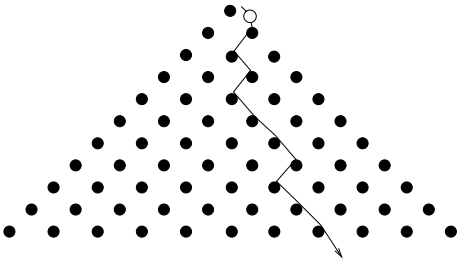


Abbildung 2.1: Galton-Brett.

links auf einen Nagel der nächsten Reihe fällt und so weiter. Wenn alles schön symmetrisch ist, fällt die Kugel jeweils mit gleicher Wahrscheinlichkeit nach links oder rechts: $p = 0.5$.

- Am Computer kann man dem Galton-Brett auch einen beliebigen Parameter p zuordnen: Man würfelt n-mal im Intervall $[0, 1]$ und ermittelt die Anzahl k, für die die Zufallszahl kleiner als p ist (das ist zum Beispiel, wie häufig die Kugel nach links gefallen ist).

Herleitung der Binomial-Verteilung: Es gibt verschiedene Kombinationen, in einer Gesamtheit von n Ereignissen k mit der Eigenschaft A zu erhalten, die sich durch die Reihenfolge des Auftretens von A unterscheiden. Zum Beispiel gibt es für $n = 3$ und $k = 2$ offensichtlich 3 mögliche Kombinationen:

$$\begin{array}{ccc} 1 & 2 & 3 \\ \hline \\ A & A & \bar{A} \\ A & \bar{A} & A \\ \bar{A} & A & A \end{array} \tag{2.1}$$

Jede einzelne Kombination zu festen Zahlen n und k hat die gleiche Wahrscheinlichkeit. Diese ergibt sich als Produkt der Wahrscheinlichkeiten, jeweils für ein bestimmtes Ereignis die Eigenschaft A oder $\bar{A}$ zu haben. Zum Beispiel würde man in der ersten Zeile von (2.1) $p \cdot p \cdot (1-p) = p^2(1-p)$ erhalten. Allgemein ergibt sich:

$$p^k \cdot (1-p)^{n-k}. \tag{2.2}$$

Um dieses Produkt der Wahrscheinlichkeiten zu bilden, muss die Wahrscheinlichkeit für das Auftreten von A unabhängig davon sein, wie häufig A bereits gezählt wurde. Zum Beipiel müssen bei einer Ziehung aus einer endlichen Anzahl von schwarzen und weissen Kugeln die Kugeln immer wieder zurückgelegt werden, damit die Wahrscheinlichkeiten für schwarz und weiss sich nicht ändern.

Die Wahrscheinlichkeit für das Auftreten irgendeiner Kombination mit k-mal der Eigenschaft A ist die Summe der Wahrscheinlichkeiten der einzelnen Kombinationen (in (2.1) also die Summe der Wahrscheinlichkeiten der 3 Zeilen, das ist $3\,p^2(1-p)$). Da jede dieser Wahrscheinlichkeiten gleich ist, muss man also nur die Anzahl der möglichen Kombinationen bestimmen.

Um nun allgemeiner die Anzahl der Kombinationen mit k-mal der Eigenschaft A zu bestimmen, beginnt man damit, zunächst k unterscheidbare Ereignisse $A_1, \ldots, A_k$ auf n Stellen zu verteilen. In (2.1) würden sich die beiden A in einer Spalte durch einen Index 1 und 2 ($A_1,\ A_2$) unterscheiden, dessen Vertauschung dann zu einer Verdoppelung der Möglichkeiten führt (von 3 auf 6). Um nun die Anzahl der Anordnungen bei k Ereignissen zu bestimmen, kann man die Ereignisse nacheinander auf die jeweils noch freien Plätze verteilen:

A_1	n	Möglichkeiten (alle Plätze sind noch frei),
A_2	$n-1$	Möglichkeiten (ein Platz ist bereits mit A_1 besetzt),
	$\ldots$	
A_k	$n-(k-1)$	Möglichkeiten ($k-1$ Plätze sind von A_1 bis A_{k-1} besetzt).

Das sind insgesamt

$$n \cdot (n-1) \ldots \cdot n-(k-1) = \frac{n!}{(n-k)!} \tag{2.3}$$

Möglichkeiten, von der jede aber in $k!$ Anordnungen der A_i auftreten (in (2.1) gibt es für die 2 A-Ereignisse jeweils 2 Permutationen). Da nach der Reihenfolge nicht unterschieden wird, ergibt sich schließlich für die Gesamtzahl der Kombinationen, die Eigenschaft A k-mal auf n Ereignisse zu verteilen:

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!} \tag{2.4}$$

Der Ausdruck $\binom{n}{k}$ beschreibt die Binomialkoeffizienten, die sich bekanntlich mit dem Pascalschen Dreieck darstellen lassen:

n										
0						1				
1					1		1			
2				1		2		1		
3			1		3		3		1	
4		1		4		6		4		1
			.		.		.		.	

$k \rightarrow$

Damit ergibt sich die Binomial-Verteilung:

$$W_k^n = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \tag{2.5}$$

Normierung: Es ist einfach zu sehen, dass die Normierung

$$\sum_{k=0}^{n} W_k^n = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = 1 \tag{2.6}$$

richtig ist, weil die Summe gerade der Formel für $(a+b)^n$ mit $a=p$ und $b=1-p$ entspricht:

$$\sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = (p+1-p)^n = 1^n = 1 \tag{2.7}$$

Mittelwert:

$$\begin{aligned}
\langle k\rangle &= \textstyle\sum_{k=0}^{n} k\cdot W_k^n \\
&= \textstyle\sum_{k=0}^{n} k\cdot \frac{n!}{(n-k)!k!}\cdot p^k\cdot(1-p)^{n-k} \\
&= \textstyle\sum_{k=1}^{n} \frac{n!}{(n-k)!(k-1)!}\cdot p^k\cdot(1-p)^{n-k} \\
&= np\cdot\textstyle\sum_{k=1}^{n} \frac{(n-1)!}{[(n-1)-(k-1)]!(k-1)!}\cdot p^{k-1}\cdot(1-p)^{n-1-(k-1)} \quad \text{mit } n-k=(n-1)-(k-1) \\
&= np\cdot\textstyle\sum_{k'=0}^{n'} \frac{n'!}{(n'-k')!k'!}\cdot p^{k'}\cdot(1-p)^{n'-k'} = np \qquad \text{mit } n'=n-1;\ k'=k-1
\end{aligned} \tag{2.8}$$

Die letzte Zeile benutzt die Normierung der Summe auf 1. Damit ergibt sich für den Mittelwert von k:

$$\langle k\rangle = np \tag{2.9}$$

Zum Beipiel ist für $p = 0.5$ wie zu erwarten $\langle k\rangle = n/2$.

Varianz: Die Varianz ist die mittlere quadratische Abweichung vom Mittelwert, die sich nach (1.44) zerlegen läßt:

$$\sigma^2 = \langle (k-\langle k\rangle)^2\rangle = \langle k^2\rangle - \langle k\rangle^2 \tag{2.10}$$

Der Erwartungswert von k^2 läßt sich ähnlich wie der Mittelwert bestimmen:

$$\begin{aligned}
\langle k^2\rangle &= \textstyle\sum_{k=0}^{n} k^2\cdot W_k^n \\
&= \textstyle\sum_{k=0}^{n} k^2\cdot \frac{n!}{(n-k)!k!}\cdot p^k\cdot(1-p)^{n-k} \\
&= \textstyle\sum_{k=1}^{n} k\frac{n!}{(n-k)!(k-1)!}\cdot p^k\cdot(1-p)^{n-k} \\
&= np\cdot\textstyle\sum_{k'=0}^{n'}(k'+1)\cdot\frac{n'!}{(n'-k')!k'!}\cdot p^{k'}\cdot(1-p)^{n'-k'} \qquad (n'=n-1;\ k'=k-1) \\
&= np\cdot\left[1+\textstyle\sum_{k'=0}^{n'} k'\cdot\frac{n'!}{(n'-k')!k'!}\cdot p^{k'}\cdot(1-p)^{n'-k'}\right] \\
&= np\cdot[1+(n-1)p]
\end{aligned} \tag{2.11}$$

Damit ergibt sich für die Varianz:

$$\sigma^2 = n\,p\,(1-p). \tag{2.12}$$

Bemerkungen: Folgende Eigenschaften der Binomial-Verteilung werden in Abb. 2.2 demonstriert:

1. Die Varianz hat für $p = 0.5$ ein Maximum:

$$\frac{d\sigma^2}{dp} = n\,[(1-p)+(-p)] = 0 \implies p = 0.5 \tag{2.13}$$

2. Die relative Breite wird mit wachsendem n kleiner:

$$\frac{\sigma}{\langle k\rangle} = \frac{\sqrt{n\,p\,(1-p)}}{n\,p} = \sqrt{\frac{1-p}{n\,p}} \sim \frac{1}{\sqrt{n}} \tag{2.14}$$

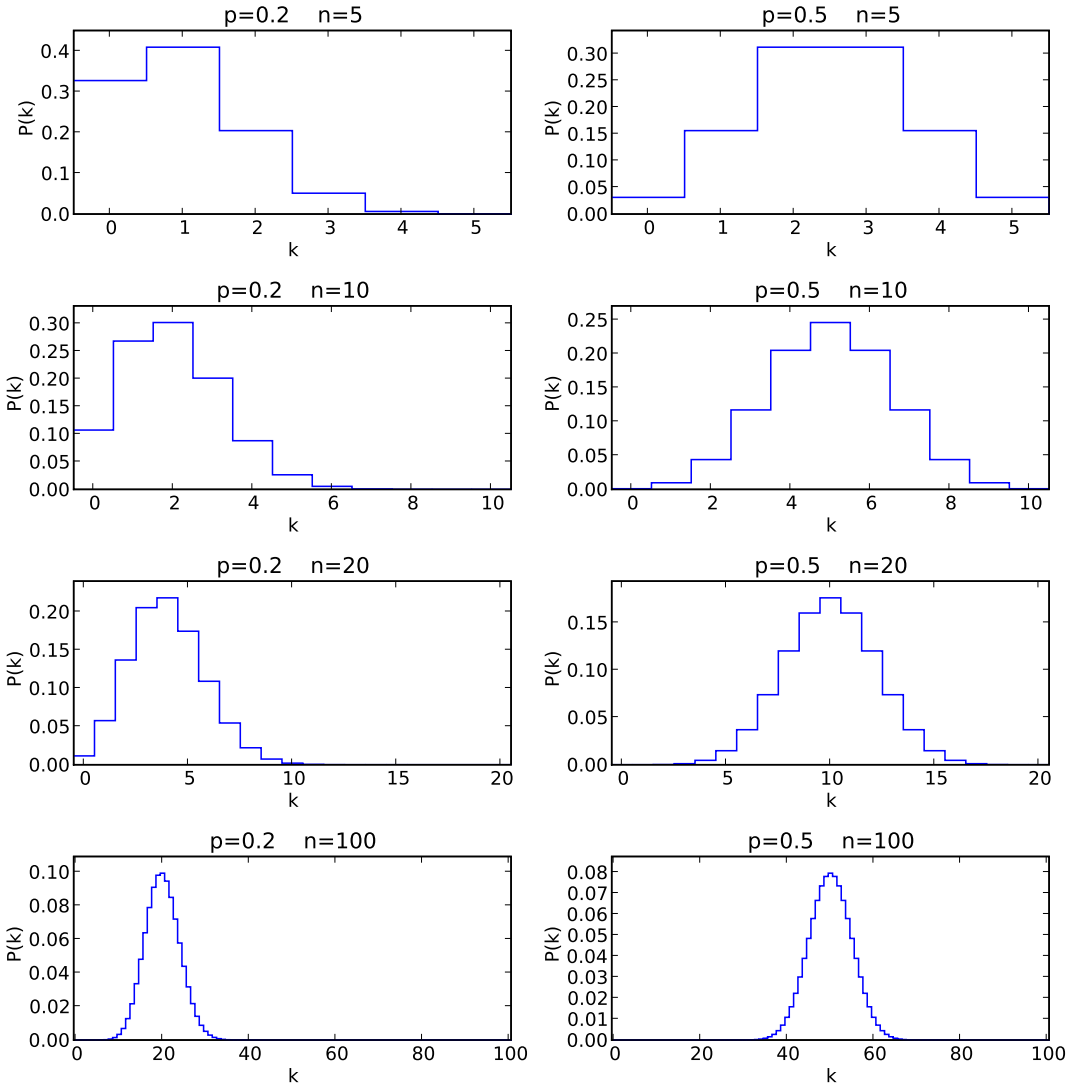


Abbildung 2.2: Beispiele von Binomial-Verteilungen mit verschiedenen Parametern n und p .

3. Für große n und $n\,p$ (p nicht zu klein) nähert sich die Binomial-Verteilung der Normalverteilung mit $\mu = np$ und $\sigma^2 = n\,p\,(1-p)$ an (das ergibt sich aus dem 'Zentralen Grenzwertsatz', siehe Abschnitt 2.6):

$$W_k^n \to W(k;\, n, p) = \frac{1}{\sqrt{2\pi\, np(1-p)}}\, \exp\left(-\frac{(k-np)^2}{2np(1-p)}\right) \tag{2.15}$$

2.2 Multinomial-Verteilung

Die Multinomial-Verteilung ist die natürliche Erweiterung der Definition der Binomial-Verteilung: Gegeben seien l Klassen von Ereignissen A_j ($j = 1, \ldots, l$) mit den Eigenschaften j und den Wahrscheinlichkeiten p_j, die sich gegenseitig ausschliessen und erschöpfend sind:

$$E = \sum_{j=1}^{l} A_j; \qquad A_i \cap A_j = \varnothing \;\; \emptyset \; i \neq j. \tag{2.16}$$

Daraus folgt für die Summe der Wahrscheinlichkeiten aller Klassen:

$$\sum_{j=1}^{l} p_j = 1 \tag{2.17}$$

Die Wahrscheinlichkeit, bei n Ereignissen gleichzeitig k_1 mit der Eigenschaft A_1, k_2 mit der Eigenschaft A_2 ... und k_l mit der Eigenschaft A_l usw. zu erhalten, ist

$$W^n_{k_1,k_2,\ldots,k_l} = n! \prod_{j=1}^{l} \frac{p_j^{k_j}}{k_j!} \tag{2.18}$$

Jedes der n Ereignisse ist jeweils in einer der l Klassen, so dass gilt:

$$\sum_{j=1}^{l} k_j = n. \tag{2.19}$$

Das bedeutet, dass die Faktoren in (2.18) nicht unabhängig voneinander sind. Der vollständige Beweis der Formel (2.18) kann durch Induktion von $l-1$ auf l durchgeführt werden.

Für $l = 2$ erhält man die Binomial-Verteilung wieder ($k_1 = k$; $k_2 = n - k$):

$$W^n_{k_1,k_2} = n! \frac{p_1^{k_1}}{k_1!} \cdot \frac{p_2^{k_2}}{k_2!} = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = W_k^n \tag{2.20}$$

Die Multinomial-Verteilung ist eine Verteilung mit mehreren Variablen (die k_j), die wir eigentlich erst im nächsten Kapitel besprechen. Im Vorgriff geben wir im Folgenden Parameter der Verteilung an, die zum Teil erst später (wie die Kovarianzmatrix) definiert werden.

Normierung: Unter Berücksichtigung der Bedingungen (2.17) und (2.19) ergibt sich für die Normierung:

$$\sum_{k_1=0}^{n} \sum_{k_2=0}^{n-k_1} \dots \sum_{k_{l-1}=0}^{n-k_1-k_2-\dots k_{l-2}} W^n_{k_1,k_2,\dots,k_l} = 1 \qquad \text{mit } k_l = n - \sum_{j=1}^{l-1} k_j \text{ und } p_l = 1 - \sum_{j=1}^{l-1} p_j \tag{2.21}$$

Mittelwert: Der Mittelwert jeder einzelnen Variablen ist:

$$\langle k_j \rangle = np_j \qquad (j = 1, \dots, l) \tag{2.22}$$

Varianz: Die Varianzen der einzelnen Variablen ergeben sich entsprechend der Binomial-Verteilung:

$$\sigma_i^2 = np_i(1 - p_i) \tag{2.23}$$

Bei mehreren Variablen treten auch Kovarianzen auf, die Korrelationen beschreiben (siehe Kapitel 3):

$$\mathrm{cov_{ij}} = -np_ip_j \tag{2.24}$$

Das Minuszeichen bedeutet eine negative Korrelation zwischen k_i, k_j (eine Änderung einer Variablen bewirkt tendentiell eine Änderung der anderen Variablen in die entgegengesetzte Richtung).

Beispiele:

- Die Häufigkeit der Buchstaben in Texten, im allgemeinen $p_i \neq p_j$, wird zur Analyse von Texten und Sprachen bestimmt.
- In Experimenten der Teilchenphysik treten in der Regel 5 Arten geladener, stabiler Teilchen mit unterschiedlichen Häufigkeiten auf (Protonen, Pionen, Kaonen, Elektronen, Myonen). Die Analyse der Häufigkeitsverteilung benötigt man zur Identifikation der Teilchen (siehe späteres Kapitel zur Entscheidung über Hypothesen).

2.3 Poisson-Verteilung

Der Grenzfall einer Binomialverteilung mit einer sehr großen Zahl von möglichen Ereignissen, die aber jeweils eine sehr kleine Wahrscheinlichkeit haben, führt zu der Poisson-Verteilung:

$$\lim_{\substack{n\to\infty \\ p\to 0}} W^n_k = P^\lambda_k \qquad (n \cdot p = \lambda \text{ endlich}) \tag{2.25}$$

Bei dem Grenzübergang zu sehr großen Zahlen n und sehr kleinen Wahrscheinlichkeiten p soll der Erwartungswert von k,

$$\langle k \rangle = \lambda = n \cdot p, \tag{2.26}$$

endlich bleiben.

Beispiele:

- Radioaktiver Zerfall: Die Zahl n der radioaktiven Kerne ist bei einer Probe meistens von der Größenordnung der Loschmidt-Zahl, also sehr groß. Die Wahrscheinlichkeit, daß einer dieser Kerne in einem festen Zeitintervall Δt zerfällt, ist dagegen sehr klein, aber die mittlere Zerfallsrate λ ist endlich.
- Die Anzahl der Sterne, die man in einem gegebenen Ausschnitt eines Teleskops bei einer bestimmten Auflösung beobachtet, hat einen bestimmten Mittelwert λ, der klein ist gegen die Gesamtzahl der Sterne. Bei einer Himmelsdurchmusterung erwartet man Fluktuationen entsprechend einer Poisson-Verteilung. Abweichungen, eventuell als Funktion der Ausschnittgröße, können auf kosmische Strukturen hinweisen.
- Die Anzahl der Gasatome in einem Volumen von der Größenordnung einiger Atomvolumina ist Poisson-verteilt.
- Die Zahl der jährlichen tödlichen Unfälle durch Pferdetritte in der Preussischen Armee ist Poisson-verteilt.
- Die Anzahl der Druckfehler auf einer Seite eines Buches ist Poisson-verteilt.

Die Poisson-Verteilung kann durch Ausführung des Grenzüberganges (2.25) aus der Binomialverteilung abgeleitet werden. Mit $\lambda = n \cdot p$ beziehungsweise $p = \lambda/n$ gilt:

$$\begin{aligned} W_k^n &= \tbinom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \\ &= \tfrac{n!}{(n-k)!\,k!} \cdot \left(\tfrac{\lambda}{n}\right)^k \cdot \left(1-\tfrac{\lambda}{n}\right)^{n-k} \\ &= \tfrac{\lambda^k}{k!} \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda} \text{ für } n\to\infty} \underbrace{\frac{n(n-1)\ldots(n-k-1)}{n^k \left(1-\frac{\lambda}{n}\right)^k}}_{\to 1 \text{ für } n\to\infty} \end{aligned} \tag{2.27}$$

Damit ergibt sich für den Limes $n \to \infty$ die Poisson-Verteilung:

$$P_k^\lambda = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \tag{2.28}$$

Ausgehend von

$$P_0^\lambda = e^{-\lambda} \tag{2.29}$$

ist vor allem zum Programmieren folgende Rekursionsformel nützlich:

$$P_{k+1}^\lambda = P_k^\lambda \cdot \frac{\lambda}{k+1} \tag{2.30}$$

Normierung: Die Poisson-Verteilung (2.28) ist richtig normiert:

$$\sum_{k=0}^{\infty} P_k^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot e^{-\lambda} = e^{-\lambda} \underbrace{\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}}_{e^\lambda} = e^{-\lambda} \cdot e^{\lambda} = 1 \tag{2.31}$$

Mittelwert: Nach Konstruktion ist der Erwartungswert von k gleich λ:

$$\langle k \rangle = \lambda, \tag{2.32}$$

was sich durch explizite Berechnung bestätigen läßt:

$$\langle k \rangle = \sum_{k=0}^{\infty} k \, \frac{\lambda^k}{k!} \cdot e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \cdot e^{-\lambda} = \lambda. \tag{2.33}$$

Varianz: Ausgehend von der Varianz für eine Binomial-Verteilung $\sigma^2 = n\,p\,(1-p)$ erhält man mit dem Grenzübergang $p \to 0$, wobei $\lambda = np$ endlich bleibt:

$$\sigma^2 = n\,p = \lambda. \tag{2.34}$$

Die Standardabweichung ist dann

$$\sigma = \sqrt{\lambda}. \tag{2.35}$$

Breite und Mittelwert der Verteilung sind also eng miteinander verknüpft.

Häufig entnimmt man als Stichprobe einer Poisson-Verteilung nur einen einzigen Wert, zum Beispiel die Zählrate N von Kernzerfällen in einem Zeitintervall. Dann ist N der beste Schätzwert für die mittlere Zerfallsrate λ und als Fehler wird der Schätzwert für die Standardabweichung benutzt:

$$\hat{\sigma} = \sqrt{N}. \tag{2.36}$$

Allerdings muss man bei der Weiterverarbeitung von Daten vorsichtig sein, weil bei Fluktuationen von N nach unten ein kleinerer Fehler folgt als bei Fluktuationen nach oben (siehe Diskussion bei 'Likelihood-Methode').

Bemerkungen: Folgende Eigenschaften sind charakteristisch für die Poisson-Verteilung (siehe Abb. 2.3):

1. Die Varianz ist gleich dem Mittelwert.
2. Für kleine Mittelwerte λ (nahe 1) ergibt sich eine asymmetrische Verteilung.
3. Für wachsende λ wird die Verteilung immer symmetrischer und nähert sich einer Gauss-Verteilung mit Mittelwert und Varianz λ (das ergibt sich wieder aus dem 'Zentralen Grenzwertsatz', siehe Abschnitt 2.6):

$$P_k^{\lambda} \to P(k;\,\lambda) = \frac{1}{\sqrt{2\pi\lambda}} \exp\left(-\frac{(k-\lambda)^2}{2\lambda}\right) \tag{2.37}$$

2.4 Gleichverteilung

Der einfachste, aber durchaus wichtige, Fall einer Wahrscheinlichkeitsverteilung einer kontinuierlichen Variablen ist die Gleichverteilung:

$$f(x) = c = const \tag{2.38}$$

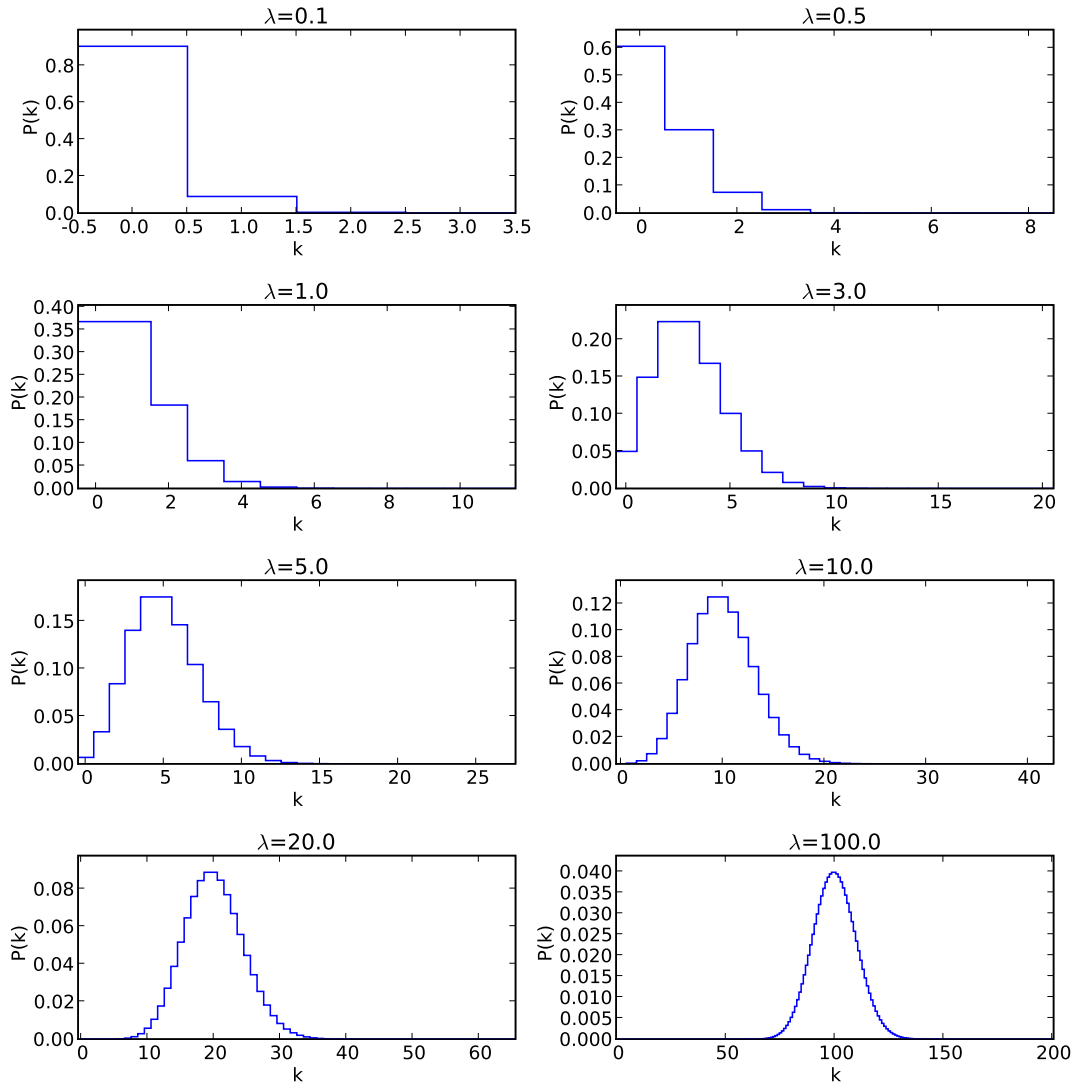
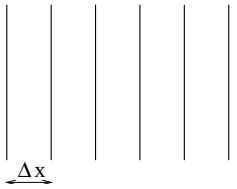


Abbildung 2.3: Beispiele von Poisson-Verteilungen mit verschiedenen Parametern λ

.

Beispiele:

- Der Winkel eines Uhrzeigers nimmt mit gleicher Wahrscheinlichkeit einen Wert zwischen 0° und 360° an.
- Viele Detektoren für Strahlung haben eine Streifenstruktur, die eine Koordinate innerhalb einer Streifenbreite festlegt:



  Bei homogener Einstrahlung ist die Koordinate des Auftreffens des Teilchens innerhalb eines Streifens gleichverteilt.
- Rundungsfehler sind gleichverteilt in dem Rundungsintervall.

Normierung:

$$1 = \int_{x_1}^{x_2} f(x)\,dx = \int_{x_1}^{x_2} c\,dx = c\,(x_2 - x_1) = c\,\Delta x \implies c = \frac{1}{\Delta x} \tag{2.39}$$

Zum Beispiel ergibt sich für eine Uhr:

$$f(\varphi) = \frac{1}{360^\circ} \tag{2.40}$$

Mittelwert:

$$\bar{x} = \langle x \rangle = \frac{1}{\Delta x} \int_{x_1}^{x_2} x\,dx = \frac{1}{2}\frac{x_2^2 - x_1^2}{x_2 - x_1} = \frac{x_1 + x_2}{2} \tag{2.41}$$

Varianz:

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2 = \frac{1}{3}\frac{x_2^3 - x_1^3}{x_2 - x_1} - \left(\frac{1}{2}\frac{x_2^2 - x_1^2}{x_2 - x_1}\right)^2 = \frac{(\Delta x)^2}{12} \tag{2.42}$$

Die Standardabweichung ist dann

$$\sigma = \frac{\Delta x}{\sqrt{12}}. \tag{2.43}$$

Das heisst, die Standardabweichung ist um eine Faktor $\sqrt{12} \approx 3.5$ besser als das Raster einer Messung.

Verteilungsfunktion: Die Verteilungsfunktion steigt linear mit x an:

$$F(x) = \frac{1}{\Delta x} \int_{x_1}^{x} d\xi = \frac{x - x_1}{\Delta x} \tag{2.44}$$

2.5 Normalverteilung

Die in der Statistik am häufigsten benutzte Verteilung ist die Gauss- oder Normalverteilung. Wir haben bereits gesehen, dass diese Verteilung aus den Binomial- und Poisson-Verteilungen im Grenzfall großer Zahlen (n bzw. λ) folgt. Wir werden weiter unten den 'zentralen Grenzwertsatz' besprechen, der solche Grenzübergänge noch allgemeiner behandelt.

Eine Normalverteilung ergibt sich, wenn viele kleine Änderungen ϵ_i aufsummiert werden. Anschaulich kann man sich das zum Beispiel anhand des Galton-Brettes (Abb. 2.1) klar machen: Die Kugel entscheidet n-mal, ob Sie links oder rechts von einem Nagel fällt, entsprechend einem Versatz um $\epsilon_i = \pm\Delta\epsilon$. Die Verteilung der Auftrefforte unter dem Brett $x = \sum_{i=1}^{n} \epsilon_i$ nähert sich einer Normalverteilung im Grenzfall großer n.

Die Normalverteilung $N(\mu, \sigma)$ ist durch die beiden Parameter Mittelwert μ und Varianz σ^2 gegeben:

$$f(x) = f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.45}$$

Normierung: Die Normierung wird durch den Faktor $(\sqrt{2\pi}\sigma)^{-1}$ sichergestellt, was sich mit folgendem bestimmten Integral ergibt:

$$\int_{-\infty}^{\infty} e^{-ax^2}\, dx = \sqrt{\frac{\pi}{a}} \tag{2.46}$$

Mittelwert: Der Mittelwert ergibt sich aus:

$$\langle x \rangle = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \tag{2.47}$$

Zur Berechnung des Integrals setzt man $x = (x-\mu)+\mu$ und erhält damit die beiden Integrale:

$$\langle x \rangle = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu) \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx}_{=0} + \underbrace{\mu \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx}_{=1} = \mu \tag{2.48}$$

Das linke Integral verschwindet, weil sich die Beiträge für $x - \mu < 0$ und die für $x - \mu > 0$ gerade aufheben.

Varianz: Die Varianz ergibt sich mit Hilfe eines weiteren bestimmten Integrals:

$$\int_{-\infty}^{\infty} x^2 e^{-ax^2}\, dx = \frac{1}{2a}\sqrt{\frac{\pi}{a}} \tag{2.49}$$

Damit erhält man:

$$\langle (x-\mu)^2 \rangle = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \sigma^2. \tag{2.50}$$

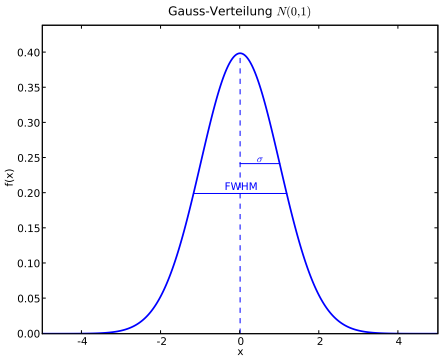


Abbildung 2.4: Standardisierte Normalverteilung $N(0,1)$.

Standardisierte Normalverteilung: Durch die Transformation

$$x \to \frac{x-\mu}{\sigma} \tag{2.51}$$

erhält man eine Normalverteilung $N(0,1)$ mit Mittelwert 0 und Varianz 1:

$$f(x) = f(x;0,1) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} \tag{2.52}$$

Eine standardisierte Normalverteilung ist in Abb. 2.4 gezeigt. Neben dem Mittelwert und der Standardabweichung σ ist auch die **volle Breite auf halber Höhe** des Maximums (FWHM = full width at half maximum) gezeigt. Diese Größe ist relativ einfach (mit Lineal und Bleistift) aus einer gemessenen Verteilung zu bestimmen. Für eine Gauss-Verteilung gibt es eine feste Beziehung zwischen FWHM und σ:

$$\frac{f(0)}{2} = \frac{1}{\sqrt{2\pi}\sigma}\, \exp\left(-\frac{(FWHM/2)^2}{2\sigma^2}\right) \implies FWHM = 2\,\sigma\,\sqrt{2\,\ln 2} \approx 2.355 \cdot \sigma \tag{2.53}$$

Verteilungsfunktion: Die Verteilungsfunktion der Normalverteilung ist nicht analytisch zu berechnen. Zahlenwerte findet man in Tabellen, in der Regel für die standardisierte Normalverteilung $N(0,1)$ als Funktion von x. Den Übergang zu Verteilungen $N(\mu,\sigma)$ findet man durch Skalieren von x mit σ und Verschieben um μ:

$$x = \frac{x'-\mu}{\sigma} \tag{2.54}$$

Statt der Verteilungsfunktion findet man auch die sogenannte Fehlerfunktion ('error function' oder "Gauss'sches Fehlerintegral") erf(x) tabelliert:

$$\begin{aligned} \mathrm{erf}(x) &= \tfrac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2} d\xi \\ \implies F(x) &= \tfrac{1}{2}\left[1 + \mathrm{erf}\left(\tfrac{x-\mu}{\sqrt{2}\sigma}\right)\right] \end{aligned} \tag{2.55}$$

Tabelle 2.1: Wahrscheinlichkeiten innerhalb von $\pm n\sigma$-Bereichen einer Normalverteilung.

a)

n	$p(\pm n\sigma)$
1	0.6827
2	0.9545
3	0.9973
4	$1 - 6.3 \cdot 10^{-5}$

b)

$p(\pm n\sigma)$	n
0.900	1.645
0.950	1.960
0.990	2.576
0.999	3.290

2.5.1 Vertrauensintervalle:

Die Verteilungsfunktion benötigt man häufig zur Bestimmung der Wahrscheinlichkeit, dass ein Ereignis innerhalb bestimmter Grenzen für x liegt. Für die Beurteilung von Messergebnissen mit normalverteilten Fehlern benutzt man zum Beispiel die Wahrscheinlichkeit, in einem zentralen 'Vertrauensintervall' von $\pm n\,\sigma$ um den Mittelwert zu liegen (Abb. 2.5a, Tab. 2.1a):

$$p(\pm n\sigma) = F(\mu + n\sigma) - F(\mu - n\sigma) = \mathrm{erf}\left(\frac{n\sigma}{\sqrt{2}\sigma}\right), \tag{2.56}$$

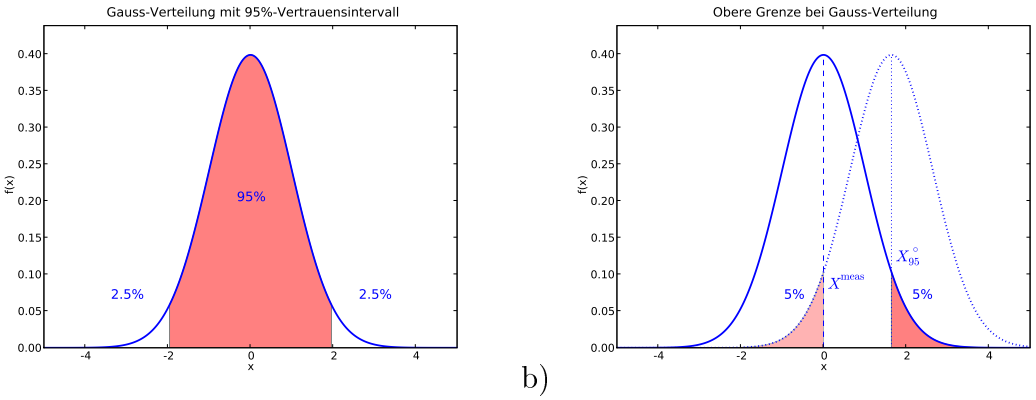
Häufig gibt man auch die Wahrscheinlichkeit, das 'Vertrauensniveau' (confidence level, c. l.), vor und fragt nach den entsprechenden Grenzen (Tab. 2.1b).

Innerhalb von 2 Standardabweichungen, $\pm 1\sigma$, um den Mittelwert liegen also 68.27 % aller Ereignisse. Häufig werden Fehler so definiert, dass 68.27 % innerhalb der Fehlergrenzen liegen, auch wenn die zugrundeliegende Verteilung nicht die Normalverteilung ist ('Standardfehler'). Bei asymmetrischen Verteilungen können die Fehler auch asymmetrisch um den Mittelwert definiert werden, zum Beispiel so, dass jeweils 16 % oberhalb und unterhalb des Fehlerbereichs liegen.

Welches Vertrauensniveau man für eine Aussage verlangt, hängt von der Problemstellung ab. Während man standardmäßig bei Messergebnissen das 1σ-Niveau angibt, verlangt man zur Festlegung von Toleranzgrenzen für Risiken, die das Leben von Menschen gefährden, viel höhere Vertrauensniveaus. Ob man nun 90 % oder 99, 9 % oder 99, 9999 % verlangt, hängt unter anderem von der 'a priori' Wahrscheinlichkeit für das Risiko, also zum Beispiel die Größe der gefährdeten Gruppe, ab ('Bayesischer Ansatz'). Wenn ein Fahrstuhl zum Beispiel im Mittel 1 Million mal während seiner Lebensdauer benutzt wird, sollte die Wahrscheinlichkeit für das Reißen des Seils kleiner als 10^{-6} sein.

Ausschließungsgrenzen: Häufig möchte man ein bestimmtes Vertrauensniveau angeben, dass bei einem gegebenen Messwert x^{mess} der wahre Wert x^{wahr} oberhalb oder unterhalb einer Grenze liegt.

> **Beispiel:** Um in der Elementarteilchenphysik die Entdeckung eines neuen Teilchens zu etablieren, wird ein Vertrauensniveau von mindestens 5 Standardabweichungen verlangt, weil jeder Physiker, der mal 1000 Histogramme mit je etwa 100 Bins angeschaut hat, eine gute Chance hat, wenigstens einen



a) b)

Abbildung 2.5: a) Fläche unter einer Gauss-Kurve, die einem Vertrauensintervall von 95% entspricht. b) Bestimmung einer oberen Grenze bei normalverteilten Fehlern, hier mit einem Vertrauensniveau von 95 %. Links ist die Verteilung um den Messwert, rechts die Verteilung um den Wert der oberen Grenze. Die schattierten Bereiche entsprechen jeweils 5 % Wahrscheinlichkeit. Siehe weitere Erläuterungen im Text.

4σ-Effekt zu beobachten. Ist dagegen ein Teilchen vorhergesagt und man findet oberhalb eines Untergrundes kein Signal, gibt man in der Regel untere Grenzen für die Häufigkeit der Erzeugung des Teilchens mit 90% oder 95% Vertrauensniveau an.

Will man zum Beispiel mit 95 % Vertrauensniveau (95 % c. l.) bei gegebenem Messwert x^{mess} eine obere Grenze für x^{wahr} angeben, stellt man die Frage: Was ist der Wert x_{95}^{o}, für den die Wahrscheinlichkeit, einen Messwert x^{mess} oder kleiner zu erhalten, 5 % beträgt. Die Grenze x_{95}^{o} wird also als Mittelwert einer Gauss-Verteilung (mit bekannter, gemessener oder geschätzter Standardabweichung) gesucht, deren Integral von $-\infty$ bis x^{mess} 5 % beträgt (Abb. 2.5b). Wegen der Symmetrie der Gauss-Verteilung kann man aber auch von einer entsprechenden Gaussverteilung um den gemessenen Wert ausgehen und x_{95}^{o} als denjenigen Wert bestimmen, für den das Integral über $x > x_{95}^{o}$ die geforderten 5 % bzw. das Komplement 95 % ergibt:

$$F(x_{95}^{o}) = 0.95 \tag{2.57}$$

Entsprechend ergibt sich für eine untere Grenze mit 95 % Vertrauensniveau:

$$F(x_{95}^{u}) = 0.05 \tag{2.58}$$

Man schreibt dann zum Beispiel:

$$x < x_{95}^{u}, \quad 95\%\ \text{c.l.} \tag{2.59}$$

Bei angenommenen gauss-verteilten Fehlern sind also die Grenzen einfach aus der Verteilungsfunktion zu bestimmen. Im allgemeinen Fall muss man aber auf die oben angegebene Definition zurückgreifen. Zum Beispiel kommt es häufig vor, dass man auf der Suche nach einem Ereignis nichts findet, also ein Nullergebnis hat. Wenn es sich um ein Zählratenexperiment handelt, ergibt sich bekanntlich für eine Poisson-Verteilung eine endliche Wahrscheinlichkeit auch bei einem nicht-verschwindenden

Tabelle 2.2: Untere und obere Grenze der Vertrauensintervalle von 90 % und 95 % für den Erwartungswert einer Posison-Verteilung gegeben, dass n Ereignisse (frei von Untergrund) gemessen wurden.

	$\epsilon = 90\,\%$		$\epsilon = 90\,\%$	
n	λ^u	λ^o	λ^u	λ^o
0	-	2.30	-	3.00
1	0.105	3.89	0.051	4.74
2	0.532	5.32	0.355	6.30
3	1.10	6.68	0.818	7.75
4	1.74	7.99	1.37	9.15
5	2.43	9.27	1.97	10.51

Mittelwert ($\lambda \neq 0$) ein Nullergebnis zu erhalten. Man kann dann nur eine obere Grenze für den wahren Wert von λ geben. Entsprechend der oben angegebene Definition fragt man für ein gefordertes Vertrauensniveau ϵ: für welchen Mittelwert λ^o_ϵ ist die Wahrscheinlichkeit die Zählrate 0 (oder kleiner) zu erhalten gerade $1 - \epsilon$:

$$p(n, \lambda) = p(0, \lambda^o_\epsilon) = \frac{(\lambda^o_\epsilon)^0}{0!} e^{-\lambda^o_\epsilon} = e^{-\lambda^o_\epsilon} \stackrel{!}{=} 1 - \epsilon \tag{2.60}$$

$$\Longrightarrow \quad \lambda^o_\epsilon = -\ln(1 - \epsilon) \tag{2.61}$$

Die Grenzen für 90 und 95 % Vertrauensniveau sind bei 0 beobachteten Ereignissen:

$$\begin{array}{rcl} \lambda^o_{90} & = & 2.30 \\ \lambda^o_{95} & = & 3.00 \end{array} \tag{2.62}$$

Für eine beobachtete Anzahl $n > 0$ ergeben sich obere und untere Grenzen λ^o und λ^u, die in Tab. 2.2 für $\epsilon = 90\,\%$ und 95 % zusammengestellt sind.

2.6 Zentraler Grenzwertsatz

Die Gauss-Verteilung hat unter allen Verteilungen eine besondere Bedeutung, weil sie für viele Verteilungen ein Grenzfall für große Zahlen darstellt. Wir hatten das bereits für die Binomial- und die Poisson-Verteilung gesehen, die beide im Grenzfall großer Mittelwerte in die Gauss-Verteilung übergehen.

Die Gauss-Verteilung kann interpretiert werden als Verteilung von Abweichungen um einen Mittelwert, die sich als Überlagerung vieler kleiner Störungen ergeben. Tatsächlich findet man, dass die Summe von n beliebigen Zufallsvariablen für große n einer Gauss-Verteilung zustrebt. In Übungsaufgabe 8 wurde das für die Summe von gleichverteilten Zufallszahlen gezeigt, wobei sich zeigte, dass die Verteilung der Summe von 12 solchen Zufallszahlen bereits sehr gut eine Gauss-Verteilung approximiert (Abb. 2.6).

Diese Eigenschaft der Gauss-Verteilung wird mathematisch im Zentralen Grenzwertsatz formuliert: Gegeben seinen n unabhängige Variablen $x_i,\ i = 1, \ldots, n$, die

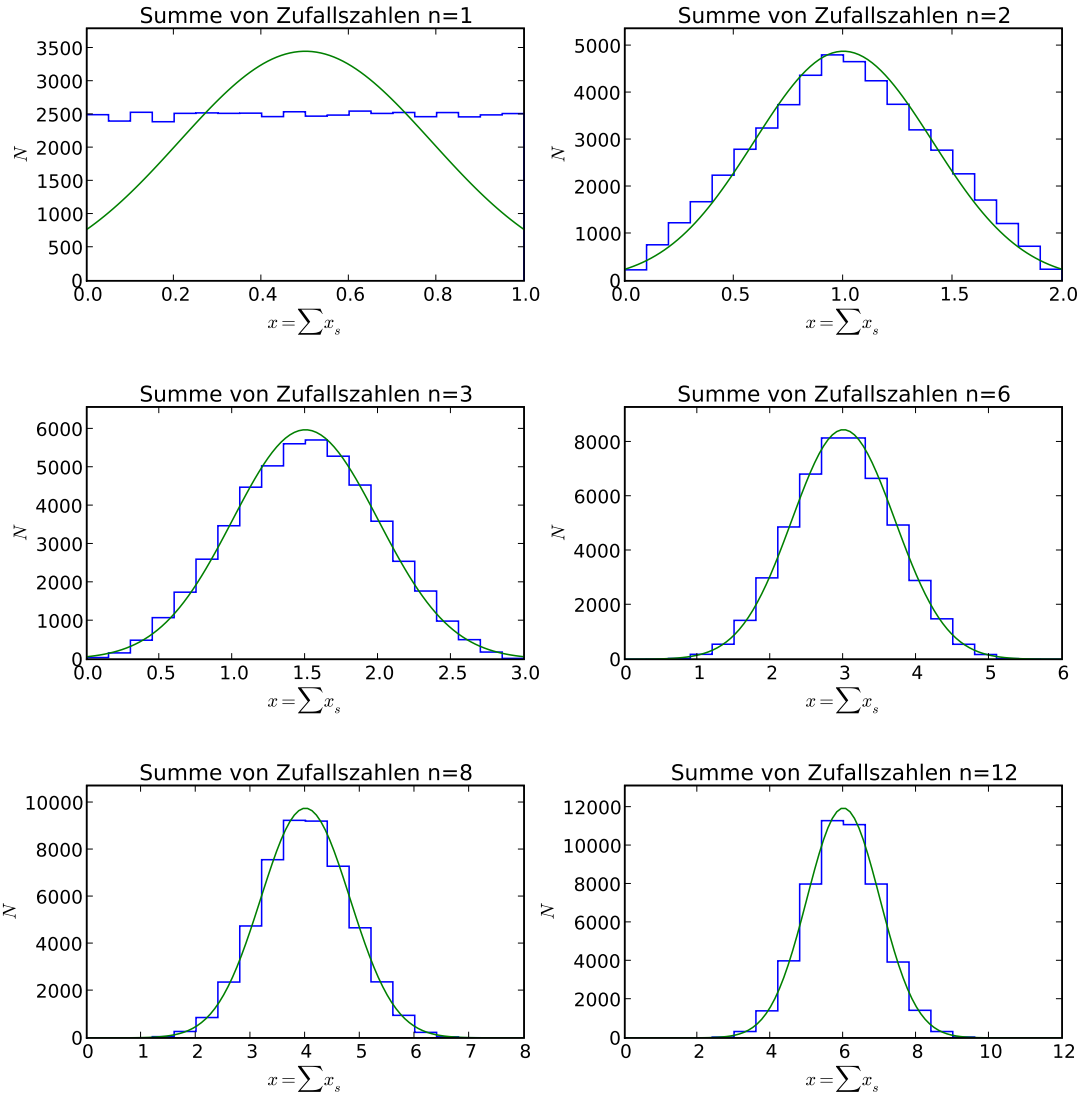


Abbildung 2.6: Beispiele von Verteilungen der Summen von n zwischen 0 und 1 gleichverteilten Zufallszahlen. Die Verteilungen werden mit Gauss-Verteilungen mit Mittelwert $\mu = n/2$ und Varianz $\sigma^2 = n/12$ verglichen.

jeweils einer Verteilung mit Mittelwert μ_i und Varianz σ_i entnommen sind (die Verteilungen sind ansonsten beliebig). Dann hat die Verteilung der Summe

$$X = \sum_{i=1}^{n} x_i \tag{2.63}$$

folgende Eigenschaften:

(i) Erwartungswert:

$$\langle X \rangle = \sum_{i=1}^{n} \mu_i; \tag{2.64}$$

(ii) Varianz:

$$\sigma_X^2 = \sum_{i=1}^{n} \sigma_i^2; \tag{2.65}$$

(iii) die Verteilung nähert sich einer Gauss-Verteilung für

$$n \to \infty. \tag{2.66}$$

Zum Beweis von (2.64) und (2.65) benutzt man die Linearität der Erwartungswertbildung: der Erwartungswert einer Summe unabhängiger Zufallszahlen ist die Summe der Erwartungswerte. Für den Erwartungswert von X ergibt sich:

$$\langle X \rangle = \left\langle \sum_i x_i \right\rangle = \sum_i \langle x_i \rangle = \sum_i \mu_i. \tag{2.67}$$

Entsprechend ergibt sich für die Varianz:

$$\begin{aligned} \sigma_X^2 &= \langle (X - \langle X \rangle)^2 \rangle = \left\langle \left(\sum_i x_i - \sum_i \mu_i \right)^2 \right\rangle = \left\langle \left(\sum_i (x_i - \mu_i) \right)^2 \right\rangle \\ &= \sum_i \left\langle (x_i - \mu_i)^2 \right\rangle + \sum_i \sum_{j \neq i} \underbrace{\langle (x_i - \mu_i)(x_j - \mu_j) \rangle}_{=0, \text{ wenn i,j unabhängig}} = \sum_i \sigma_i^2 \end{aligned} \tag{2.68}$$

Der Beweis der wichtigen Aussage (2.66) ist schwieriger und kann in Statistikbüchern nachgelesen werden, zum Beispiel [1, 2]. Abbildung 2.6 zeigt die Summe gleichverteilter Variablen, die sich der Gauss-Verteilung mit wachsender Anzahl Variabler annähert.

Kapitel 3

Verteilungen mehrerer Variablen

3.1 Eigenschaften von Verteilungen mehrerer Variablen

Im allgemeinen muss man Wahrscheinlichkeiten für mehrere Variable, die häufig auch voneinander abhängen, gleichzeitig betrachten.

Beispiele:

- Wir hatten im letzten Kapitel bereits die Multinomial-Verteilung als Beispiel einer Verteilung, die von mehreren diskreten Variablen abhängt, kennengelernt.
- Die Dichte einer Ladungswolke um eine Glühkathode hat eine dreidimensionale Verteilung.
- Ein System von n Teilchen hat eine Wahrscheinlichkeitsdichte in dem $6n$-dimensionalen Orts-Impulsraum (= Phasenraum). Zum Beispiel sind für ein ideales Gas die Ortskoordinaten gleichverteilt und die Impulsverteilung ist durch die Maxwell-Verteilung mit der Temperatur als Parameter gegeben.

3.1.1 Wahrscheinlichkeitsdichte, Verteilungsfunktion, Randverteilung

Wir betrachten n Zufallsvariable $x_1, x_2, \ldots, x_n$, die wir in einem n-Tupel

$$\vec{x} = (x_1, x_2, \ldots, x_n)^T \tag{3.1}$$

zusammenfassen.

Wahrscheinlichkeitsdichte: Die Wahrscheinlichkeitsdichte $f(\vec{x})$ liefert die differentielle Wahrscheinlichkeit an einem Punkt $\vec{x}$:

$$dp(\vec{x}) = f(\vec{x}) dx_1\, dx_2 \ldots dx_n \tag{3.2}$$

Die Normierung erfolgt über den n-dimensionalen Raum Ω in dem f definiert oder ungleich Null ist:

$$\int_\Omega f(\vec{x}) dx_1\, dx_2\, \ldots\, dx_n = 1 \tag{3.3}$$

Verteilungsfunktion: Die Verteilungsfunktion ergibt sich analog zum eindimensionalen Fall:

$$F(\vec{x}) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(\vec{\xi}) d\xi_1\, d\xi_2\, \ldots\, d\xi_n = 1 \tag{3.4}$$

Umgekehrt lässt sich die Wahrscheinlichkeitsdichte aus der Verteilungsfunktion ableiten:

$$f(\vec{x}) = \frac{\partial^n}{\partial x_1\, \partial x_2\, \ldots\, \partial x_n} F(\vec{x}). \tag{3.5}$$

Randverteilung: Die Randverteilung einer Variablen x_i ist die Projektion der Wahrscheinlichkeit auf die i-te Koordinate, das heisst man betrachtet die Verteilung von x_i gemittelt über alle anderen Variablen. Zum Beispiel ist die Randverteilung von x_1:

$$h_1(x_1) = \int_{-\infty}^{+\infty} dx_2 \int_{-\infty}^{+\infty} dx_3 \ldots \int_{-\infty}^{+\infty} dx_n\, f(\vec{x}) \tag{3.6}$$

Beispiel: Die Aufenthaltswahrscheinlichkeit des Elektrons in einem Wasserstoffatom wird in der Regel durch Kugelkoordinaten (r, θ, ϕ) angegeben. Wenn man nur an der radialen Abhängigkeit interessiert ist, erhält man die Randverteilung von r:

$$\rho_r(r) = \int_{-1}^{+1} d\cos\theta \int_0^{2\pi} d\phi\, \rho(r, \theta, \phi) \tag{3.7}$$

3.1.2 Bedingte Wahrscheinlichkeitsdichten, Selektionsschnitte

Häufig möchte man Wahrscheinlichkeitsdichten betrachten unter der Bedingung, dass eine der Variablen einen bestimmten Wert hat, zum Beispiel $x_1 = x_{10}$ (Abb. 3.1a):

$$f^*(x_2,\, x_3,\, \ldots,\, x_n | x_1 = x_{10}) = \frac{f(x_1 = x_{10}, x_2, \ldots, x_n)}{h_1(x_1 = x_{10})} \tag{3.8}$$

Das entspricht einer Umnormierung der Wahrscheinlichkeitsdichte auf eine n-1–dimensionale Hyperfläche, die durch $x_1 = x_{10}$ festgelegt ist.

Tatsächlich gibt man in der Praxis meistens ein endliches Intervall $x_{1L} < x_1 < x_{1H}$ vor und die Wahrscheinlichkeitsdichte für $x_2,\, x_3,\, \ldots,\, x_n$ muss auf diesen beschränkten n–dimensionalen Unterraum umnormiert werden (Abb. 3.1b):

$$f^*(x_2,\, x_3,\, \ldots,\, x_n | x_{1L} < x_1 < x_{1H}) = \frac{\int_{x_{1L}}^{x_{1H}} f(x_1, x_2, \ldots, x_n) dx_1}{\int_{x_{1L}}^{x_{1H}} h_1(x_1) dx_1} \tag{3.9}$$

Solche Einschränkungen von Variablenbereichen ist bei multi-dimensionalen Datensätzen ein Standardverfahren zur Bereinigung der Daten von Untergrund und

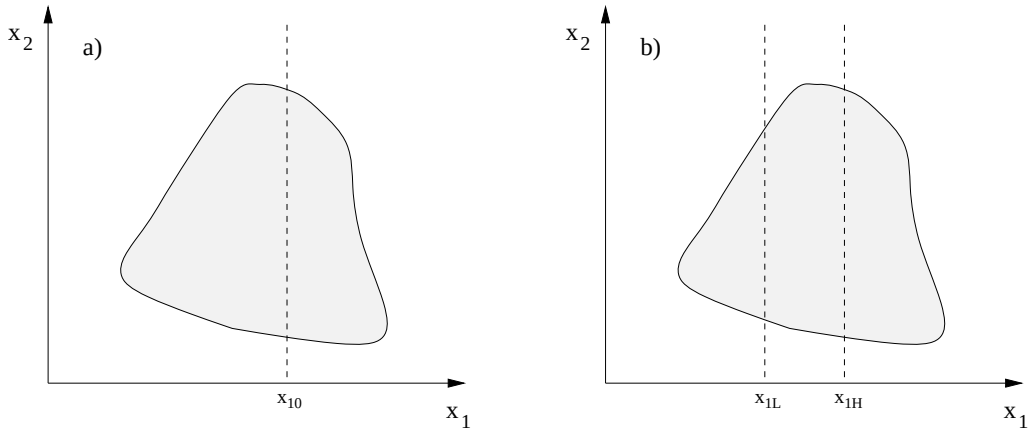


Abbildung 3.1: Bedingte Wahrscheinlichkeiten: a) Definition einer 'Hyperebene' durch $x_1 = x_{10}$, b) Schnitt in der Variablen x_1.

zur Untersuchung von Abhängigkeiten der Variablen untereinander. Häufig versucht man Signale, die auf einem Untergrund sitzen, dadurch statistisch signifikanter zu machen, indem man Bereiche, die einen relativ hohen Untergrundbeitrag liefern wegschneidet (Selektionsschnitte).

3.2 Erwartungswerte

Erwartungswert und Varianz einer Funktion: Der Erwartungswert einer Funktion g der Zufallsvariablen $\vec{x} = (x_1, x_2, \ldots, x_n)$, die die Wahrscheinlichkeitsdichte $f(\vec{x})$ haben, ist analog zum eindimensionalen Fall definiert:

$$E\left(g(\vec{x})\right) = \langle g(\vec{x})\rangle = \int_\Omega g(\vec{x})\, f(\vec{x})\, dx_1\, dx_2\, \ldots\, dx_n \tag{3.10}$$

Entsprechend ist die Varianz der Funktion g:

$$V(g(\vec{x}) = E\left((g(\vec{x}) - E(g(\vec{x}))^2\right) = \int_\Omega (g(\vec{x}) - \langle g(\vec{x})\rangle)^2\, f(\vec{x})\, dx_1\, dx_2\, \ldots\, dx_n \tag{3.11}$$

Momente: In Erweiterung der Definition für die Momente einer eindimensionalen Verteilung in Abschnitt 1.2.2 werden Momente einer mehrdimensionalen Verteilung als Erwartungswerte von Produkten von Potenzen der Zufallszahlen definiert:

1. Momente um den Ursprung:

$$\lambda_{l_1 l_2 \ldots l_n} = E\left(x_1^{l_1} \cdot x_2^{l_2} \cdot \ldots \cdot x_n^{l_n}\right) \tag{3.12}$$

2. Zentrale Momente:

$$\mu_{l_1 l_2 \ldots l_n} = E\left((x_1 - \mu_1)^{l_1} \cdot (x_2 - \mu_2)^{l_2} \cdot \ldots \cdot (x_n - \mu_n)^{l_n}\right) \tag{3.13}$$

Dabei sind die niedrigsten Momente die Mittelwerte μ_i der Zufallsvariablen x_i, die den niedrigsten Momenten mit $l_i = 1, l_k = 0$ für $k \neq i$ entsprechen:

$$\mu_i = \int_\Omega x_i\, f(\vec{x})\, dx_1\, dx_2\, \ldots\, dx_n \tag{3.14}$$

3.3 Kovarianzmatrix

3.3.1 Definition und Eigenschaften der Kovarianzmatrix

Die Momente mit $l_i = l_j = 1;\ l_k = 0$ für $k \neq i,\ k \neq j$ oder $l_i = 2;\ l_k = 0$ für $i = j$ und $k \neq i$ werden in einer sogenannten Kovarianzmatrix V_{ij} zusammengefasst:

$$V_{ij} = \mu_{0\ldots\underbrace{1}_{i}\ldots\underbrace{1}_{j}\ldots 0} = E\left((x_i - \mu_i)(x_j - \mu_j)\right) \tag{3.15}$$

$$V_{ii} = \mu_{0\ldots\underbrace{2}_{i}\ldots 0\ldots 0} = E\left((x_i - \mu_i)^2\right) \tag{3.16}$$

Die Kovarianzmatrix hat folgende Eigenschaften:

1. Die Matrix ist symmetrisch:
$$V_{ij} = V_{ji}. \tag{3.17}$$

2. Für $i = j$ ergibt sich die Varianz von x_i:
$$V_{ii} = E\left((x_i - \mu_i)^2\right) = E(x_i^2) - \left(E(x_i)\right)^2 = \sigma_i^2 \geq 0. \tag{3.18}$$

3. Die nicht-diagonalen Elemente, $i \neq j$, sind die Kovarianzen:
$$V_{ij} = \mathrm{cov}(x_i, x_j) = E\left((x_i - \mu_i)(x_j - \mu_j)\right) = E(x_i\, x_j) - E(x_i)\, E(x_j) \begin{matrix} \geqq \\ \leqq \end{matrix} 0. \tag{3.19}$$

3.3.2 Beispiel: Multi-dimensionale Gaussverteilung

Durch Verallgemeinerung der Varianz σ^2 auf die Kovarianzmatrix wird eine mehrdimensionale Gauss- oder Normalverteilung definiert:

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n\ \det(V)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T\, V^{-1}\, (\vec{x} - \vec{\mu})\right) \tag{3.20}$$

Bei zwei Variablen x_1, x_2 ist die Kovarianzmatrix:

$$V = \begin{pmatrix} \sigma_1^2 & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_1, x_2) & \sigma_2^2 \end{pmatrix} \tag{3.21}$$

Die inverse Kovarianzmatrix ist:

$$V^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \left(\mathrm{cov}(x_1, x_2)\right)^2} \begin{pmatrix} \sigma_2^2 & -\mathrm{cov}(x_1, x_2) \\ -\mathrm{cov}(x_1, x_2) & \sigma_1^2 \end{pmatrix} \tag{3.22}$$

Für einen festen Wert des Exponenten in (3.20) beschreibt $f(x)$ eine Kontur mit fester Wahrscheinlichkeitsdichte

$$f_{Kontur} = f(x|(\vec{x} - \vec{\mu})^T\, V^{-1}\, (\vec{x} - \vec{\mu}) = const). \tag{3.23}$$

Im Falle der multi-dimensionalen Gauss-Verteilung sind die Konturen konstanter Wahrscheinlichkeitsdichte n-dimensionale Ellipsoide.

Wenn die Kovarianzmatrix und damit auch ihre inverse Matrix diagonal sind, folgt für den Exponenten der Gauss-Verteilung (3.20):

$$(\vec{x}-\vec{\mu})^T\, V^{-1}\, (\vec{x}-\vec{\mu}) = \sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{\sigma_i^2} \tag{3.24}$$

Es treten also keine gemischten Terme $x_i \cdot x_j$ mit $i \neq j$ auf. Deshalb lässt sich in diesem Fall die mehrdimensionale Gauss-Verteilung (3.20) in ein Produkt eindimensionaler Gauss-Verteilungen zerlegen:

$$f(\vec{x}) = \prod_{i=1}^{n} f_i(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i-\mu_i)^2}{2\,\sigma_i^2}\right) \tag{3.25}$$

Da V und V^{-1} symmetrische, positiv definite Matrizen sind, lässt sich immer eine orthogonale Transformation $x_i \rightarrow x_i'$ finden, so dass V' und V'^{-1} diagonal sind (Hauptachsentransformation):

$$\vec{x}^T\, V^{-1}\, \vec{x} = \vec{x}^T\, U^{-1} U\, V^{-1}\, U^{-1}\, U\vec{x} \tag{3.26}$$

Für orthogonale Transformationen gilt $U^T = U^{-1}$. Die Transformation U wird so bestimmt, dass $U\, V^{-1}\, U^{-1}$ diagonal ist.

Häufig sind auf Computersystemen bereits Generatoren für gauss-verteilte Zufallszahlen vorhanden. Um mehrdimensionale Gauss-Verteilungen zu erzeugen, bestimmt man zunächst die Transformation U, die V^{-1} diagonal macht. Die Diagonalelemente $\sigma_i'^2$ und die transformierten Mittelwerte $\mu_i' = U_{ij}\,\mu_j$ sind die Parameter von n unabhängigen Gauss-Verteilungen. Entsprechend diesen Verteilungen erzeugt man nun n unabhängige gauss-verteilte Zufallszahlen x_i', die dann mittels $x_i = U_{ij}^{-1}\, x_j' = U_{ji}\, x_j'$ zurücktransformiert werden.

3.3.3 Kovarianzen von Stichproben

In Analogie zu der Schätzung der Varianz aus einer Stichprobe in (1.52) werden die Kovarianzen geschätzt. Die Korrelation zwischen zwei Variablen x^j, x^k, deren Verteilung an den Messpunkten i abgetastet wird, ergeben sich zu:

$$\mathrm{cov}(x^j, x^k) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i^j - \bar{x}^j)(x_i^k - \bar{x}^k) \tag{3.27}$$

3.3.4 Kovarianzmatrix von unabhängigen Variablen

Wenn die Zufallsvariablen x_i unabhängig sind, faktorisiert die Wahrscheinlichkeitsdichte:

$$f(\vec{x}) = f_1(x_1) \cdot f_2(x_2) \cdot \ldots \cdot f_n(x_n) \tag{3.28}$$

Wie bei der Gauss-Verteilung (3.25) ist auch im allgemeinen Fall die Kovarianzmatrix von unabhängigen Variablen diagonal. Um die Kovarianzmatrix auszurechnen, berechnen wir zunächst den Erwartungswert von $x_i\, x_j$:

$$E(x_i\, x_j) = \int x_i\, f_i(x_i)\, dx_i \cdot \int x_j\, f_j(x_j)\, dx_j \cdot \prod_{k\neq i; k\neq j} \underbrace{\int f_k(x_k)\, dx_k}_{=1} = E(x_i) \cdot E(x_j) \tag{3.29}$$

Damit ergibt sich:

$$\mathrm{cov}(x_i, x_j) = E\left((x_i - \mu_i)(x_j - \mu_j)\right) = \underbrace{E(x_i\, x_j) - E(x_i)\, E(x_j) = 0}_{(3.29)} \tag{3.30}$$

Für unabhängige Variable verschwinden also die Kovarianzen:

$$x_i,\, x_j \text{ unabhängig} \implies \mathrm{cov}(x_i, x_j) = 0 \tag{3.31}$$

Die Umkehrung dieses Satzes gilt nicht im Allgemeinen. Man sieht an (3.30), dass die Kovarianzen verschwinden, wenn sich die Terme $(x_i - \mu_i)(x_j - \mu_j)$ im Mittel auslöschen. Das kann auf verschiedenste Weisen passieren. Zum Beispiel heben sich in Abb. 3.2b gerade die Kovarianzen der rechten und linken Hälfte der Verteilung auf (in der linken Hälfte ergibt sich eine positive Korrelation und in der rechten eine negative). Die Kovarianz der gesamten Verteilung verschwindet also, obwohl es offensichtlich eine Abhängigkeit von x_1 und x_2 gibt.

3.3.5 Korrelationen

Wenn die Kovarianzen nicht verschwinden, nennt man die entsprechenden Variablen korreliert. Als Maß für die Stärke der Korrelation definiert man den **Korrelationskoefizienten**:

$$\rho(x_i,\, x_j) = \frac{V_{ij}}{\sqrt{V_{ii}\, V_{jj}}} = \frac{\mathrm{cov}(x_i, x_j)}{\sigma_i \cdot \sigma_j} \tag{3.32}$$

Durch die Normierung auf die Standardabweichungen ergibt sich für den Wertebereich von ρ:

$$-1 \leq \rho(x_i,\, x_j) \leq +1 \tag{3.33}$$

Je mehr der Korrelationskoeffizient von Null abweicht, umso besser kann man aus der Kenntnis einer Variablen die andere vorhersagen (Abb. 3.2):

$$\begin{array}{lcll} \rho(x_i,\, x_j) \to +1 & \implies & x_i \to +x_j & \text{(positiv korreliert)} \\ \rho(x_i,\, x_j) \to \pm 0 & \implies & x_i,\ x_j \text{ unabhängig} & \text{(nicht korreliert)} \\ \rho(x_i,\, x_j) \to -1 & \implies & x_i \to -x_j & \text{(negativ korreliert)} \end{array} \tag{3.34}$$

Beispiele:

1. Ein Teilchen, das wie Abb. 3.3 durch eine Materieschicht geht, wird unter einem Winkel θ gestreut und erfährt eine Ablage Δx. Streuwinkel und Ablage sind positiv korreliert.

2. Ein Anthropologe untersucht 5 Funde von Neandertalerknochen. Er vergleicht die Längen der Oberarm- mit der der Oberschenkelknochen und möchte seinen naheliegenden Verdacht, dass beide korreliert sind, statistisch erhärten.

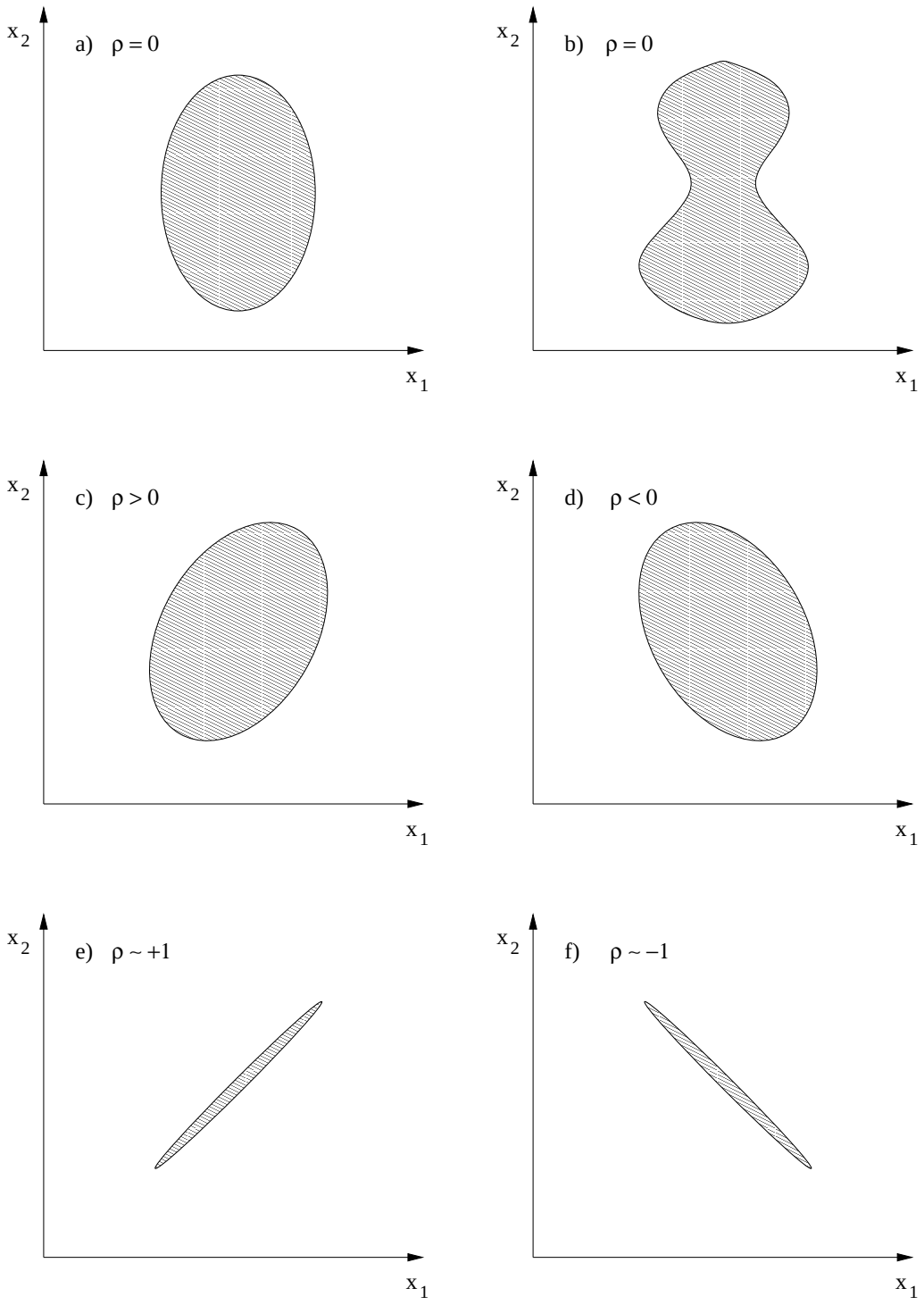


Abbildung 3.2: Verteilungsformen mit unterschiedlichem Korrelationskoeffizienten ρ.

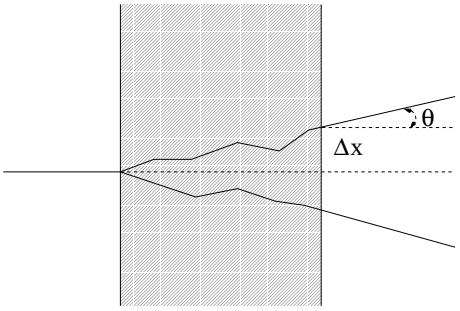


Abbildung 3.3: Streuung von Teilchen in einer Materieschicht, zum Beispiel α-Teilchen in einer Goldfolie wie bei dem Rutherford-Experiment.

Die vorliegenden Daten sind (l^a, l^b sind die Längen jeweils der Arm- und Beinknochen):

Fund	l^a [mm]	l^b [mm]	l^{a2} [mm^2]	l^{b2} [mm^2]	$l^a l^b$ [mm^2]
1	312	430	97344	184900	134160
2	335	458	112225	209764	153430
3	286	407	81796	165649	116402
4	312	440	97344	193600	137280
5	305	422	93025	178084	128710
Mittel	310.0	431.4	96346.8	186399.4	133996.4
$\sigma_{l^{a,b}}$	17.56	19.15		$\mathrm{cov}(l^a, l^b)$	328.0

(3.35)

Die letzten drei Spalten enthalten die Berechnung von l^{a2}, l^{b2} und $l^a \cdot l^b$ und deren Mittelwerte, die dann in die Berechnung der Kovarianzmatrix eingehen. Entsprechend (3.27) ergibt sich:

$$\mathrm{cov}(l^a, l^b) = E(l^a \cdot l^b) - E(l^a)\,E(l^b) = \frac{5}{5-1}\left(\overline{l^a l^b} - \overline{l^a} \cdot \overline{l^b}\right) \tag{3.36}$$

Der Faktor 5/4 korrigiert wie bei der Berechnung der Varianz einer Stichprobe darauf, dass bezüglich des Mittelwertes bereits die quadratischen Abweichungen minimiert werden. Einsetzen der Zahlen aus der Tabelle ergibt:

$$\mathrm{cov}(l^a, l^b) = 328.0 \implies \rho(l^a, l^b) = \frac{\mathrm{cov}(l^a, l^b)}{\sigma_l^a \cdot \sigma_l^b} = 0.975 \tag{3.37}$$

Die Korrelation in der Größe der Arm- und Beinknochen ist also sehr hoch.

3.4 Lineare Funktionen von mehreren Zufallsvariablen

In den folgenden Abschnitten werden Funktionen von mehreren Zufallsvariablen betrachtet. Wir interessieren uns insbesondere für die Berechnung einfacher Erwartungswerte dieser Funktionen, wie Mittelwerte und Varianzen. Die Berechnung der

Varianz einer Funktion von Zufallsvariablen wird für die Fehlerfortplanzung von Messungen benutzt.

Ein besonders einfacher Fall ist eine lineare Funktion von mehreren Variablen. Wir werden im folgenden häufig auch bei nicht-linearen Funktionen durch Linearisierung um einen Entwicklungspunkt die Ergebnisse für lineare Funktionen benutzen.

Es sei g eine lineare Funktion der n Zufallsvariablen $\vec{x} = (x_1, \ldots, x_n)$:

$$g(\vec{x}) = \sum_{i=1}^{n} a_i \, x_i \tag{3.38}$$

Erwartungswert: Der Erwartungswert der Funktion ist:

$$E\left(g(\vec{x})\right) = \sum_{i=1}^{n} a_i \underbrace{E\left(x_i\right)}_{=\mu_i} = \sum_{i=1}^{n} a_i \, \mu_i \tag{3.39}$$

Varianz:

$$\begin{aligned} V\left(g(\vec{x})\right) &= E\left((g(\vec{x}) - E\left(g(\vec{x})\right))^2\right) = E\left(\left(\textstyle\sum_i a_i \, x_i - \sum_i a_i \, \mu_i\right)^2\right) \\ &= E\left(\left(\textstyle\sum_i a_i \left(x_i - \mu_i\right)\right)^2\right) = \textstyle\sum_i \sum_j a_i \, a_j \, E\left((x_i - \mu_i)(x_j - \mu_j)\right) \\ &= \textstyle\sum_i \sum_j a_i \, a_j \, V_{ij} \end{aligned} \tag{3.40}$$

Dabei ist V_{ij} die Kovarianzmatrix der Zufallsvariablen $\vec{x}$. Mit der Beziehung $V_{ij} = V_{ji}$ lässt sich die Varianz von g durch die Varianzen und die Kovarianzen ausdrücken:

$$V(g(\vec{x})) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i \, a_j \, V_{ij} \tag{3.41}$$

Wenn die x_i unabhängig sind, ist $V_{ij} = 0$ für $i \neq j$ und die Varianz von g ergibt sich nur aus den Varianzen der x_i:

$$V\left(g(\vec{x})\right) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 \tag{3.42}$$

Beispiele:

1. Eine Stichprobe $x_1, \ldots, x_n$ aus einer Verteilung mit dem Mittelwert μ und Varianz σ^2 kann man als einen Satz von n unabhängigen Zufallsvariablen interpretieren, die alle den gleichen Mittelwert $\mu_i = \mu$ und die gleiche Varianz $\sigma_i^2 = \sigma^2$ haben. Das arithmetische Mittel der x_i ist eine lineare Funktion der x_i:

   $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.43}$$

   Der Erwartungswert des Mittelwertes ist dann:

   $$E\left(\bar{x}\right) = \frac{1}{n} \sum_{i=1}^{n} E\left(x_i\right) = \frac{1}{n} \cdot n \, \mu = \mu \tag{3.44}$$

Das heisst, das arithmetischen Mittel einer Stichprobe ist eine 'erwartungstreue' Schätzung des Erwartungswertes μ der entsprechenden Verteilung, aus der die Stichprobe gezogen wurde.

Die Varianz des arithmetischen Mittels ist (die Kovarianzen fallen weg, weil die x_i unabhängig sind):

$$V(\bar{x}) = \sigma_{\bar{x}}^2 = \left(\frac{1}{n}\right)^2 \sum_i \sigma_i^2 = \left(\frac{1}{n}\right)^2 n\,\sigma^2 = \frac{\sigma^2}{n} \tag{3.45}$$

Damit hat man das bekannte Ergebnis, dass der Fehler des Mittelwertes von n Messungen um $1/\sqrt{n}$ kleiner als der Fehler der Einzelmessung ist:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.46}$$

2. Im allgemeinen hat die Varianz einer Funktion von zwei Zufallsvariablen,

$$g(x, y) = a\,x + b\,y, \tag{3.47}$$

folgende Form:

$$V(a\,x + b\,y) = a^2 \underbrace{V_{xx}}_{=\sigma_x^2} + b^2 \underbrace{V_{yy}}_{=\sigma_y^2} + 2\,a\,b \underbrace{V_{xy}}_{=\mathrm{cov}(x,y)} = a^2\,\sigma_x^2 + b^2\,\sigma_y^2 + 2ab\,\sigma_x\sigma_y\,\rho(x,y) \tag{3.48}$$

Dabei kann der Korrelationskoeffizient $\rho(x, y)$ Werte von -1 bis +1 annehmen.

3.5 Nicht-lineare Funktionen von Zufallsvariablen

3.5.1 Eine Funktion von einem Satz von Zufallsvariablen

In diesem Abschnitt wollen wir allgemeine Funktionen g der Zufallsvariablen betrachten:

$$g = g(x_1, \ldots, x_n). \tag{3.49}$$

Um die Ergebnisse des vorigen Abschnitts benutzen zu können, linearisieren wir die Funktion in der Umgebung der Mittelwerte $\vec{\mu}$:

$$g(\vec{x}) = g(\vec{\mu}) + \sum_{i=1}^{n} (x_i - \mu_i) \left.\frac{\partial g}{\partial x_i}\right|_{\vec{x}=\vec{\mu}} + \ldots \tag{3.50}$$

Erwartungswert: Der Erwartungswert der Funktion g ist in der linearen Näherung:

$$E\left(g(\vec{x})\right) = E\left(g(\vec{\mu})\right) + \sum_{i=1}^{n} \underbrace{E\left(x_i - \mu_i\right)}_{=0} \left.\frac{\partial g}{\partial x_i}\right|_{\vec{x}=\vec{\mu}} = E\left(g(\vec{\mu})\right) = g(\vec{\mu}) \tag{3.51}$$

Der Erwartungswert der Funktion $g(\vec{x})$ ist also diese Funktion an der Stelle der Erwartungswerte von $\vec{x}$:

$$E\left(g(\vec{x})\right) = g(\vec{\mu}) \tag{3.52}$$

Varianz:

$$\begin{aligned} V\left(g(\vec{x})\right) &= E\left((g(\vec{x}) - E\left(g(\vec{x})\right))^2\right) \\ &= E\left((g(\vec{x}) - g(\vec{\mu}))^2\right) \\ &= E\left(\left(\sum_i (x_i - \mu_i)\frac{\partial g}{\partial x_i}\right)^2\right) \\ &= \sum_i \sum_j \frac{\partial g}{\partial x_i}\frac{\partial g}{\partial x_j} E\left((x_i - \mu_i)(x_j - \mu_j)\right) \\ &= \sum_i \sum_j \frac{\partial g}{\partial x_i}\frac{\partial g}{\partial x_j} V_{ij} \end{aligned} \tag{3.53}$$

Das entspricht also genau dem Ergebnis (3.40), wenn man statt der Koeffizienten a_i die partiellen Ableitungen $\partial g/\partial x_i$ einsetzt.

In Matrixschreibweise definiert man den Spaltenvektor:

$$\vec{a} = \begin{pmatrix} \frac{\partial g}{\partial x_1} \\ \vdots \\ \frac{\partial g}{\partial x_n} \end{pmatrix} \tag{3.54}$$

Damit ergibt sich für die Varianz:

$$V\left(g(\vec{x})\right) = \sigma^2\left(g(\vec{x})\right) = \vec{a}^T\, V(\vec{x})\, \vec{a} \tag{3.55}$$

Zum Beispiel erhält man für $n = 2$:

$$\sigma^2\left(g(\vec{x})\right) = \left(\frac{\partial g}{\partial x_1}\right)^2 \sigma_1^2 + \left(\frac{\partial g}{\partial x_2}\right)^2 \sigma_2^2 + 2\frac{\partial g}{\partial x_1}\frac{\partial g}{\partial x_2}\,\mathrm{cov}(x_1, x_2) \tag{3.56}$$

Das ist also die bekannte Formel, die auch für Fehlerfortpflanzung benutzt wird.

3.5.2 Mehrere Funktionen von einem Satz von Zufallszahlen

Wir betrachten jetzt den allgemeineren Fall, dass m Funktionen $g = (g_1, \ldots, g_m)$ von den gleichen n Zufallszahlen $(x_1, \ldots, x_n)$ abhängen:

$$\vec{g}(\vec{x}) = \begin{pmatrix} g_1(\vec{x}) \\ \vdots \\ g_m(\vec{x}) \end{pmatrix} \tag{3.57}$$

Ein häufig auftretendes Beispiel ist eine Koordinatentransformation der Zufallsvariablen: die transformierten Variablen sind im allgemeinen eine Funktion aller ursprünglichen Variablen.

Die Erwartungswerte der Funktionen g_j und deren Varianzen ergeben sich für jede Funktion einzeln. Neu kommt jetzt allerdings hinzu, dass die Funktionen untereinander korreliert sein können und damit nicht-verschwindende Kovarianzen haben.

Wir linearisieren wieder jede der Funktionen ($k = 1, \ldots, m$):

$$g_k(\vec{x}) = g_k(\vec{\mu}) + \sum_{i=1}^{n} (x_i - \mu_i)\frac{\partial g_k}{\partial x_i}\bigg|_{\vec{x}=\vec{\mu}} + \ldots \tag{3.58}$$

Mit

$$S_{ki} = \left.\frac{\partial g_k}{\partial x_i}\right|_{\vec{x}=\vec{\mu}} \tag{3.59}$$

ergibt (3.58):

$$\begin{aligned} g_k(\vec{x}) &= g_k(\vec{\mu}) + \textstyle\sum_{i=1}^{n}(x_i - \mu_i)\, S_{ki} \\ \text{oder} & \\ \vec{g}(\vec{x}) &= \vec{g}(\vec{\mu}) + S\,(\vec{x} - \vec{\mu}) \end{aligned} \tag{3.60}$$

Dabei sind $\vec{x},\ \vec{\mu}$ Spaltenvektoren und die Jacobische Funktionalmatrix S ist in Matrixschreibweise:

$$S = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g_m}{\partial x_1} & \frac{\partial g_m}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_n} \end{pmatrix} \tag{3.61}$$

Erwartungswert: Die Erwartungswerte der Funktionen $\vec{g}(\vec{x})$ ergibt sich wie für eine einzelne Funktion (3.51):

$$E\left(\vec{g}(\vec{x})\right) = \vec{g}(\vec{\mu}) \tag{3.62}$$

Varianz:

$$\begin{aligned} V_{kl}\left(\vec{g}(\vec{x})\right) &= E\left[(g_k(\vec{x}) - E\left[g_k(\vec{x})\right])(g_l(\vec{x}) - E\left[g_l(\vec{x})\right])\right] \\ &= \textstyle\sum_i \sum_j \frac{\partial g_k}{\partial x_i}\frac{\partial g_l}{\partial x_j} \underbrace{E\left((x_i - \mu_i)(x_j - \mu_j)\right)}_{=V_{ij}(\vec{x})} \\ &= \textstyle\sum_i \sum_j \frac{\partial g_k}{\partial x_i}\frac{\partial g_l}{\partial x_j} V_{ij}\,(\vec{x}) = \sum_i \sum_j S_{ki}\, S_{lj}\, V_{ij}\,(\vec{x}) \\ \Longrightarrow \qquad V\left(\vec{g}(\vec{x})\right) &= S \cdot V\,(\vec{x}) \cdot S^T \end{aligned} \tag{3.63}$$

Dabei sind in der letzten Zeile alle Größen Matrizen.

Um das obige Beispiel einer Variablentransformation aufzugreifen: Die Matrix S kann man beispielsweise so bestimmen, dass die Transformation $\vec{x} \to \vec{g}$ die Kovarianzmatrix $V\,(\vec{g})$ diagonal macht, die neuen Variablen g_i also nicht korreliert sind.

Beispiel: Fehlerfortpflanzung bei Koordinatenwechsel.
Auf einem Koordinatenmesstisch werden rechtwinklige Koordinaten (x, y) mit den Auflösungen

$$\begin{aligned} \sigma_x &= 1\,\mu\text{m} \\ \sigma_y &= 3\,\mu\text{m} \end{aligned} \tag{3.64}$$

gemessen. Da die Messungen der beiden Koordinaten unabhängig sein sollen, ist die Kovarianzmatrix diagonal:

$$V(x, y) = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \tag{3.65}$$

Für die weitere Auswertung sollen die Messpunkte in Polarkoordinaten (r, ϕ) ausgedrückt werden:

$$\begin{array}{rcl} x &=& r\,\cos\phi \\ y &=& r\,\sin\phi \end{array} \quad \Longrightarrow \quad \begin{array}{rcl} r &=& \sqrt{x^2+y^2} \\ \phi &=& \arctan\frac{y}{x} \end{array} \tag{3.66}$$

Wir wollen nun berechnen, wie sich der Fehler der x, y-Messungen auf r, ϕ fortpflanzt und bestimmen deshalb die Kovarianzmatrix für die Variablen r, ϕ. Die Funktionalmatrix für die Transformation ist:

$$S = \begin{pmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y} \\ \frac{\partial \phi}{\partial x} & \frac{\partial \phi}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{x}{r} & \frac{y}{r} \\ -\frac{y}{r^2} & \frac{x}{r^2} \end{pmatrix} \tag{3.67}$$

Damit transformiert sich die Kovarianzmatrix wie folgt:

$$V(r,\phi) = S \cdot V(x,y) \cdot S^T = \begin{pmatrix} \frac{1}{r^2}(x^2\sigma_x^2 + y^2\sigma_y^2) & \frac{xy}{r^3}(-\sigma_x^2+\sigma_y^2) \\ \frac{xy}{r^3}(-\sigma_x^2+\sigma_y^2) & \frac{1}{r^4}(y^2\sigma_x^2 + x^2\sigma_y^2) \end{pmatrix} \tag{3.68}$$

Ausgedrückt in Polarkoordinaten ergibt sich für die Kovarianzmatrix:

$$\begin{array}{rcl} V(r,\phi) &=& \begin{pmatrix} \sigma_r^2 & \mathrm{cov}(r,\phi) \\ \mathrm{cov}(r,\phi) & \sigma_\phi^2 \end{pmatrix} \\ &=& \begin{pmatrix} \cos^2\phi\,\sigma_x^2 + \sin^2\phi\,\sigma_y^2 & \frac{\sin\phi\,\cos\phi}{r}(-\sigma_x^2+\sigma_y^2) \\ \frac{\sin\phi\,\cos\phi}{r}(-\sigma_x^2+\sigma_y^2) & \frac{1}{r^2}(\sin^2\phi\,\sigma_x^2 + \cos^2\phi\,\sigma_y^2) \end{pmatrix} \end{array} \tag{3.69}$$

Man sieht, dass die Kovarianzmatrix auch in Polarkoordinaten diagonal ist, wenn die x- und y-Messgenauigkeit gleich, also $\sigma_x = \sigma_y$, ist. Die Kovarianzen verschwinden auch für die Spezialfälle $\phi = 0°, 90°$, das heisst für Punkte auf der x- bzw. y-Achse:

$$V(r,\phi=0°) = \begin{pmatrix} \sigma_r^2 = \sigma_x^2 & \mathrm{cov}(r,\phi) = 0 \\ \mathrm{cov}(r,\phi) = 0 & \sigma_\phi^2 = \frac{1}{r^2}\sigma_y^2 \end{pmatrix} \tag{3.70}$$

$$V(r,\phi=90°) = \begin{pmatrix} \sigma_r^2 = \sigma_y^2 & \mathrm{cov}(r,\phi) = 0 \\ \mathrm{cov}(r,\phi) = 0 & \sigma_\phi^2 = \frac{1}{r^2}\sigma_x^2 \end{pmatrix} \tag{3.71}$$

Man kann jetzt auch wieder umgekehrt die Varianzen der Zufallsvariablen x und y berechnen, wenn die Kovarianzmatrix in Polarkoordinaten vorliegt. Will man zum Beispiel die Varianz von x am Punkt $(1,1)$, also $(r = \sqrt{2},\ \phi = 45°)$, berechnet man zunächst

$$\sigma_r^2 = 5, \qquad \sigma_\phi 2 = \frac{5}{2}, \qquad \mathrm{cov}(r,\phi) = \frac{4}{\sqrt{2}} \tag{3.72}$$

Damit ergibt sich beispielsweise für σ_x^2 (siehe (3.56)):

$$\begin{array}{rcl} \sigma_x^2 &=& \left(\frac{\partial x}{\partial r}\right)^2\sigma_r^2 + \left(\frac{\partial x}{\partial \phi}\right)^2\sigma_\phi^2 + 2\,\frac{\partial x}{\partial r}\frac{\partial x}{\partial \phi}\,\mathrm{cov}(r,\phi) \\ &=& \cos^2\phi\,\sigma_r^2 + r^2\sin^2\phi\,\sigma_\phi^2 - 2\,r\,\cos\phi\,\sin\phi\,\mathrm{cov}(r,\phi) \\ &=& \frac{5}{2} + \frac{5}{2} - \frac{8}{2} = 1 \qquad (= \sigma_x^2) \end{array} \tag{3.73}$$

Es ergibt sich also korrekt wieder der Wert $\sigma_x^2 = 1$, der hineingesteckt wurde. Hier sieht man, dass man im allgemeinen die Kovarianzen nicht vernachlässigen kann: ohne Berücksichtigung der Kovarianz hätte sich $\sigma_x^2 = 5$ ergeben.

3.6 Transformationen von Zufallsvariablen

In dem obigen Beispiel hatten wir einen Transformation der Zufallsvariablen x, y auf r, ϕ und die daraus folgende Transformation der Varianzen betrachtet. Wir fragen nun, wie sich die Wahrscheinlichkeitsdichten transformieren, wenn man zu anderen Variablen übergeht. Variablentransformationen macht man unter anderem auch um einfachere Wahrscheinlichkeitsdichten zu erhalten, zum Beispiel Gleichverteilungen für eine Simulation (siehe Abschnitt 1.3).

Wir betrachten zunächst den Fall, dass eine einzelne Variable in eine andere transformiert wird:

$$x \to z, \qquad f(x) \to g(z) \tag{3.74}$$

In einem Interval dx, das in dz übergeht, müssen die Wahrscheinlichkeiten vor und nach der Transformation gleich sein:

$$dp = f(x)\,dx = g(z)\,dz \implies g(z) = f(x(z)) \left|\frac{dx}{dz}\right| \tag{3.75}$$

Im rechten Ausdruck wird der Betrag der Ableitung genommen, damit die Wahrscheinlichkeit positiv bleibt.

Für n Variable mit der Transformation

$$(x_1, \ldots, x_n) \to (z_1, \ldots, z_n), \qquad f(x_1, \ldots, x_n) \to g(z_1, \ldots, z_n) \tag{3.76}$$

ergibt sich die Bedingung:

$$f(\vec{x})\,dx_1 \ldots dx_n = g(\vec{z})\,dz_1 \ldots dz_n \implies g(\vec{z}) = f(\vec{x}(\vec{z})) \left|\frac{\partial(x_1, \ldots, x_n)}{\partial(z_1, \ldots, z_n)}\right| \tag{3.77}$$

Der rechte Ausdruck ist die Funktional- oder Jacobi-Determinante:

$$\left|\frac{\partial(x_1, \ldots, x_n)}{\partial(z_1, \ldots, z_n)}\right| = \det \begin{pmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} & \cdots & \frac{\partial x_1}{\partial z_n} \\ \frac{\partial x_2}{\partial z_1} & \frac{\partial x_2}{\partial z_2} & \cdots & \frac{\partial x_2}{\partial z_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial z_1} & \frac{\partial x_n}{\partial z_2} & \cdots & \frac{\partial x_n}{\partial z_n} \end{pmatrix} \tag{3.78}$$

Beispiele:

1. In der Physik kommt häufig die Transformation auf krummlinige Koordinaten vor. Zum Beispiel transformiert sich bei dem Übergang von kartesischen auf Kugelkoordinaten, $(x, y, z) \to (r, \theta, \phi)$, das Volumenelement bekanntlich wie

   $$dx\,dy\,dz \to r^2 \sin\theta\,dr\,d\theta\,d\phi, \tag{3.79}$$

   so dass sich die Jacobi-Determinante zu $r^2 \sin\theta$ ergibt.

2. Ein schnelles geladenes Teilchen emittiert sogenannte Bremsstrahlung, wenn eine Kraft auf das Teilchen wirkt, wie beim Durchgang durch Materie oder

in elementaren Wechselwirkungen. Die Wahrscheinlichkeitsdichte für die Abstrahlungsrichtung θ relativ zur Teilchenrichtung hat etwa folgende Form:

$$w(\theta) = w_0 \frac{\sin\theta}{1 - \beta\cos\theta} \tag{3.80}$$

Dabei ist $\beta = v/c$ die Teilchengeschwindigkeit in Einheiten der Lichtgeschwindigkeit. Für Elektronen ist β schon bei relativ niedrigen Energien sehr nahe 1, zum Beispiel für $E = 1\,\mathrm{GeV}$ ist $1 - \beta = 1.3 \cdot 10^{-7}$. In diesem Fall 'hochrelativistischer' Teilchen ist der Ausdruck $1/(1 - \beta\cos\theta)$ bei $\theta = 0$ nahezu divergent. Dieses Verhalten wird auch nicht durch den $\sin\theta$-Term in (3.80) gedämpft, weil das Winkelelement $\sin\theta\, d\theta = d\cos\theta$ bei $\theta = 0$ endlich bleibt.

Eine Simulation der Abstrahlung wird also zum Beipiel mit der 'Hit and Miss' Methode sehr ineffektiv. Man wird also eine Transformation suchen, die das Polverhalten dämpft. Tatsächlich kann man (3.80) auf eine Gleichverteilung transformieren. Entsprechend Abschnitt 1.3 machen wir den Ansatz (u ist eine zwischen 0 und 1 gleichverteilte Zufallsvariable):

$$w(\theta)\, d\theta = du \implies u = \int_0^\theta w(\vartheta)\, d\vartheta = W(\theta) = \frac{w_0}{\beta} \ln \frac{1 - \beta\cos\theta}{1 - \beta}, \tag{3.81}$$

wobei $W(\theta)$ die Verteilungsfunktion ist. Der Normierungsfaktor w_0 ergibt sich aus der Integration von $w(\theta)$ über den gesamten Wertebereich:

$$\frac{1}{w_0} = \int_0^\pi w(\vartheta)\, d\vartheta = W(\pi) = \frac{1}{\beta} \ln \frac{1 + \beta}{1 - \beta}, \tag{3.82}$$

Die Transformation $\theta \to u$ ergibt sich aus der Inversion von (3.81):

$$\theta = \arccos\left[\frac{1}{\beta}\left((1 - \beta)\, e^{\frac{\beta u}{w_0}} - 1\right)\right] \tag{3.83}$$

Nehmen wir weiterhin an, dass die azimuthale Winkelverteilung der Strahlung durch Polarisationseffekte (die Elektronenspins könnten zum Beispiel transversal zu ihrer Flugrichtung polarisiert sein) sinusförmig moduliert wird:

$$w'(\theta, \phi) = w_0' \frac{\sin\theta\, \sin\phi}{1 - \beta\cos\theta} \tag{3.84}$$

Eine entsprechende Transformation von ϕ im Interval 0 bis π auf eine zwischen 0 und 1 gleichverteilte Variable v erhält man wie in (3.81):

$$\alpha\, \sin\phi\, d\phi = dv \implies v = \frac{\int_0^\phi \sin\varphi\, d\varphi}{\int_0^\pi \sin\varphi\, d\varphi} = \frac{\cos\phi + 1}{2} \tag{3.85}$$

Dabei ist $\alpha = 1/2$ die Normierungskonstante und es gilt $w_0' = w_0\, \alpha$. Die gesamte Variablentransformation ist damit:

$$\begin{aligned} \theta &= \arccos\left[\tfrac{1}{\beta}\left((1 - \beta)\, e^{\frac{\beta u}{w_0}} - 1\right)\right] \\ \phi &= \arccos\left(2v - 1\right) \end{aligned} \tag{3.86}$$

Daraus ergibt sich die Funktionaldeterminante:

$$\left|\frac{\partial(\theta,\phi)}{\partial(u,v)}\right| = \det\begin{pmatrix} \frac{\partial\theta}{\partial u} & \frac{\partial\theta}{\partial v} \\ \frac{\partial\phi}{\partial u} & \frac{\partial\phi}{\partial v} \end{pmatrix} = \frac{1}{w_0'}\frac{1-\beta\,\cos\theta}{\sin\theta\,\sin\phi} = \frac{1}{w'(\theta,\phi)} \tag{3.87}$$

Es ist natürlich kein Zufall, dass die Jacobi-Determinante gerade das Reziproke der ursprünglichen Dichteverteilung ergibt, weil ja gerade auf eine Gleichverteilung transformiert werden sollte.

Kapitel 4

Stichproben und Schätzungen

4.1 Stichproben, Verteilungen und Schätzwerte

Eine physikalische Messung ist eine endliche Stichprobe aus einer Grundgesamtheit, die endlich oder unendlich sein kann. Im allgemeinen möchte man bei der Weiterverarbeitung der Messergebnisse eine **Reduktion der Daten** auf die wesentliche Information erreichen. Diese Information steckt in der mathematischen Beschreibung der Verteilung der Grundgesamtheit, die durch – hoffentlich endlich viele – Parameter beschrieben werden kann. Man versucht nun die Verteilungen zu bestimmen, indem man Schätzwerte für diese Parameter aus der Messung ableitet. Eine allgemeine Methode zur Schätzung von Parametern ist die Maximum-Likelihood-Methode (Kapitel 6).

Zum Beispiel weiss man beim radioaktiven Zerfall,

$$N(t) = N_0 e^{-\lambda t}, \tag{4.1}$$

dass der einzige Parameter die Zerfallswahrscheinlichkeit (oder mittlere Lebensdauer) λ ist, die man als Mittelwert aus der gemessenen Häufigkeitsverteilung $N(t)$ bestimmt. Die Messwerte haben sonst keine weitere wesentliche Information (wenn man weiss, dass sie einem Zerfallsgesetz folgen).

Eine Stichprobe von n Messungen aus einer Grundgesamtheit mit der Wahrscheinlichkeitsdichte $f(x)$

$$\vec{x} = (x_1, \ldots, x_n) \tag{4.2}$$

kann man als eine n-dimensionale Zufallsvariable auffassen und ihr eine Wahrscheinlichkeitsdichte

$$g(\vec{x}) = g(x_1, \ldots, x_n) \tag{4.3}$$

zuordnen (siehe Beispiel 1 in Abschnitt 3.4). Damit die **Stichprobe zufällig** ist, muss gelten:

(i) Die x_i sind unabhängig

$$\Longrightarrow \; g(\vec{x}) = g_1(x_1) \cdot g_2(x_2) \ldots g_n(x_n) \tag{4.4}$$

(ii) Jeder Messwert x_i hat die Wahrscheinlichkeitsdichte der Grundgesamtheit:

$$g_i(x_i) = f(x) \tag{4.5}$$

Diese Eigenschaften sind durchaus nicht immer gegeben. Zum Beispiel ändert sich die Wahrscheinlichkeitsdichte, wenn man aus einer endlichen Grundgesamtheit Stichproben entnimmt ohne zurückzulegen (Karten aus einem Kartenstapel usw.).

4.2 Eigenschaften von Schätzwerten

Schätzwerte S sind Funktionen der Messwerte (Stichprobenfunktion):

$$S = S(x_1, \ldots, x_n), \tag{4.6}$$

und sind damit selbst wieder Zufallsvariable (die nächste Messreihe ergibt im allgemeinen ein etwas anderes Resultat für S). Als Beispiel hatten wir in Abschnitt 3.4 (Beispiel 1) das arithmetische Mittel als Zufallsvariable behandelt:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.7}$$

Es gibt gewisse Freiheiten Schätzwerte zu definieren. Optimale Eigenschaften von Schätzwerten erhält man mit folgenden Forderungen:

1. **Erwartungstreue:** Unabhängig von der Anzahl der Messwerte soll der Erwartungs des Schätzwerts für einen Parameter λ gleich dem Parameter sein:

$$E(S_\lambda(x_1, \ldots, x_n)) = \lambda \tag{4.8}$$

In Abschnitt 3.4 (Beispiel 1) hatten wir gesehen, dass das arithmetische Mittel in diesem Sinne erwartungstreu (unverzerrt, unbiased) ist.

Beispiel: Als weiteres Beispiel wollen wir die Varianz einer Verteilung mit Mittelwert μ und Varianz σ aus einer Stichprobe abschätzen. Dazu betrachten wir zunächst den Erwartungswert der quadratischen Abweichungen vom Mittelwert der Stichprobe:

$$\begin{aligned} E\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) &= E\left(\sum_{i=1}^{n}(x_i - \mu + \mu - \bar{x})^2\right) \\ &= \sum_{i=1}^{n}\left[\underbrace{E\left((x_i-\mu)^2\right)}_{\sigma^2} - \underbrace{E\left((\bar{x}-\mu)^2\right)}_{\sigma^2/n}\right] \\ &= n\left[\sigma^2 - \tfrac{\sigma^2}{n}\right] = (n-1)\,\sigma^2 \end{aligned} \tag{4.9}$$

$$\Longrightarrow \frac{1}{n-1} E\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \sigma^2 \tag{4.10}$$

Dabei wurde für die Varianz des Mittelwertes der Stichprobe, σ^2/n, das Ergebnis von (3.45) benutzt. Der Ausdruck

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{4.11}$$

auch empirische Varianz genannt, ist also eine erwartungstreue Schätzung der Varianz der Verteilung, weil für alle n gilt:

$$E(s^2) = \sigma^2. \tag{4.12}$$

Interpretation des Faktors $1/(n-1)$: Aus den n unabhängigen Messungen wurde zunächst der Parameter $\bar{x}$ bestimmt, dadurch geht ein Freiheitsgrad für die Bestimmung weiterer Parameter verloren. Die Anzahl der Freiheitsgrade ist die Anzahl der unabhängigen Messungen minus der Anzahl der bestimmten Parameter, hier also $n_F = n-1$. Aus der zweiten Zeile in (4.9) sieht man auch, dass die Minderung der Freiheitsgrade mit der Varianz σ^2/n des geschätzten Mittelwertes zusammenhängt.

2. **Konsistenz:** Eine Schätzung wird konsistent genannt, wenn die Varianz des Schätzwertes für große Stichproben gegen Null geht:

$$\lim_{n\to\infty} \sigma^2(S(x_1, \ldots, x_n)) = 0 \tag{4.13}$$

Beispiel: Für die Schätzung der Varianz des arithmetischen Mittels einer Stichprobe hatten wir in Abschnitt 3.4 (Beispiel 1) gefunden:

$$\sigma^2(\bar{x}) = \frac{\sigma^2(x)}{n} \tag{4.14}$$

Das arithmetische Mittel ist damit einen konsistente Schätzung des Mittelwertes der Verteilung.

3. **Effektivität:** Es seien S_1 und S_2 zwei Schätzungen des gleichen Parameters λ. Man sagt, S_1 ist effektiver als S_2, wenn gilt:

$$E\left[(S_1 - \lambda)^2\right] = \sigma^2(S_1) < E\left[(S_2 - \lambda)^2\right] = \sigma^2(S_2) \tag{4.15}$$

Diejenige Schätzung S_i, für die die Varianz minimal wird, nutzt also die vorhandenen Information am effektivsten.

Beispiel: Die Stichprobenfunktionen

$$S = \sum_{i=1}^{n} a_i\, x_i \qquad \text{mit } \sum_{i=1}^{n} a_i = 1 \tag{4.16}$$

sind für sonst beliebige a_i erwartungstreue Schätzungen des Mittelwertes μ:

$$E(S) = E\left(\sum_{i=1}^{n} a_i\, x_i\right) = \sum_{i=1}^{n} a_i\, E(x_i) = \sum_{i=1}^{n} a_i\, \mu = \mu \tag{4.17}$$

Es stellt sich aber heraus, dass S für $a_i = 1/n$ minimale Varianz hat, wenn alle Varianzen gleich sind, $\sigma_i = \sigma$ für alle x_i. Dann ergibt sich:

$$\sigma^2(S) = \sum_{i=1}^{n} a_i^2\, \sigma^2(x_i) = \sigma^2(x) \sum_{i=1}^{n} a_i^2 \tag{4.18}$$

Es bleibt also zu zeigen, dass $A = \sum a_i^2$ für $a_i = 1/n$ minimal wird. Durch die Bedingung $\sum_{i=1}^{n} a_i = 1$ sind nur $n-1$ der a_i unabhängig, so dass sich ein a_i eliminieren lässt:

$$A = \sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n-1} a_i^2 + \left(1 - \sum_{i=1}^{n-1} a_i\right)^2 \tag{4.19}$$

Die Extremwertbedingung ergibt:

$$\frac{\partial A}{\partial a_i} = 2\,a_i - 2\,\underbrace{\left(1 - \sum_{i=1}^{n-1} a_i\right)}_{a_n} = 2\,(a_i - a_n) = 0 \implies a_i = a_n \implies a_i = \frac{1}{n} \quad \forall\, i \tag{4.20}$$

4. **Robustheit:** Die Schätzwerte sollen möglichst gegen Annahmen "falscher" Verteilungen stabil sein.

   Zum Beispiel sind apparative Auflösungen nicht immer gauss-förmig, sondern haben zusätzlich zu einem Gauss-Anteil "nicht-gaussische" Ausläufer. Um stabile Parameter für Mittelwert und Auflösung zu erhalten, hilft häufig ein Abschneiden nach oben und unten (zum Beispiel könnte man die jeweils 20% kleinsten und größten Werte einer Messung wegschneiden). Eine andere Möglichkeit ist, die Verteilung der Messwerte mit einer angenommenen Verteilung in einem begrenzten Bereich anzupassen. Zum Beispiel passt man häufig Gauss-Kurven an Auflösungsverteilungen innerhalb 1 bis 2 Standardabweichungen um den Mittelwert an.

   Beispiel: In den meisten Teilchenexperimenten werden Energieverlustmessungen (dE/dx) zur Identifikation der Teilchen durchgeführt. Da die Fluktuationen sehr groß sein können und die dE/dx-Verteilung ('Landau-Verteilung') lange Ausläufer zu hohen Energien hat, werden sehr viele Messungen gemacht, manchmal einige hundert, und dann gemittelt. Der Mittelwert wird deutlich stabiler, wenn man zum Beipiel die kleinsten 10% und die größten 20% der Messwerte wegschneidet ('truncated mean').

   Robustheit ist schwieriger als die anderen Kriterien für Schätzungen zu behandeln, weil man hier "Unwissen" zu berücksichtigen versucht.

4.3 Stichproben aus Normalverteilungen; χ^2-Verteilung

Wir betrachten Stichproben $(x_1, \ldots, x_n)$ vom Umfang n aus einer Normalverteilung

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}} \tag{4.21}$$

mit Mittelwert μ und Standardabweichung σ. Dann folgt die Stichprobenfunktion

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \tag{4.22}$$

folgender Verteilung ($\lambda = n/2$):

$$f(\chi^2) = \frac{1}{\Gamma(\lambda)\, 2^\lambda} \, (\chi^2)^{\lambda-1} \, e^{-\frac{\chi^2}{2}}. \tag{4.23}$$

Die Γ-Funktion ist tabelliert zu finden. Mit der Definition

$$\Gamma(x+1) = \int_0^\infty t^x \, e^{-t} \, dt \tag{4.24}$$

findet man folgende Eigenschaften:

$$\begin{array}{rcll} \Gamma(1) & = & 1 & \\ \Gamma(x+1) & = & x\,\Gamma(x) & \\ \Gamma(n+1) & = & n! & \text{(n ganzzahlig)} \end{array} \tag{4.25}$$

Der Beweis, dass die in (4.22) definierte Größe χ^2 der Verteilung (4.23) folgt, ist zum Beispiel in [1] nachzulesen.

Erwartungswert und Varianz der χ^2-Verteilung: Den **Erwartungswert von** χ^2 erhält man aus (4.22) mit $\sigma^2 = E((x_i - \mu)^2)$:

$$E(\chi^2) = n, \tag{4.26}$$

wobei n hier die Anzahl der Messungen und im allgemeinen die Anzahl der Freiheitsgrade ist.

In den meisten Fällen ist der Parameter μ in der χ^2-Funktion (4.22) nicht bekannt und wird durch den Mittelwert der Stichprobe $\bar{x}$ geschätzt. Die χ^2-Funktion wird damit entsprechend definiert:

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} \tag{4.27}$$

Mit der empirischen Varianz s^2 ergibt sich:

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} = (n-1)\,\frac{s^2}{\sigma^2} \tag{4.28}$$

Da der Erwartungswert von s^2 nach (4.12) gleich σ^2 ist, ist der Erwartungswert der χ^2-Funktion bezüglich $\bar{x}$:

$$E(\chi^2) = n - 1 = n_F \tag{4.29}$$

Im allgemeinen wird in (4.27) $\bar{x}$ der Erwartungswert der Messgröße x_i sein, der eventuell von mehreren geschätzten Parametern abhängt, zum Beispiel wenn an die x_i eine Ausgleichsfunktion angepasst wird (siehe nächstes Kapitel). Die Anzahl der

Freiheitsgrade ist dann allgemein die Anzahl der Messwerte minus die Anzahl n_P der aus der Stichprobe bestimmten Parameter:

$$n_F = n - n_P \tag{4.30}$$

Die **Varianz von** χ^2 ist [1]:

$$\sigma^2(\chi^2) = E\left((\chi^2)^2\right) - \left(E(\chi^2)\right)^2 = 2n. \tag{4.31}$$

Hier wie im folgenden soll $n = n_F$ als Anzahl der Freiheitsgrade, der Parameter der χ^2-Verteilung, verstanden werden.

Eigenschaften der χ^2-Verteilung: Beispiele von χ^2-Verteilungen für verschiedene n sind in Abb. 4.1 gezeigt. Bei $\chi^2 = 0$ findet man folgendes Verhalten:

$$\begin{array}{lrcl} n = 1: & f(\chi^2) & = & \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\chi^2}} e^{-\frac{\chi^2}{2}} \quad \rightarrow \quad \infty \quad \text{für } \chi^2 \rightarrow 0 \\ n = 2: & f(0) & = & \frac{1}{2} \\ n \geq 3: & f(0) & = & 0 \end{array} \tag{4.32}$$

Für $n = 1$ hat die χ^2-Verteilung also einen Pol bei $\chi^2 = 0$. Die Verteilungsfunktion $F(\chi^2)$ bleibt aber endlich.

Für große n wird die Verteilung zunehmend symmetrischer und geht, entsprechend dem 'zentralen Grenzwertsatz' (Abschnitt 2.6), in eine Normalverteilung mit $\mu = n$ und $\sigma = \sqrt{2n}$ über.

Stichproben aus nicht gleichen Normalverteilungen: Gegenüber (4.22) und (4.28) kann man die χ^2-Funktion auch auf Messwerte mit unterschiedlichen Erwartungswerten μ_i bzw. $\bar{x_i}$ und Standardabweichungen σ_i verallgemeinern:

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{4.33}$$

Das ist leicht einzusehen, weil die reduzierten Variablen

$$x_i' = \frac{x_i - \mu_i}{\sigma_i} \tag{4.34}$$

alle der gleichen Normalverteilung $N(0, 1)$ mit $\mu = 0$ und $\sigma = 1$ folgen.

Der χ^2-Test: Die Stichprobenfunktion χ^2 wird zum Testen der Zuverlässigkeit ('confidence level') einer Messung benutzt. Man erwartet, dass jeder Freiheitsgrad im Mittel eine Varianz σ^2 hat, also eine Einheit zum χ^2 beiträgt:

$$\chi^2/n_f \approx 1 \tag{4.35}$$

Größere Abweichungen von dieser Erwartung deuten darauf hin, dass das angenommenen Gauss-Model oder die Schätzung der Parameter μ, σ für die Daten nicht richtig sind oder dass es einen nicht-gaussischen Untergrund gibt.

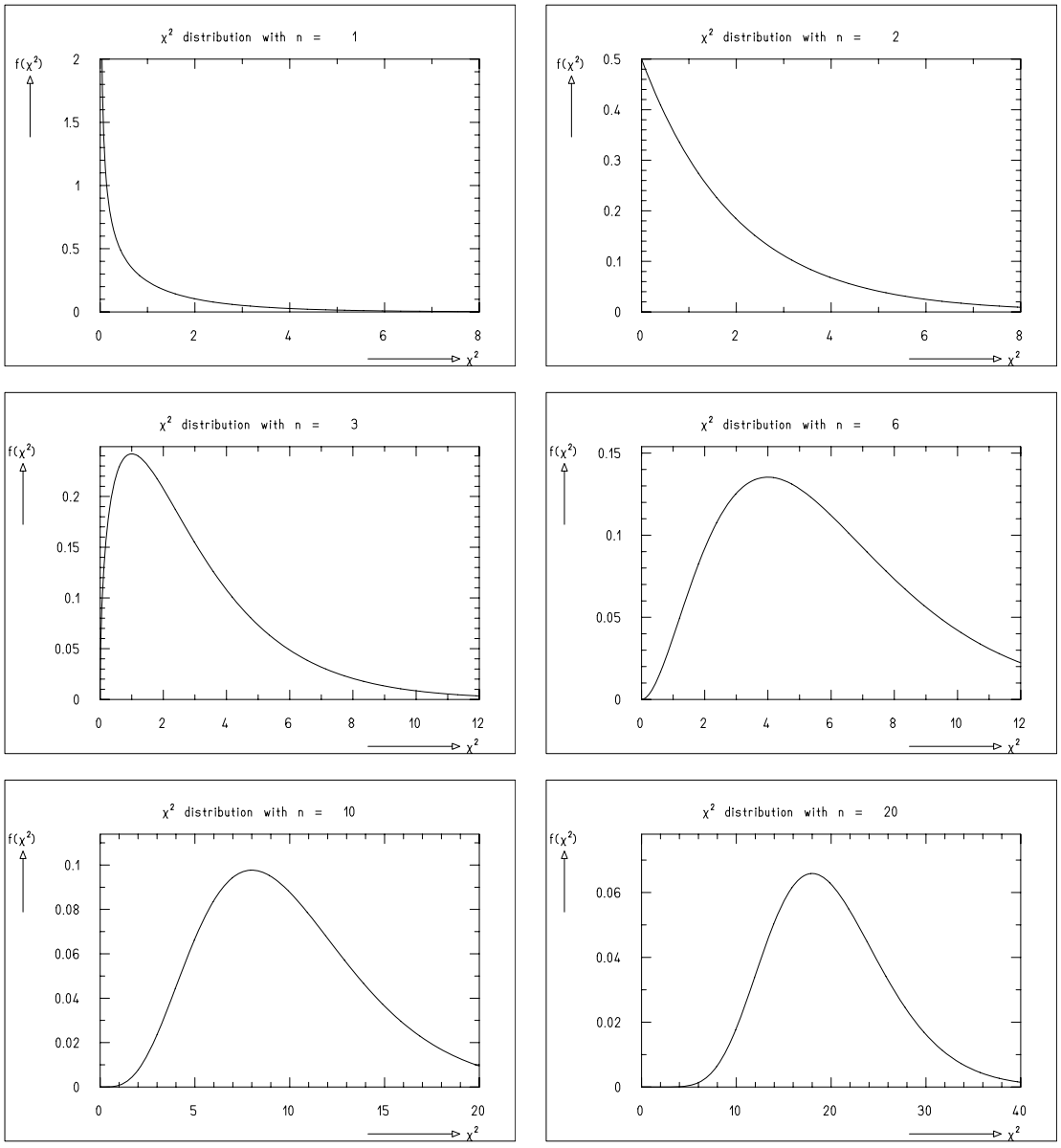


Abbildung 4.1: χ^2-Verteilungen für verschiedene Freiheitsgrade n (erstellt mit dem Programm s2sd [1]).

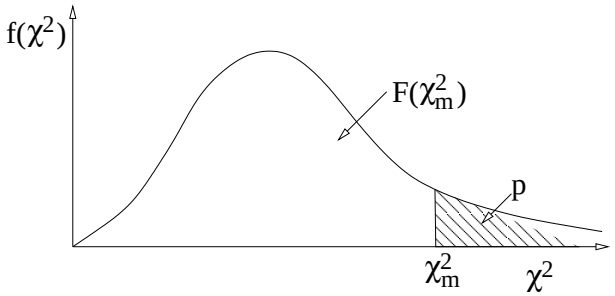


Abbildung 4.2: Definition des p-Wertes für einen gemessenen χ^2-Wert χ^2_m.

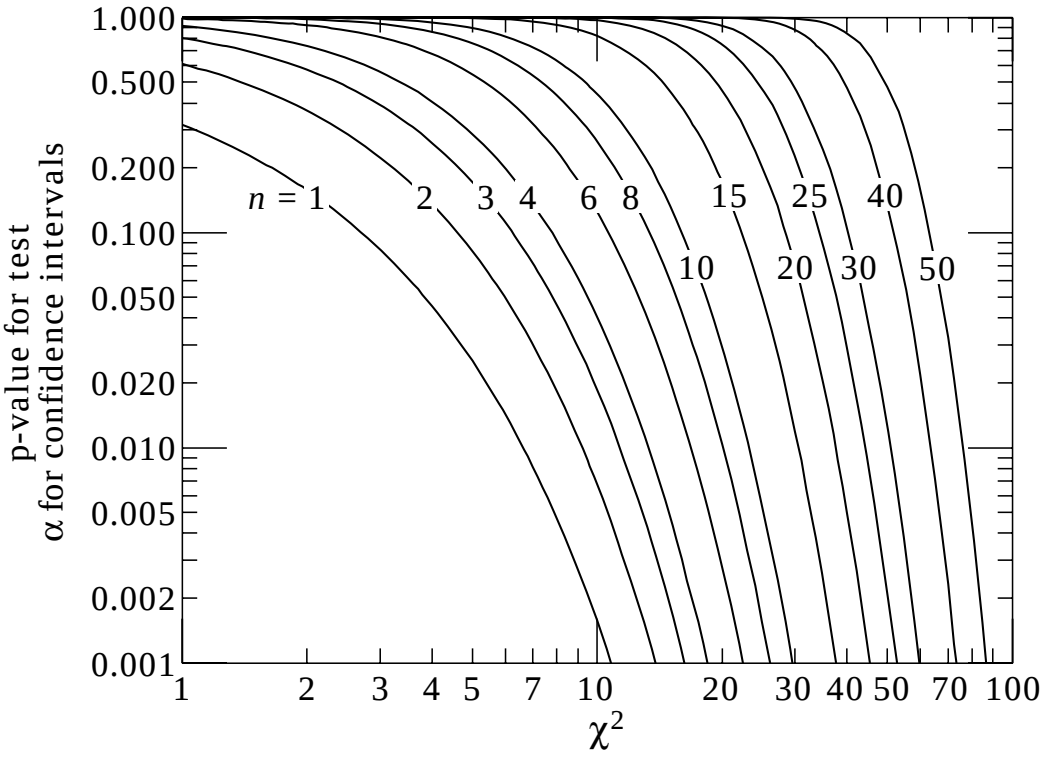


Abbildung 4.3: Der p-Wert beziehungsweise das Vertrauensniveau α als Funktion des χ^2-Wertes für verschiedene Freiheitsgrade $n = n_F$ (aus PDG [15]).

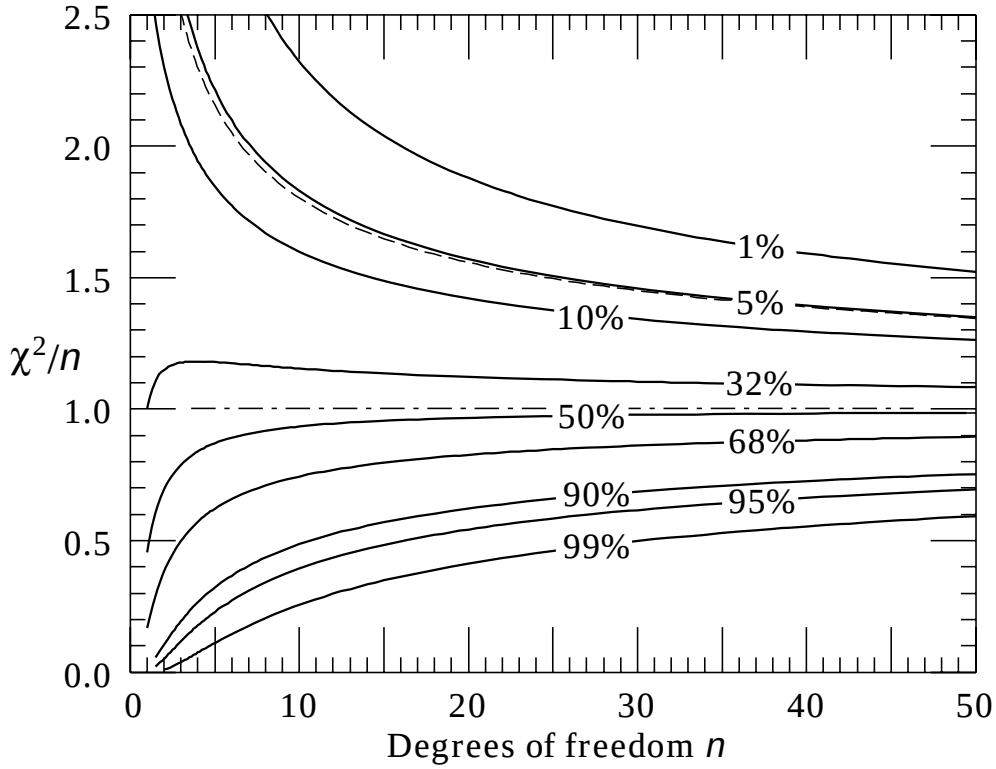


Abbildung 4.4: Das "reduzierte χ^2", χ^2/n_F, für verschiedene Vertrauensniveaus α als Funktion des Freiheitsgrades $n = n_F$. Für große n_F geht die $\alpha = 50\%$-Kurve asymptotisch gegen 1, das heisst, die χ^2-Verteilung wird immer symmetrischer (aus PDG [15]).

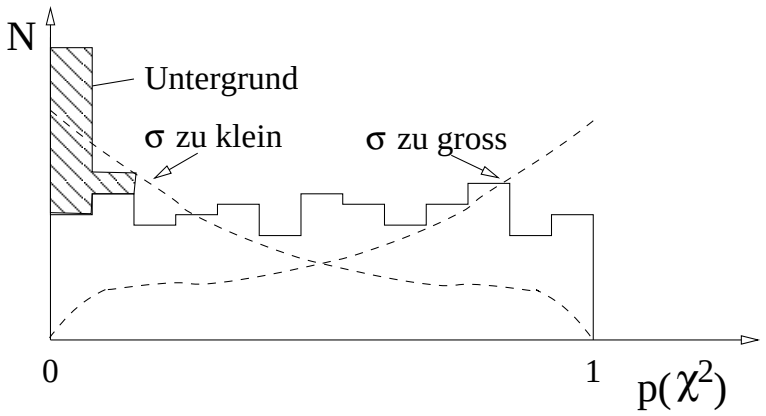


Abbildung 4.5: Typische Verteilung des p-Wertes. Über- oder Unterschätzungen der Fehler führen zu Abweichungen von der Gleichverteilung. Der Untergrund sammelt sich nahe $p = 0$.

Quantitativ gibt man die Zuverlässigkeit einer Messung beziehungsweise den Grad der Übereinstimmung mit dem Gauss-Modell durch Angabe des Integrals über die χ^2-Verteilung oberhalb des gemessenen χ^2-Wertes χ^2_m (Abb. 4.2) an:

$$p = 1 - F(\chi^2_m), \tag{4.36}$$

wobei F die Verteilungsfunktion ist. Der durch (4.36) definierte, so genannte p-Wert gibt also die Wahrscheinlichkeit an, dass bei den gemachten Annahmen eine Messung einen schlechteren χ^2-Wert, also $\chi^2 > \chi^2_m$, ergeben würde. Einen gemessenen χ^2-Wert kann man mit einem χ^2-Wert für ein vorgegebenes Vertrauensniveau α,

$$\alpha = 1 - F(\chi^2_\alpha), \tag{4.37}$$

vergleichen. Das Vertrauen in die Messung wird also größer, wenn das gemessene χ^2 kleiner wird. Bei welchem χ^2-Wert ein bestimmter p-Wert oder Vertrauensniveau erreicht wird, hängt von der Anzahl der Freiheitsgrade n_F ab. Man findet diese Angaben in Tabellen und graphischen Darstellungen (Abb. 4.3 und 4.4).

Die Wahrscheinlichkeitsdichte von $F(\chi^2)$ und damit auch von $p = 1 - F(\chi^2)$ ist gleichverteilt zwischen 0 und 1. Die Stichprobenfunktionen $F(\chi^2)$ und p sind dabei als Zufallsvariable zu betrachten. Wenn man sehr viele Messungen gemacht hat, die einen χ^2-Tests erfüllen sollen, kann man die gemessene p-Verteilung graphisch darstellen (Abb. 4.5). Abweichungen von einer Gleichverteilung haben meistens folgende Ursachen:

- das Gauss-Modell ist falsch oder
- die Standardabweichungen σ_i sind zu groß ($\Rightarrow$ Verschiebung zu großen p) oder
- die Standardabweichungen σ_i sind zu klein ($\Rightarrow$ Verschiebung zu kleinen p) oder
- es gibt nicht-gaussischen Untergrund.

Der Untergrund häuft sich bei kleinen Werten von p und kann mit einem Schnitt auf p entfernt werden (typische Forderung: $p > \alpha$ mit $\alpha = O(1\%)$).

> **Beispiel:** In Teilchenreaktionen werden in der Regel die Impulse und Richtungen der der beobachteten Teilchen mit gewissen Fehlern gemessen. Zusammen mit einer Hypothese für die Massen kann man Impuls- und Energieerhaltung mit einem χ^2-Test überprüfen. Ereignisse, bei denen wenigstens ein Teilchem dem Nachweis entgangen ist, werden sich bei einem kleinen Vertrauensniveau p-Wert ansammeln.

Man sollte sich klar machen, dass grundsätzlich alle Werte von p gleich häufig auftreten. Es ist also nicht von vornherein ein Wert von p nahe 1 besser als einer nahe 0. Selektionsschnitte auf p sollten ausschließlich durch das Untergrundverhalten bestimmt sein.

Die Bestimmung von Vertrauensintervallen wird im Zusammenhang mit Maximum-Likelihood-Schätzungen (Kapitel 6) und im speziellen Kapitel über Signifikanzanalysen (Kapitel 8) noch einmal aufgegriffen.

Kapitel 5

Monte-Carlo-Methoden

5.1 Einführung

Als Monte-Carlo-Methoden (MC-Methoden) werden Verfahren bezeichnet, mit denen numerische Probleme mit Hilfe von wiederholtem Ziehen von Zufallsstichproben aus bekannten Verteilungen gelöst werden. Diese Methoden werden häufig zur Simulation von mathematischen, physikalischen, biologischen, technischen oder ökonomischen Systemen benutzt, insbesondere wenn deterministische Algorithmen zu aufwendig oder vielleicht garnicht möglich sind.

Komplexe Simulationsprogramme, wie zum Beispiel die Simulation von Luftschauern hochenergetischer kosmischer Strahlung, die Simulation von Klimamodellen oder eines Öko-Systems, benötigen leistungsfähige Computer. Trotz des enormen Anstiegs von Schnelligkeit und Kernspeicherplatz der Rechner in den letzten Jahren sind viele Probleme nur mit vereinfachenden Annahmen zu simulieren. Zum Beispiel können globale Klimamodelle erst seit ein paar Jahren mit einigermaßen aussagekräftigen Ergebnissen simuliert werden.

Typische Anwendungen findet die MC-Methode zur Lösung folgender Probleme:

- Numerische Lösung von Integralen: viele Anwendungen lassen sich letztlich auf die Lösung von Integralen zurückführen. Zum Beispiel ist die Nachweiswahrscheinlichkeit eines Detektors für eine bestimmte Teilchenreaktion definiert als ein Integral über den Phasenraum der Reaktion in den Grenzen der Akzeptanz des Detektors gewichtet mit Verlustwahrscheinlichkeiten für einzelne Teilchen (in der Realität stellt sich das Problem im Allgemeinen noch komplexer dar, zum Beispiel durch kinematische Migrationen durch Streuung und Energieverlust).

- Simulation von dynamischen Prozessen: zum Beispiel Bewegungsabläufe von mechanischen Systemen in der Technik, Produktionsabläufe in der Wirtschaft oder die Entwicklung des Wetters.

- Simulation von Gleichgewichtszuständen, zum Beispiel in der statistischen Physik oder bei dem Einsatz bestimmter Typen neuronaler Netze. Diese Anwendung ist hier getrennt aufgeführt, weil dafür spezielle Methoden entwickelt wurden (zum Beispiel der Metropolis-Algorithmus).

- Statistische Untersuchung von Zufallsverteilungen, die analytisch nicht oder nur schwer zu behandeln sind. Dazu gehört zum Beispiel auch die Bestimmung von Fehlern einer Messung indem man das Experiment vielfach simuliert und den Fehler durch die Schwankung der simulierten Ergebnisse abschätzt ('bootstrap' Methode).

Auch vor der Entwicklung leistungsfähiger Computer wurden Simulationen zur Lösung komplexer mathematischer Probleme als 'analoge Simulationen' eingesetzt, wie zum Beispiel die Optimierung von Fahrzeugformen in Windkanälen oder die Lösung gekoppelter Differentialgleichungen mit Pendelsystemen. Ein schönes Beispiel, dass auf zufälligen Stichproben beruhende Simulationen auch ohne Computer gemacht werden können, ist das Buffonsche Nadelexperiment zur Bestimmung der Zahl π:

> **Beispiel:** Auf ein Blatt Papier mit parallelen Linien im Abstand g werden Nadeln der Länge l so geworfen, dass ihre Lage und Richtung zufällig ist. Die Wahrscheinlichkeit, dass eine Nadel eine Linie kreuzt, hängt wegen der Rotationssymmetrie der Nadelorientierung mit der Zahl π zusammen:
>
> $$p = \frac{2l}{g\pi} \qquad \Longrightarrow \qquad \pi = \frac{2l}{gp} \tag{5.1}$$
>
> Die Wahrscheinlichkeit p wird nun experimentell durch das Werfen von Nadeln bestimmt.

Häufig entspricht die Aufgabenstellung der Lösung eines Integrals in einem multidimensionalen Raum mit komplizierten Integrationsgrenzen. Mit der MC-Methode wird das Integral gelöst, indem man diskrete Punkte in dem Raum nach dem Zufallsprinzip würfelt.

Das Integral kann nun auf verschiedene Weise ausgewertete werden. Nach der einfachsten Methode werden die Punkte gleichverteilt in dem Raum erzeugt und die Integrandenfunktion wird an den diskreten Punkten aufaddiert. Das entspricht der numerischen Lösung des Integrals durch eine endliche Summe über Intervalle. Hier könnte man fragen, ob es nicht grundsätzlich am günstigsten ist, eine feste Intervalaufteilung zu machen, wodurch der Fehler des Integrals mit der Anzahl N der Intervalle abfallen würde. Dagegen fällt bei einer zufälligen Wahl der Punkte der Fehler nur wie $1/\sqrt{N}$ ab. Bei einer einzelnen Dimension ist eine gleiche Verteilung der Punkte auf jeden Fall optimaler. Allerdings ist es in höheren Dimensionen für auf einem regulären Gitter angeordnete Punkte nicht mehr richtig, dass der Fehler mit $1/N$ abnimmt, was an den Korrelationen der untereinander liegt. Da bei der MC-Methode der Fehler immer mit $1/\sqrt{N}$ abnimmt, wird die MC-Methode dimensionsabhängig (und problemabhängig) optimaler ('Monte-Carlo-Paradoxon'). Darüber hinaus bietet die MC-Methode bei komplexen Problemen viele bedenkenswerte weitere Vorteile. Ein ganz wichtiger Vorteil der Benutzung von Zufallsvariablen ist die Möglichkeit, die Simulation beliebig fortzusetzen und damit die Genauigkeit zu erhöhen. Bei diskreter Intervalschachtelung würde ein nächster Schritt mindestens eine Halbierung der Intervalabmessungen bedeuten, was die Rechenzeit bei einer Dimension n um einen Faktor 2^n verlängern würde (also schon ein Faktor von etwa 1000 bei 10 Dimensionen).

Bei der Standard-MC-Methode zur Lösung eines Integrals werden die Punkte in dem Raum mit der durch die normierte Integrandenfunktion gegebenen Wahrscheinlichkeitsdichte erzeugt. Man erhält dann "Ereignisse" mit der entsprechenden Wahrscheinlichkeitsdichte, die dann auch weiteren Analysen unterworfen werden können, was eine hohe Flexibilität bei dem Vergleich der Simulation mit gemessenen Daten ergibt.

In diesem Kapitel werden verschiedene Methoden zur Erzeugung von Stichproben mit bestimmten Wahrscheinlichkeitsdichten und optimale Methoden zur Bestimmung von Integralen besprochen. Für die Anwendung der MC-Methode benötigt man Generatoren von (Pseudo)-Zufallszahlen, deren Eigenschaften wir zunächst kurz besprechen wollen. Wir orientieren uns in diesem kapitel besonders an [4]; einen guten Überblick gibt auch der Artikel [16].

5.2 Zufallszahlengeneratoren

In der Regel geht man von einem Zufallszahlengenerator aus, der bei jedem Aufruf eine neue Zahl z, die im Intervall $[0, 1]$ gleichverteilt ist, zurückgibt. Aus diesen Zufallszahlen werden die Zufallsvariablen des betrachteten Problems erzeugt.

Die Zufallszahlen werden fast ausschließlich durch geeignete Algorithmen als 'Pseudozufallszahlen' im Rechner erzeugt. Ein Problem ist, dass wegen der digitalen Darstellung der reellen Zahlen mit einer endlichen Bit-Anzahl, die Zahlengeneratoren im allgemeinen eine **Periodizität** haben können. Man versucht die Periode möglichst lang zu machen, um große Ereignismengen unabhängig erzeugen zu können. Gute Generatoren sollten auch keine Korrelationen in der Abfolge der Zufallszahlen aufweisen, um Muster in einem multi-dimensonalen Raum zu vermeiden.

Da Zufallszahlengeneratoren im Prinzip 'deterministisch' sind, ist eine Wiederholbarkeit von Rechnungen, die statistisch unabhängige Fortsetzung und die parallele Ausführung auf verschiedenen Rechnern möglich. Die Zufallszahlengeneratoren liefern dafür so genannte 'seeds', Zahlen mit denen man einen Generator an wohldefinierten Stellen einer Zufallszahlenfolge initiieren kann.

5.2.1 Multiplikativ kongruentielle Generatoren

Es gibt eine Vielzahl von Algorithmen zur Erzeugung von Pseudozufallszahlen. Viele der in der Vergangenheit sehr popolären Zufallsgeneratoren gehören zur Klasse der multiplikativ oder gemischt kongruentiellen Generatoren (engl. linear congruential generator, LCG). Das Prinzip soll hier kurz erläutert werden. Eine Zufallszahl erzeugt ein LCG über die Rekursionsrelation:

$$x_{i+1} = (ax_i + b) \mod m \tag{5.2}$$

wobei Modul m, Faktor a, Inkrement b und Startwert x_1 die Zufallsequenz vollständig bestimmen. In der Praxis hängen die Eigenschaften eines LCG sensitiv von der Wahl dieser Parameter ab. Für $m = 2^k$ mit $k \geq 4$ ist die maximale Periode eines LCG bei optimaler Wahl der Parameter $m/4$.

LCG haben deutliche Schwächen, z.B. kleine Periode der Sequenz (und noch geringere Perioden für nicht signifikante Stellen), sowie deutliche Korrelation von

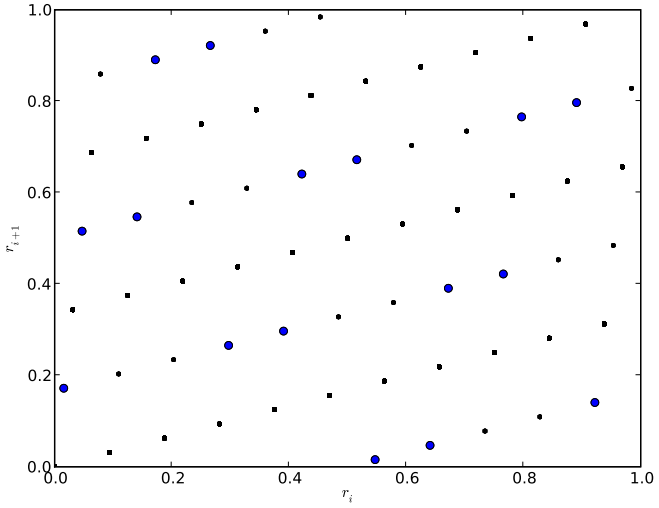


Abbildung 5.1: Iterationsfunktion eines LCG mit $m = 64$, $a = 11$, $b = 0$. Für diesen LCG gibt zwei Sequenzen mit Periode 16 (d.h. der maximalen Periode für $m = 64$), die zusammen alle ungeraden Zahlen $< m$ enthalten. Die fetten Punkte entsprechen einer dieser Sequenzen. Gerade Zahlen als Startwerte liefern (teilweise deutlich) kürzere Perioden.

aufeinander folgenden Zufallszahlen. Letztere Eigenschaft führt bei Erzeugung von mehrdimensionalen Tupeln zu ungewünschten Stukturen ("Hyperebenen"). Abbildung 5.1 zeigt die Verteilung von aufeinanderfolgenden Zufallzahlen für $m = 64$, $a = 11$, $b = 0$. Einige, aber nicht alle, Probleme der LCG werden durch Verwendung von verallgemeinerten kongruentiellen Generatoren umgangen.

5.2.2 Mersenne-Twister

Als der "Zufallszahlengenerator der Wahl" hat sich in den letzten Jahren der Mersenne-Twister etabliert. Der Algorithmus gehört zu der Klasse der *twisted generalised feedback shift register*. Er zeichnet sich unter anderem durch eine extrem lange Periode aus ($2^{19937} - 1 \approx 4.3 \cdot 10^{6001}$), erzeugt sehr gute Gleichverteilungen (nachgewiesen bis zur Dimension 623), und ist dennoch schneller als andere (hinreichend gute) Zufallsgeneratoren. Er wird in allen modern Programmbibliotheken wie z.B. der *GNU Scientific Library* eingesetzt. Der Zustand des Generators wird eindeutig durch 624+1 Integer-Zahlen (32bit) beschrieben, für deren Initialisierung in der Regel ein einfacher LCG Algorithmus verwendet wird.

5.2.3 Quasi-Zufallszahlen

Um das Konvergenzverhalten von Monte Carlo Integrationsalgorithmen zu verbessen, werden gelegenlich Quasi-Zufallszahlen eingesetzt. Quasi-Zufallszahlen Generatoren (QZG) sind deterministische Algorithmen die eine gleichmäßigere Füllung des Integrationsvolumen mit Punkten garantiert, ohne dass die Punkte wie auf einem re-

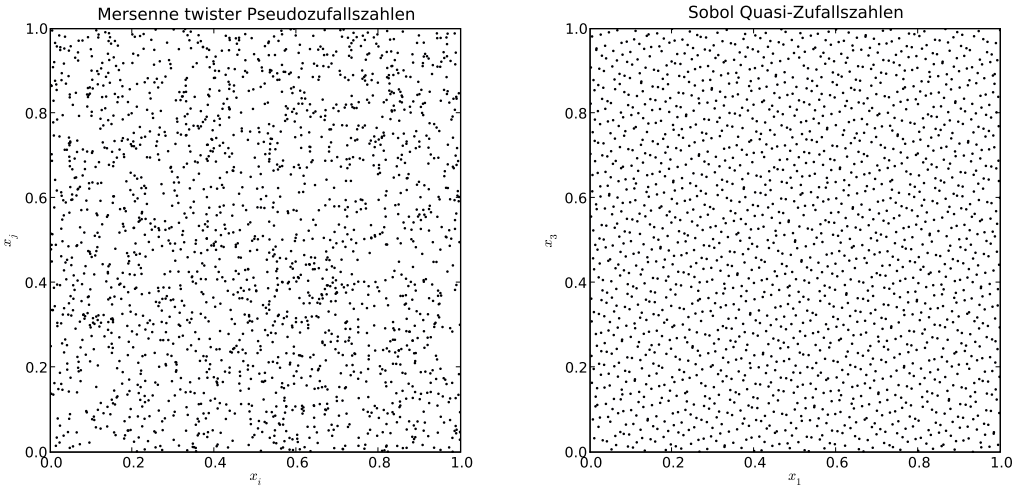


Abbildung 5.2: Pseudozufallszahlen erzeugt durch einen Mersenne-Twister, und Quasi-Zufallzahlen nach der Sobol-Sequenz in 3 Dimensionen.

gulären Gitter korreliert sind. In Abbildung 5.2 werden Zufallsverteilung erzeugt mit dem Mersenne-Twister mit Quasi-Zufallzahlen verglichen. Quasi-Zufallzahlen basieren häufig auf der van-der-Corput-Sequenz (Halton-Sequenz) oder auf der Sobol-Sequenz.

Quasi-Zufallzahlen eignen sich **nur** zum Integrieren. Auch muss die Dimension des Problems bereits bei der Erzeugung der Zufallszahlen feststehen. Zur Verbesserung der Genauigkeit der Integration muss an der Stelle der Sequenz, an der unterbrochen wurde, fortgesetzt werden. QZG stellen also einen besonderen Anspruch an die Disziplin des Programmierers.

5.3 Monte-Carlo-Erzeugung von Ereignissen

Einfache Beispiele für die Erzeugung von Ereignissen entsprechend vorgegebenen Zufallsverteilungen sind bereits in den Kapiteln 1 und 3 gegeben worden. Der Vollständigkeit halber werden wir die dort eingeführten Inversions- und Hit-and-Miss-Methoden wiederholen und auf multi-dimensionale Verteilungen erweitern.

Das prinzipielle Problem ist die Zuordnung von Sätzen $\vec{z} = (z_1, z_2, \ldots z_n)$ von jeweils im Einheitsinterval gleichverteilten Zufallsvariablen zu den Variablen $\vec{x} = (x_1, x_2, \ldots x_n)$ einer Wahrscheinlichkeitsdichte $f(\vec{x})$, so dass die entsprechenden Ereignisse $\vec{x}_j$ $(j = 1, \ldots, N)$ Stichproben der Verteilung $f(\vec{x})$ sind. Das ist im Allgemeinen keine leichte Aufgabe, insbesondere wenn die Variablen x_i untereinander korreliert sind.

5.3.1 Inversionsmethode

Eine Methode haben wir bereits in Abschnitt 1.3.1 für eine einzelne Variable eingeführt. Zusammengefasst sind die Ergebnisse: Aus der Forderung, dass bei einer

Transformation einer Zufallsvariablen x mit einer Wahrscheinlichkeitsdichte $f(x)$ auf eine im Einheitsinterval gleichverteilte Zufallsvariable z die differentielle Wahrscheinlichkeit gleich bleiben muss,

$$f(x)dx = dz, \tag{5.3}$$

ergibt sich

$$z = F(x) \qquad \text{und die Umkehrung :} \qquad x = F^{-1}(z), \tag{5.4}$$

wobei $F(x)$ die Verteilungsfunktion von $f(x)$ ist (die ja zwischen 0 und 1 gleichverteilt ist). Falls F analytisch invertierbar ist, ist damit das Problem gelöst. Es ist im Prinzip auch möglich, numerisch zu invertieren, was aber häufig zeitaufwendiger ist, als die Anwendung anderer Methoden, die zum Teil im Folgenden besprochen werden.

Im mehr-dimensionalen Fall muß man im Allgemeinen schrittweise vorgehen, beginnend zum Beispiel mit der Erzeugung der Variablen x_1. Aus der Randverteilung $h_1(x_1)$ (siehe (3.6)) zu dieser Variablen kann man mit der Inversionsmethode die Zufallsvariable $x_{1,j}$ (j ist das jeweilige Ereignis) erzeugen. Da im Allgemeinen die übrigen Variablen von x_1 abhängig sein können, muss man im weiteren die bedingte Wahrscheinlichkeit für $x_1 = x_{1,j}$ wie in (3.8) definiert betrachten:

$$f^*(x_2,\, x_3,\, \ldots,\, x_n | x_1 = x_{1,j}) = \frac{f(x_1 = x_{1,j}, x_2, \ldots, x_n)}{h_1(x_1 = x_{1,j})}$$

Diese Verteilung wird nun benutzt, um mit der Inversionsmethode die Variable x_2 zu erzeugen. Die Schritte wiederholen sich bis zur letzten Variablen x_n. Am einfachsten ist natürlich, wenn in jedem Schritt die Invertierung analytisch gemacht werden kann; grundsätzlich ist aber auch eine numerische Invertierung möglich. Das Verfahren lässt sich, wie in Abschnitt 1.3.1 besprochen, entsprechend auch auf diskrete Verteilungen anwenden.

Beispiel: Ein Beispiel für die eindimensionale Lebensdauerverteilung ist in Abschnitt 1.2.1 gegeben worden (Gleichungen (1.35 - 1.37)).

Für eine zwei-dimensionalen Verteilung wurde ein Beispiel in Abschnitt 3.6 gegeben (allerdings für den einfacheren Fall unabhängiger Variablen).

Für eine zwei-dimensionale, unkorrelierten Gauss-Verteilung lässt sich durch Transformation der kartesischen Koordinaten (x_1, x_2) auf Polarkoordinaten (r, ϕ) (Box-Müller-Transformation) analytisch invertierbare Verteilungsfunktionen erhalten (was im Eindimensionalen nicht der Fall ist). Man erhält dann mit der oben angegebenen Methode und nach Rücktransformation die unabhängig gauss-verteilten Zufallszahlen (x_1, x_2) als Funktion der im Einheitsinterval gleichverteilten Zufallzahlen (z_1, z_2):

$$\begin{aligned} x_1 &= \sigma\sqrt{-\ln z_1^2} \cdot \sin(2\pi z_2) \\ x_2 &= \sigma\sqrt{-\ln z_1^2} \cdot \cos(2\pi z_2) \end{aligned} \tag{5.5}$$

Tabelle 5.1: Erzeugungsalgorithmen von Zufallszahlen einiger wichtiger Wahrscheinlichkeitsverteilungen aus gleichverteilten Zufallszahlen z in $[0, 1[$.

Wahrscheinlichkeitsdichte	Wertebereich	Algorithmus
$f(x) = \dfrac{1}{b-a}$	$[a, b[$	$x = (b-a) \cdot z + a$
$f(x) = 2x$	$[0, 1[$	$x = \max(z_1, z_2)$ or $x = \sqrt{z}$
$f(x) \sim x^{r-1}$	$[a, b[$	$x = [(b^r - a^r) \cdot z + a^r]^{1/r}$
$f(x) \sim \dfrac{1}{x}$	$[a, b[$	$a \cdot (b/a)^z$
$f(x) = \dfrac{1}{x^2}$	$]1, \infty]$	$x = 1/z$
$f(x) = \dfrac{1}{k} e^{-x/k}$	$]0, \infty]$	$x = -k \ln z$
$f(x) = xe^{-x}$	$]0, \infty]$	$x = -\ln(z_1 \cdot z_2)$
$f(x) = -\ln x$	$[0, 1[$	$x = z_1 \cdot z_2$
Gauss: $f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{x^2}{2\sigma^2}}$	$[-\infty, \infty]$	$x = \sigma\sqrt{-\ln z_1^2} \cdot \cos(2\pi z_2)$
Breit–Wigner: $f(x) = \dfrac{\Gamma}{2\pi} \cdot \dfrac{1}{(x-\mu)^2 + (\Gamma/2)^2}$	$[-\infty, \infty]$	$x = [\tan \pi(z - 0.5)] \cdot \Gamma/2 + \mu$

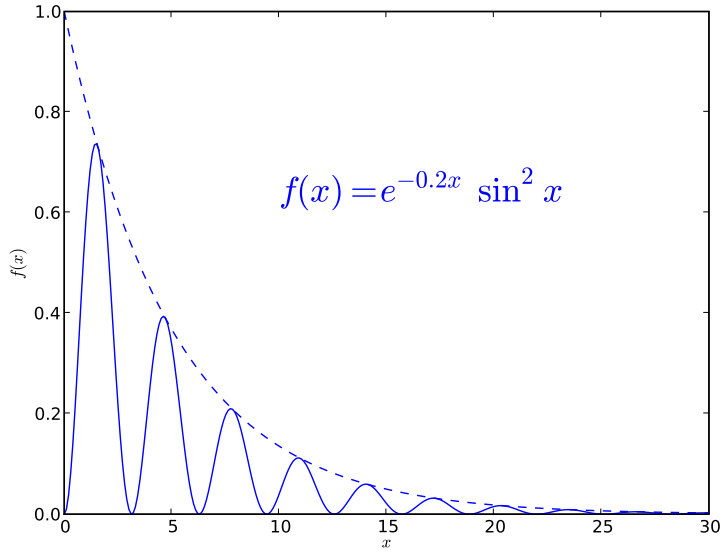


Abbildung 5.3: Verteilung mit Majorante.

5.3.2 'Hit-and-Miss'-Verfahren

Das in Abschnitt 1.3.2 besprochene 'Hit-and-Miss'-Verfahren für eine einzelne Variable lässt sich einfach auf mehrere Dimensionen übertragen: Man definiert in n Dimensionen einen Quader der den gesamten Wertebereich der Zufallsvariablen x_i enthält. In diesem Quader würfelt man gleichverteilt Punkte; zu jedem Punkt wird eine zusätzliche Variable z_f gleichverteilt zwischen 0 und f_{max}, dem Maximalwert der Funktion $f(\vec{x})$ (oder größer), gewürfelt. Wenn $z_f < f_{max}$ gilt, wird der Punkt akzeptiert ('hit') oder verworfen ('miss').

Wie im ein-dimensionalen Fall gilt auch hier, dass die Methode ineffizient wird, wenn es große Unterschiede in den Funktionswerten gibt. Das lässt sich verbessern, indem das gesamte Quadervolumen in kleinere Quader mit geringeren Funktionsschwankungen unterteilt wird. Die Anzahl der Versuche in jedem Unterquader ist proportional zu dessen Volumen. Man wird also zunächst durch Würfeln in der diskreten Verteilung $p_i = V_i/V_{tot}$ (entsprechend Abb. 1.7) das entsprechende Volumen auswählen, in dem man das nächste Ereignis erzeugt. Der Übergang von diskreten Intervalschachtelungen zu einer kontinuierlichen Umhüllenden der Wahrscheinlichkeitsdichte wird mit der im Folgenden beschriebenen Majorantenmethode vollzogen.

5.3.3 Majorantenmethode

Für eine Majorante $g(\vec{x})$ einer Funktion $f(\vec{x})$ gilt:

$$g(\vec{x}) \geq f(\vec{x}) \quad \forall\ \vec{x} \tag{5.6}$$

Man sucht sich nun eine Majorante $g(\vec{x})$ von $f(\vec{x})$, für die man einfacher als für $f(\vec{x})$ Ereignisse erzeugen kann, zum Beispiel mit der Inversionsmethode. Wie bei der 'Hit-

and-Miss'-Methode akzeptiert man dann ein Ereignis j mit der Wahrscheinlichkeit

$$w_j = \frac{f(\vec{x}_j)}{g(\vec{x}_j)}. \tag{5.7}$$

Dazu würfelt man gleichverteilt zwischen 0 und $g(\vec{x}_j)$ und verwirft die Ereignisse, wenn die gewürfelte Zahl größer als $f(\vec{x}_j)$ ist.

5.3.4 Wichtung der Ereignisse

Statt Ereignisse zu verwerfen, kann man einem Ereignis auch ein Gewicht geben. Bei der Majoranten-Methode ist es das in (5.7) definierte Gewicht $w_j = f(\vec{x}_j)/g(\vec{x}_j)$; beim einfachen 'Hit-and-Miss'-Verfahren ist es $w_j = f(\vec{x}_j)/f_{max}$ (die Majorante ist hier die Konstante f_{max}).

Auch wenn man zunächst ungewichtete Ereignisse erzeugt, kann es vorteilhaft sein, bei einer Datenanalyse auf jeden Fall zu jedem Ereignis einen Speicherplatz für ein Gewicht mitzuführen. Häufig möchte man nämlich unterschiedliche Modelle, zum Beispiel unterschiedliche Matrixelemente eines Wirkungsquerschnitts, austesten, was durch 'Umwichten' der Ereignisse einfach und ökonomisch möglich ist. Wurde ein Ereignis entsprechend der Wahrscheinlichkeitsdichte $f(\vec{x})$ erzeugt und soll nun der Wahrscheinlichkeitsdichte $g(\vec{x}_j)$ folgen, wird jedes Ereignisgewicht mit dem entsprechenden Verhältnis multipliziert:

$$w'_j = w_j \cdot \frac{g(\vec{x}_j)}{f(\vec{x}_j)} \tag{5.8}$$

Bei der Verarbeitung von gewichteten Ereignissen, zum Beispiel bei grafischen Darstellungen, müssen die Gewichte immer berücksichtigt werden. Das gilt insbesondere auch bei der Fehlerrechnung. Wenn man zum Beispiel N Einträge in einem Interval eines Histogramms hat, ist der Fehler bei ungewichteten Ereignissen $\sqrt{N}$, entsprechend der Poisson-Statistik. Bei gewichteten Ereignissen ist der entsprechende Eintrag

$$N_w = \sum_{j=1}^{N} w_j. \tag{5.9}$$

Den Fehler von N_w kann man durch Fehlerfortpflanzung bestimmen. Man nimmt dazu an, dass man N unabhängige Ereignisse hat, jedes mit dem Poisson-Fehler $\sigma_j = \sqrt{1} = 1$ (zu einem einzelnen Ereignis, $N_j = 1$). Dann ergibt die Fehlerfortpflanzung:

$$\sigma^2(N_w) = \sum_{j=1}^{N} \left(\frac{\partial N_w}{\partial N_j}\right)^2 \sigma_j^2 = \sum_{j=1}^{N} w_j^2 \tag{5.10}$$

Der relative Fehler ist

$$\frac{\sigma(N_w)}{N_w} = \frac{\sqrt{\sum_{j=1}^{N} w_j^2}}{\sum_{j=1}^{N} w_j}. \tag{5.11}$$

Gewichte sollten nicht stark variieren, weil sonst die statistischen Fluktuationen sehr groß werden können.

5.4 Monte-Carlo-Integration

Die Monte-Carlo-Methode kann zur numerischen Bestimmung von Integralen benutzt werden. Dazu werden in dem Definitionsbereich der Integrandenfunktion Zufallsereignisse generiert, deren Gesamtheit das Integral bestimmt. Im einzelnen können dazu die Methoden herangezogen werden, die im vorigen Abschnitt zu Erzeugung von Ereignissen benutzt wurden. Für die Bestimmung eines Integrals ist es wichtig, mit welcher Methode am effektivsten eine gewünschte Genauigkeit erreicht werden kann. Wir werden im Folgenden die verschiedene Methoden daraufhin untersuchen.

Wir nehmen an, dass die Integrandenfunktion $f(\vec{x})$ im Integrationsvolumen Ω^n nur positive oder nur negative Werte annimmt:

$$f(\vec{x}) \geq 0 \qquad \text{oder} \qquad f(\vec{x}) \leq 0, \qquad \vec{x} \in \Omega^n. \tag{5.12}$$

Falls das nicht der Fall ist, muss Ω^n in entsprechende Bereiche zerlegt werden, in denen f nur ein Vorzeichen hat. In solchen Bereichen kann $f(\vec{x})$ nach einer geeigneten Normierung als Wahrscheinlichkeitsdichte interpretiert werden, bezüglich der MC-Ereignisse generiert werden können.

5.4.1 Majoranten-Methode mit Hit-or-Miss

Mit dem Hit-or-Miss-Verfahren ergibt sich eine Schätzung des gesuchten Integrals I zu:

$$I = \int_{\Omega^n} f(\vec{x})\, dx_1 \ldots dx_n \approx I_{ref} \frac{k}{N} \tag{5.13}$$

Dabei ist I_{ref} das Integral einer Majorantenfunktion $g(x)$ über Ω^n (im einfachsten Hit-or-Miss-Verfahren ist $g(x) = f_{max}$ eine Konstante); N ist die Anzahl der in Ω^n generierten MC-Ereignisse und k die der akzeptierten.

Die Effizienz der MC-Erzeugung

$$\epsilon = \frac{k}{N} \tag{5.14}$$

ist ein Parameter der Binomialverteilung von k mit der Varianz

$$\sigma_k^2 = N\epsilon(1-\epsilon). \tag{5.15}$$

Damit kann der relative Fehler des Integrals abgeschätzt werden:

$$\frac{\sigma_I}{I} = \frac{\sigma_k}{k} = \sqrt{\frac{1-\epsilon}{k}}. \tag{5.16}$$

Der Fehler wird also klein, wenn die Anzahl der akzeptierten MC-Ereignisse groß wird, und hat mit $1/\sqrt{k}$ das erwartete Poisson-Verhalten. Der Fehler nimmt auch mit wachsender Effizienz ϵ ab und kann sogar im Grenzfall $\epsilon \to 1$ ganz verschwinden. Das bedeutet, dass die Majorantenfunktion $g(x)$ den Integranden $f(x)$ möglichst gut approximieren sollte. Falls eine analytisch integrierbare Majorantenfunktion nicht gefunden werden kann, kann der gesamte Integrationsbereich so zerlegt werden, dass in jedem einzelnen Untervolumen die Integrandenfunktion nicht stark schwankt und deshalb die Effizienzen hoch sein können. Wir werden dieses Verfahren unter den 'varianz-reduzierende Methoden' weiter unten näher betrachten.

5.4.2 MC-Integration mit Ereigniswichtung

Wenn das Ziel nur die Bestimmung des Integrals unter der Funktion $f(\vec{x})$ ist und die gleichzeitige Gewinnung einer Ereignisstichprobe keine Rolle spielt, gibt es eigentlich keinen Grund Ereignisse nach dem Hit-or-Miss-Verfahren zu verwerfen.

Der einfachste Fall ist die Summation von in Ω^n gleichverteilten Zufallsereignissen gewichtet mit ihren jeweiligen Funktionswerten. Das ist zwar sehr ähnlich dem Quadraturverfahren mit gleichmässigen Intervallen, die MC-Methode ist aber, wie bereits in der Einleitung zu diesem Kapitel angesprochen, bei höheren Dimensionen vorteilhafter. Mit dem Integrationsvolumen V ergibt sich die Schätzung des Integrals zu:

$$I \approx \frac{V}{N} \sum_{j=1}^{N} f(\vec{x}_j) = V\,\overline{f} \tag{5.17}$$

Mit der Varianz des Mittelwertes

$$\sigma_{\bar{f}}^2 = \frac{1}{N(N-1)} \sum_{j=1}^{N} (f(\vec{x}_j) - \bar{f})^2 \tag{5.18}$$

ergibt sich der relative Fehler des Integrals zu

$$\frac{\sigma_I}{I} = \frac{\sigma_{\bar{f}}}{\bar{f}} \tag{5.19}$$

5.4.3 Varianz-reduzierende Verfahren

Die Varianz des Integrals ist also proportional zu der Varianz der Funktion f im Integrationsvolumen. Deshalb sind Methoden zur Reduzierung der Varianz ein wichtiges Hilfsmittel bei der numerischen Integration mit der MC-Methode. Eine Möglichkeit der Varianzreduktion ist die Anwendung des Majoranten-Verfahrens. Die entsprechend der Majoranten $g(\vec{x})$ erzeugten Ereignisse werden ohne Verwerfen wie in (5.13), aber gewichtet, aufsummiert:

$$I = \int_{\Omega^n} f(\vec{x})\, dx_1 \ldots dx_n = I_{ref} \frac{1}{N} \sum_{j=1}^{N} w_j \tag{5.20}$$

Das entspricht (5.13), wenn man die Zahl der akzeptierten Ereignisse k durch $N_w = \sum_{j=1}^{N} w_j$, die Summe der Gewichte, ersetzt. Der Fehler des Integrals ist dann:

$$\frac{\sigma_I}{I} = \frac{\sigma_{N_w}}{N_w} = \frac{\sqrt{\sum_{j=1}^{N} w_j^2}}{\sum_{j=1}^{N} w_j}, \tag{5.21}$$

wobei (5.11) auf der rechten Seite eingesetzt wurde. Am geringsten wird der Fehler, wenn die Gewichte alle gleich sind (Beweis wie für die Effizienz des Mittelwertes, Gleichung (4.16) und folgende), das heißt, dass die Majorante g dem Integranden f sehr gut folgt. Es sei $w_j = 1/N$ für alle Ereignisse j. Dann ergibt sich für den relativen Integralfehler in (5.21):

$$\frac{\sigma_I}{I} = \frac{1}{\sqrt{N}}. \tag{5.22}$$

Das ist offensichtlich eine untere Grenze für den Fehler.

5.4.4 Stratified Sampling ('Geschichtete Stichproben')

Es sind verschiedene Methode der Varianzreduktion entwickelt worden, die nicht voraussetzen, dass man die zu integrierende Funktion gut kennt, insbesondere, dass man keine Majorante finden muss. Die Idee ist, Untervolumen so zu definieren, dass die Varianzen jeweils klein werden ('stratified sampling', 'geschichtete Stichproben'). Die Varianzen können beliebig klein gemacht werden, wenn man in beliebig viele Untervolumen aufteilt. Dem steht bei der praktischen Ausführung der Rechenzeitaufwand und der Bedarf an Speicherplatz limitierend entgegen. Die Frage stellt sich dann eher so: Wenn N Ereignisse erzeugt und auf m Untervolumen aufgeteilt werden sollen, wie finde ich die Untervolumengrenzen, die die Varianz minimieren.

Betrachtet man, z.B., eine Aufteilung des Integrationsvolumen V in zwei gleich große Untervolumen V_a und V_b, und würfelt in jedes Volumen $n/2$ Punkte, dann kann die Varianz durch die Varianzen der Untervolumen σ_a^2 und σ_b^2 ausgedrückt werden

$$\begin{aligned}\frac{\sigma_{(2)}^2}{n} &= \frac{1}{4}\left(\frac{\sigma_a^2}{n/2}+\frac{\sigma_b^2}{n/2}\right)\\ &= \frac{\sigma_a^2+\sigma_b^2}{2n}\end{aligned} \tag{5.23}$$

Zum Vergleich bestimmt sich die Varianz der Integration im Gesamtvolumen durch (die x_i für $i=[1,n/2]$ sollen im Untervolumen V_a und die x_i für $i=[n/2+1,n]$ im Untervolumen V_b liegen)

$$\begin{aligned}{\sigma_{(1)}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-\bar{f})^2\\ &= \frac{1}{n}\left\{\sum_{i=1}^{n/2}\big((f(x_i)-\bar{f}_a)+(\bar{f}_a-\bar{f})\big)^2+\dots\right\}\\ &= \frac{1}{n}\left\{\sum_{i=1}^{n/2}\Big((f(x_i)-\bar{f}_a)^2+2\overbrace{(f(x_i)-\bar{f}_a)}^{0}(\bar{f}_a-\bar{f})+(\bar{f}_a-\bar{f})^2\Big)+\dots\right\}\\ &= \left\{\underbrace{\frac{1}{n}\sum_{i=1}^{n/2}(f(x_i)-\bar{f}_a)^2}_{\sigma_a^2/2}+\frac{1}{2}\big(\bar{f}_a-\bar{f}\big)^2+\dots\right\}\\ &= \frac{\sigma_a^2}{2}+\frac{\sigma_b^2}{2}+\frac{1}{2}\big(\bar{f}_a-\bar{f}\big)^2+\frac{1}{2}\big(\bar{f}_b-\bar{f}\big)^2\\ &= \frac{\sigma_a^2}{2}+\frac{\sigma_b^2}{2}+\frac{1}{2}\left[\bar{f}_a-\frac{1}{2}(\bar{f}_a+f_b)\right]^2+\frac{1}{2}\left[\bar{f}_b-\frac{1}{2}(\bar{f}_a+f_b)\right]^2\\ &= \frac{\sigma_a^2}{2}+\frac{\sigma_b^2}{2}+\frac{1}{4}\big[\bar{f}_a-f_b\big]^2\end{aligned} \tag{5.24}$$

Die Aufteilung in Volumina, die die Varianz minimieren, erfolgt im Allgemeinen iterativ mit einer Regel, die große Varianzbeiträge minimiert. Ein Beispiel sind folgende Regeln:

- unterteile das Integrationsvolumen in 2 gleiche Untervolumen,
- erzeuge Punkte in beiden Volumina,
- berechne aus den Funktionswerten des Integranden an diesen Punkten die Varianzen für jedes Untervolumen getrennt,
- unterteile das Untervolumen mit der größten Varianz weiter,
- wiederhole die Unterteilung des Untervolumens mit der jeweils größten Varianz,
- breche die Prozedur ab, wenn ein vorgegebenes Abbruchkriterium erfüllt ist (zum Beispiel Erreichen einer Höchstzahl an Untervolumen oder Unterschreiten einer Grenze für die maximale Varianz).

In den Untervolumina kann während des Aufteilungsprozesses weiter gewürfelt werden, um mit Verfeinerung des Rasters die Punktdichte zu erhöhen. Es kann gezeigt werden, dass die optimale Anzahl der Punkte N_k im Untervolumen gegeben ist durch die Bedingung:

$$\frac{N_k}{\sigma_k} = \text{const} \tag{5.25}$$

Ziel des Algorithmus ist es eine Unterteilung in Volumen mit gleicher Varianz zu finden. Dann kann in alle Volumen die gleiche Anzahl an Punkten gewürfelt werden.

Das Integral ergibt sich dann wie in (5.17), mit zusätzlicher Summation über m Untervolumen:

$$I \approx \sum_{k=1}^{m} \frac{V_k}{N_k} \sum_{j=1}^{N_k} f(\vec{x}_j) = \sum_{k=1}^{m} V_k \, \overline{f}_k \tag{5.26}$$

Der geschätzte Fehler des Integrals ist dann:

$$\sigma_I = \sqrt{\sum_{k=1}^{m} V_k^2 \, \frac{\sigma_{\bar{f}_k}^2}{N_k}} \tag{5.27}$$

Die Formel zeigt, dass der absolute Fehler mit der Verkleinerung der Varianzen sinkt (der relative Fehler fällt bei gegebener Volumenaufteilung weiterhin wie $1/\sqrt{N}$).

Eine typische Zerlegung eines Integrationsbereiches durch einen auf "stratified sampling" basierenden Monte-Carlo-Integrationsalgorithmus (Divonne) ist in Abb. 5.4 dargestellt.

Beispiel: Den Effekt des 'stratified samplings' kann man gut an folgendem Beispiel klarmachen: Man betrachte eine lineare Funktion $f(x) = x$ in dem Interval $[0, 1]$ und berechne das Integral durch Würfeln von gleichverteilten x-Werten und Summation der Funktionswerte. Das Würfeln soll einmal im gesamten Interval $[0, 1]$ und dann getrennt in den Intervallen $V_a = [0, 0.5]$ und $V_b = [0.5, 1]$ ausgeführt werden.

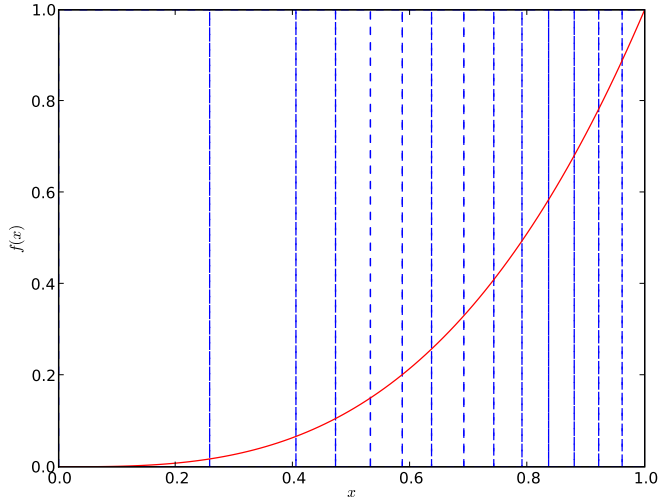


Abbildung 5.4: Zerlegung des Integrationsbereiches durch den Monte-Carlo-Integrationsalgorithmus Divonne (stratified sampling) für die Funktion $f(x) = x^3$.

Die Varianz der Funktion im Intervall $[x_1, x_2]$ ist $\sigma_f^2 = (x_2 - x_1)^2/12$. Die Varianz des Integrals ohne Unterteilung in Untervolumen ist dann

$$\sigma_{I,(1)}^2 = \frac{1}{12\,N} \tag{5.28}$$

Durch Unterteilung in zwei Volumen erhalten wir $\sigma_a^2 = \sigma_b^2 = 1/48$. Die Varianz des Integrals berechnet sich dann aus Gl. (5.27) zu

$$\sigma_{I,(2)}^2 = \left(\frac{1}{2}\right)^2 \frac{\sigma_a^2}{N/2} + \left(\frac{1}{2}\right)^2 \frac{\sigma_b^2}{N/2} = \frac{1}{48\,N} \tag{5.29}$$

Kapitel 6

Die Maximum-Likelihood-Methode

In diesem und dem nächsten Kapitel werden wir Methoden untersuchen, mit denen für Daten von Stichproben eine möglichst optimale theoretische Beschreibung beziehungsweise ein passendes Modell gefunden werden kann. Es kann sich dabei um diskrete Modell-Hypothesen oder um Funktionen der Messwerte handeln. Funktionen werden im allgemeinen durch geeignete Wahl von Parametern an die Messungen angepasst. Die Prozedur der Anpassung optimaler Parameter oder der Wahl einer Hypothese sollte gleichzeitig ein quantitatives Kriterium für die Güte der Beschreibung der Daten im Vergleich zu anderen möglichen Hypothesen bieten.

Die 'Maximum-Likelihood-Methode' (ML-Methode) ist in verschiedener Hinsicht die allgemeinste Methode zur Parameterschätzung mit vielen optimalen Eigenschaften. Eine speziellere Methode ist die sogenannte 'Methode der kleinsten Quadrate', die auf dem χ^2-Test für normal-verteilte Messwerte beruht (siehe nächstes Kapitel). Die 'Methode der kleinsten Quadrate' entspricht der 'Maximum-Likelihood-Methode' für den Spezialfall, dass die Stichproben aus Normalverteilungen stammen. Deshalb diskutieren wir im folgenden zunächst das ML-Prinzip.

6.1 Das Maximum-Likelihood-Prinzip

Es sei wieder eine Stichprobe $x_1, \ldots, x_n$ vom Umfang n gegeben, wobei jedes x_i im allgemeinen für einen ganzen Satz von Variablen stehen kann.

Wir wollen jetzt die Wahrscheinlichkeit für das Auftreten dieser Stichprobe berechnen unter der Annahme, dass die x_i einer Wahrscheinlichkeitsdichte $f(x|\theta)$ folgen, die durch einen Satz von Parametern $\theta = \theta_1, \ldots, \theta_m$ bestimmt ist. Wenn die Messungen zufällig sind (siehe die Gleichungen (4.4, 4.5) in Abschnitt 4.1), ist diese Wahrscheinlichkeit das Produkt der Wahrscheinlichkeiten für das Auftreten jedes einzelnen Elementes der Stichprobe:

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{6.1}$$

Die so definierte Stichprobenfunktion heisst **Likelihood-Funktion** und ist als Wahrscheinlichkeitsdichte für Stichproben $x_1, \ldots, x_n$ auf deren Definitionsbereich Ω normiert:

$$\int_\Omega L(x_1, \ldots, x_n|\theta)\, dx_1 \ldots dx_n = 1 \tag{6.2}$$

Das gilt für alle θ, solange $f(x_i|\theta)$ richtig normiert ist. Es ist wichtig zu realisieren, dass L nicht auf den θ-Bereich normiert ist. Andererseits betrachtet man L bei der Suche nach optimalen Parametern als eine Funktion der Parameter, die im Optimierungsprozess variiert werden.

Das ML-Prinzip lässt sich nun wie folgt formulieren:

> Wähle aus allen möglichen Parametersätzen θ denjenigen Satz $\hat{\theta}$ als Schätzung, für den gilt:

$$L(x_1, \ldots, x_n|\hat{\theta}) \geq L(x_1, \ldots, x_n|\theta) \qquad \forall\, \theta \tag{6.3}$$

Das Prinzip läuft also auf die Aufgabe hinaus, das Maximum von L in bezug auf die Parameter zu finden. Die Parameter können diskret oder kontinuierlich sein. Im diskreten Fall muss das die maximale Likelihood-Funktion bezüglich diskreter Hypothesen gefunden werden. Wenn die Parameter kontinuierlich sind kann man gängige numerische Methoden zum Auffinden des Maximums als Funktion der Parameter benutzten. Da L als Produkt von Wahrscheinlichkeiten sehr kleine Zahlenwerte haben kann, benutzt man aus numerischen Gründen meistens den Logarithmus der Likelihood-Funktion, die sogenannte Log-Likelihood-Funktion:

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \ln L(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta) \tag{6.4}$$

Die Maximierungsbedingungen (bei kontinuierlichen Parametern) lauten dann für die Log-Likelihood-Funktion, zunächst für nur einen Parameter θ:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(x_i|\theta) = 0 \qquad \Longrightarrow\ \hat{\theta} \tag{6.5}$$

$$\left.\frac{\partial^2 \mathcal{L}}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0 \tag{6.6}$$

Die Verallgemeinerung auf mehrere Parameter $\theta = \theta_1, \ldots, \theta_m$ lautet:

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \ln f(x_i|\theta) = 0 \qquad \Longrightarrow\ \hat{\theta} \tag{6.7}$$

$$\left.\frac{\partial^2 \mathcal{L}}{\partial \theta_i \theta_j}\right|_{\theta=\hat{\theta}} = U_{ij}(\hat{\theta}) \text{ negativ definit} \tag{6.8}$$

Die Matrix U ist negativ definit, wenn alle Eigenwerte kleiner 0 sind. Falls Gleichung (6.7) auf ein lineares Gleichungssystem führt, kann man die Lösung durch Matrixinversion erhalten. Im allgemeinen sind die Gleichungen nicht-linear und man muss eine numerische, meistens iterative Methode zur Lösung finden. Wir werden Lösungsverfahren im Zusammenhang mit der 'Methode der kleinsten Quadrate' im nächsten Kapitel besprechen.

Beispiele:

1. Schätzung der mittleren Lebensdauer: Die Abfolge der Zerfälle eines radioaktiven Präparates habe die Wahrscheinlichkeitsdichte

$$f(t|\tau) = \frac{1}{\tau}\, e^{-t/\tau}, \tag{6.9}$$

die als einzigen Parameter die mittlere Lebensdauer τ enthält. In einer Messung werden n Zerfälle mit den Zeiten $t_i,\ i = 1, \ldots, n$ gemessen. Die Likelihood-Funktion dieser Stichprobe ist:

$$L(t_1, \ldots, t_n|\tau) = \prod_{i=1}^{n} \frac{1}{\tau}\, e^{-t_i/\tau} \qquad \Longrightarrow \qquad \mathcal{L}(t_1, \ldots, t_n|\tau) = \sum_{i=1}^{n} \left(-\ln \tau - \frac{t_i}{\tau} \right) \tag{6.10}$$

Die Maximierung von $\mathcal{L}$ ergibt den ML-Schätzwert für τ:

$$\frac{\partial \mathcal{L}}{\partial \tau} = \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \qquad \Longrightarrow \qquad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i = \bar{t} \tag{6.11}$$

mit

$$\left. \frac{\partial^2 \mathcal{L}}{\partial \tau^2} \right|_{\tau = \hat{\tau}} = -\frac{n}{\hat{\tau}^2} < o \tag{6.12}$$

Die ML-Schätzung der mittleren Lebensdauer ist also das arithmetische Mittel der gemessenen Zeiten.

2. Schätzung der Parameter einer Gauss-Verteilung: Eine Stichprobe $x_i,\ i = 1, \ldots, n$ aus einer Normalverteilung $N(\mu,\, \sigma)$ hat die Likelihood-Funktion:

$$L(x_1, \ldots, x_n|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x_i-\mu)^2}{2\,\sigma^2}} \tag{6.13}$$

$$\Longrightarrow \mathcal{L}(x_1, \ldots, x_n|\mu, \sigma^2) = \frac{1}{2} \sum_{i=1}^{n} \left(-\ln \sigma^2 - \ln 2\pi - \frac{(x_i-\mu)^2}{2\,\sigma^2} \right) \tag{6.14}$$

Die Maximierung in Bezug auf beide Parameter fordert:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \tag{6.15}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \sum_{i=1}^{n} \left(-\frac{1}{2\,\sigma^2} + \frac{1}{2\,\sigma^4} (x_i - \mu)^2 \right) = 0 \tag{6.16}$$

Die Lösung des Gleichungssystems ergibt:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \tag{6.17}$$

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{6.18}$$

Die ML-Schätzung des Mittelwertes ist also wieder das arithmetische Mittel. Die Schätzung der Varianz ist allerdings verzerrt, denn der Erwartungswert ist nicht unabhängig von n (siehe dazu Abschnitt 4.2):

$$E(\hat{\sigma^2}) = \left(1 - \frac{1}{n}\right) \sigma^2 \tag{6.19}$$

Die Schätzung ist aber 'konsistent', weil der Erwartungswert der Schätzung für große n gegen den zu schätzenden Parameter konvergiert.

6.2 ML-Methode für Histogramme

In den Beispielen im vorigen Abschnitt wurde die Likelihood-Funktion als Produkt der Wahrscheinlichkeiten der einzelnen Ereignisse konstruiert ('unbinned likelihood'). Häufig werden Messdaten auch als Histogramme dargestellt, das heißt, die Häufigkeit von Ereignissen als Funktion einer Variablen wird für endliche Intervalle ('bins') dieser Variablen aufgetragen.

Beispiel: In Abb. 6.1 sind die Raten von beobachteten Myonpaaren, die in Proton-Kern-Reaktionen von einem separierten Vertex kommen, gegen deren invariante Masse pro Masseninterval aufgetragen. Die einzelnen Zählraten sind hier als Punkte mit Fehlerbalken eingezeichnet (könnten aber auch als Histogrammbalken dargestellt werden), ein getrennt gemessener Untergrund wird zusätzlich als Histogramm eingezeichnet. Man beobachtet bei etwa 3.1 GeV das Signal für den Zerfall $J/\psi \to \mu^+\mu^-$ mit einer etwa gauss-förmigen Massenverteilung auf einem näherungsweise konstanten Untergrund. Eine Funktion bestehend aus der Summe einer Normalverteilung und einem konstanten Untergrund wurde mit der ML-Methode an die Verteilung angepasst. Die Funktion hat bis zu 4 Parameter: Höhe, Breite und Lage der Normalverteilung und eine Konstante für den Untergrund. Statt der Höhe der Normalverteilung definiert mal vorteilhafter das Integral unter der Signalkurve, weil das direkt die gesuchte Anzahl der J/ψ-Mesonen ergibt und sich damit eine Umrechnung mit eventuell korrelierten Parameterfehlern vermeiden läßt.

Für die Bestimmung der Likelihood-Funktion, die wir für die Anpassung brauchen, nehmen wir an, dass die Raten N_i in jedem Interval i poisson-verteilt sind. Wir vergleichen diese Raten mit der Hypothese $\lambda_i(\theta)$, die wir als Mittelwert der Anpassungsfunktion $f(x|\theta)$ um die Intervalmitte x_i bestimmen:

$$\lambda_i(\theta) = \langle f(x|\theta) \rangle_{[x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}]} \tag{6.20}$$

Die Likelihood-Funktion wird dann aus den Poisson-Wahrscheinlichkeiten für die Beobachtung von N_i Ereignissen bei gegebenem Erwartungswert λ_i in jedem Interval i konstruiert:

$$L(\theta) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda_i} \lambda_i^{N_i}}{N_i!} \quad \Rightarrow \ln L(\theta) = \sum_{i=1}^{n} \left(-\lambda_i + N_i \ln \lambda_i - \ln(N_i!)\right) \tag{6.21}$$

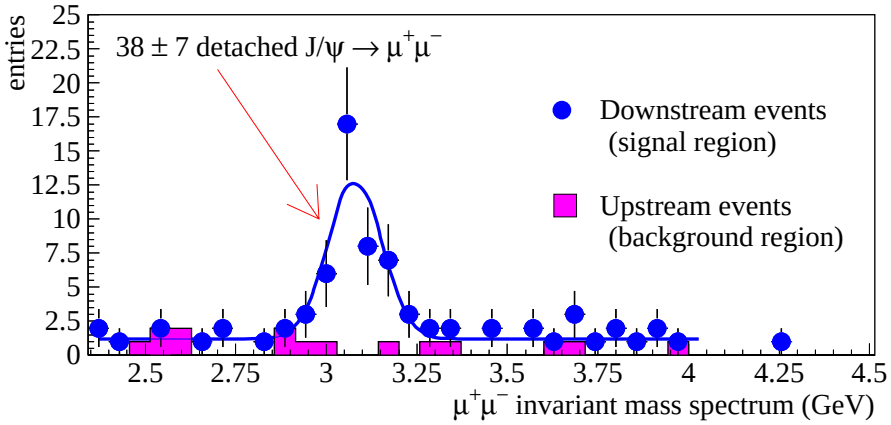


Abbildung 6.1: Massenverteilung von Myonpaaren in Proton-Kern-Reaktionen (HERA-B-Experiment), die einen gemeinsamen Vertex mit Abstand ('detached') zum Primärvertex haben. Die Myonpaare wurden als Kandidaten für Zerfälle von J/ψ-Mesonen, die wiederum aus Zerfällen von langlebigen B-Mesonen stammen, selektiert. Die Verteilung wird durch eine Normalverteilung für das J/ψ-Signal über einem konstanten Untergrund beschrieben.

Der letzte Term ist durch die Messung gegeben und hängt nicht von den zu optimierenden Parametern θ ab. Die zu maximierende Log-Likelihood-Funktion reduziert sich deshalb auf:

$$\ln L(\theta) = \sum_{i=1}^{n} (-\lambda_i + N_i \ln \lambda_i) \tag{6.22}$$

Wenn jedes einzelne Ereignis tatsächlich gemessen wurde und nicht durch den Messprozess bereits der Eintrag in Histogramme erfolgt, kann man alternativ zu dieser 'binned likelihood' Methode natürlich auch die Likelihood-Funktion mit den Wahrscheinlichkeiten der einzelnen Ereignisse konstruieren ('unbinned likelihood'). Die 'unbinned likelihood' kann im Allgemeinen mehr Information ausnutzen.

Bemerkung: Häufig wird die Poisson-Verteilung für die Raten durch eine Normalverteilung approximiert, um dann als Log-Likelihood-Funktion die χ^2-Funktion anpassen zu können (siehe nächstes Kapitel). Bei kleinen Zählraten, insbesondere mit Null-Einträgen in Intervallen, führt das in der Regel zu verfälschten Ergebnissen. Aber auch bei Zählraten, für die die Gauss-Approximation gut ist, gibt es ein Problem: Das Integral unter der Anpassungskurve wird regelmässig unterschätzt, wenn die Fehler durch $1/\sqrt{N_i}$ abgeschätzt werden. Damit werden Fluktuationen nach unten durch einen kleineren Fehler stärker bewichtet als Fluktuationen nach oben. Im Mittel zieht das dann die Anpassungskurve nach unten. Wenn man unbedingt eine χ^2-Anpassung machen will, kann man als Abhilfe den Fehler iterativ mit dem aktuellen Anpassungswert λ_i als $1/\sqrt{\lambda_i}$ festlegen.

6.3 Berücksichtigung von Zwangsbedingungen

Oft sind bei einer Anpassung einer Funktion an Messdaten Zwangsbedingungen zu berücksichtigen. Zwangsbedingungen kommen häufig bei kinematischen Anpassun-

gen vor: zum Beispiel ist in einer e^+e^--Annihilation im Schwerpunktsystem die Summe der Impulse gleich null und die Summe der Energien gleich zweimal die Strahlenergie. Daraus resultieren 4 Zwangsbedingungen, die durch weitere Bedingungen, wie Massen- oder Vertexbedingungen an Untersysteme von Teilchen, ergänzt werden können. Jede Zwangsbedingung kann zur Eliminierung eines Parameters benutzt werden, zum Beispiel kann man mit der gerade erwähnten Impulserhaltung 3 Impulskomponenten eliminieren. Häufig ist das aber nicht erwünscht, zum Beispiel um bei der Anpassung die äquvalente Behandlung der Parameter zu gewährleisten oder um schwierigen Eliminierungs-Algorithmen aus dem Weg zu gehen.

6.3.1 Methode der Lagrange-Multiplikatoren

Die k_c Zwangbedingungen ('constraints') eines Anpassungsproblems werden als Funktionen $c_j(\theta)$ $(j = 1, \ldots, k_c)$ definiert, die verschwinden, wenn die jeweilige Bedingung erfüllt ist. Wie in der klassischen Mechanik lassen sich die Bedingungen mit der Methode der Lagrange-Multiplikatoren in die Likelihood-Funktion einbeziehen:

$$\mathcal{L} = \ln L = \sum_{i=1}^{m} \ln f(x_i|\theta) \; - \; \sum_{j=1}^{k_c} \lambda_j \, c_j(\theta). \tag{6.23}$$

Die k_c Lagrange-Multiplikatoren λ_j werden wie zusätzliche Parameter behandelt, bezüglich der die Likelihood-Funktion ebenfalls zu minimieren ist. Zu den m Maximierungsbedingungen in (6.7)

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 0 \tag{6.24}$$

kommen noch die k_c Bedingungen

$$\frac{\partial \mathcal{L}}{\partial \lambda_j} = c_j(\theta) = 0. \tag{6.25}$$

Das Verschwinden der Funktionen $c_j(\theta)$ ergibt sich also aus der Maximierungsbedingung bezüglich der Lagrange-Multiplikatoren.

6.3.2 Zwangsbedingungen als Zufallsverteilungen

Insbesondere wenn Zwangsbedingungen nicht scharf definiert sind oder nur mit begrenzter Genauigkeit bekannt sind, kann man die Abweichungen als Zufallsverteilung behandeln. Mit einer angenommenen Normalverteilung mit der Breite δ_j für die Verteilung von c_j um Null ergibt sich in der Log-Likelihood-Funktion ein χ^2-artiger Zusatz:

$$\mathcal{L} = \ln L = \sum_{i=1}^{m} \ln f(x_i|\theta) \; - \; \frac{1}{2} \sum_{j=1}^{k_c} \frac{c_j^2(\theta)}{\delta_j^2}. \tag{6.26}$$

Diese Art der Implementierung der Zwangbedingungen kann auch im Falle scharf definierter Zwangsbedingungen vorteilhaft sein, weil die Anzahl der Parameter kleiner wird. In diesem Fall würde man die δ_j genügend klein machen (eventuell auch adaptiv während des Maximierungsprozesses).

6.3.3 Erweiterte ML-Methode

Es gibt Probleme, bei denen sich aus einer ML-Anpassung gleichzeitig die Anzahl der zu erwartenden Ereignisse ergibt und diese Anzahl mit der Anzahl der tatsächlich beobachteten Ereignisse in Übereinstimmung gebracht werden soll. Will man zum Beispiel von n Ereignissen bestimmen, welcher Bruchteil jeweils aus einer von drei angenommenen Reaktionen stammt, sollte gleichzeitig die Summe der jeweiligen Anzahlen gleich n sein: $n = n_1 + n_2 + n_3$. Man kann nun diese Bedingung als einen zusätzlichen Faktor in die Likelihood-Funktion einsetzen, und zwar entsprechend der Poisson-Verteilung als Wahrscheinlichkeit, dass bei einem Erwartungswert λ tatsächlich n Ereignisse beobachtet werden. Die Likelihood-Funktion (6.1) mit normierten Wahrscheinlichkeiten $f(x|\theta)$ wird dann erweitert zu:

$$L(x_1, \ldots, x_n|\theta) = \frac{\lambda^n \, \mathrm{e}^{-\lambda}}{n!} \prod_{i=1}^{n} f(x_i|\theta) \tag{6.27}$$

Daraus folgt für die Log-Likelihood-Funktion:

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = n \ln \lambda - \lambda + \sum_{i=1}^{n} \ln f(x_i|\theta), \tag{6.28}$$

wobei der für die Maximierung irrelevante Term $(-\ln n!)$ weggelassen wurde.

Mit der Umrechnung

$$n \ln \lambda + \sum_{i=1}^{n} \ln f(x_i|\theta) = \sum_{i=1}^{n} (\ln f(x_i|\theta) + \ln \lambda) = \sum_{i=1}^{n} \ln (\lambda f(x_i|\theta)) \tag{6.29}$$

kann eine Funktion $g(x|\theta) = \lambda f(x|\theta)$ definiert werden, deren Normierung λ ist:

$$\int_\Omega g(x|\theta) dx = \lambda \int_\Omega f(x|\theta) dx = \lambda \tag{6.30}$$

Damit wird aus (6.28) die gängige Form der erweiterten Likelihood-Funktion (EML):

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{n} \ln g(x_i|\theta) - \int_\Omega g(x|\theta) dx \tag{6.31}$$

Dass $\mathcal{L}$ tatsächlich maximal wird, wenn der zusätzliche Term in (6.31) n ergibt, kann man sich folgendermaßen klar machen: Wir skalieren die Funktion $g(x|\theta)$ mit einem Faktor β und fragen uns, für welchen Wert von β die Likelihood-Funktion maximal wird:

$$\mathcal{L} = \sum_{i=1}^{n} \ln (\beta \, g(x_i|\theta)) - \int_\Omega \beta \, g(x|\theta) dx \tag{6.32}$$

Die Maximierungsbedingung bezüglich β lautet:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \frac{n}{\beta} - \int_\Omega g(x|\theta) dx = 0 \qquad \Longrightarrow \qquad \beta = \frac{n}{\int_\Omega g(x|\theta) dx} \tag{6.33}$$

Man sieht also, dass für das tatsächlich gewählte $\beta = 1$ die Likelihood-Funktion für

$$n = \int_\Omega g(x|\theta)dx \tag{6.34}$$

maximal wird. Man kann sich vergewissern, dass diese Normierungsbedingung sogar exakt erfüllt wird, obwohl wir bei der Herleitung der EML von einer Poisson-Verteilung ausgegangen waren. Zusätzlich lernt man von diesem Beweis, dass man β auch anders wählen und damit andere Normierungsbedingungen erhalten kann. Naheliegend wäre zum Beispiel $\beta = 1/n$, womit sich nach (6.33) $\int_\Omega g(x|\theta)dx = 1$ ergibt[1].

Beispiel: Wir greifen das oben angeführte Beispiel auf: n gemessene Ereignisse sollen m verschiedenen Reaktionen zugeordnet werden, für jede Reaktion j gibt es die normierte Wahrscheinlichkeit $f_j(x)$, dass das Ereignis aus dieser Reaktion stammt. Die Funktion g wird dann definiert:

$$g(x|n_1,\ldots,n_m) = \sum_{j=1}^{m} n_j f_j(x) \quad \Longrightarrow \quad \int_\Omega g(x|n_1,\ldots,n_m)dx = \sum_{j=1}^{m} n_j \tag{6.35}$$

Mit der erweiterten Likelihood-Funktion

$$\mathcal{L}(x_1,\ldots,x_n|n_1,\ldots,n_m) = \sum_{i=1}^{n} \ln g(x_i|n_1,\ldots,n_m) - \sum_{j=1}^{m} n_j \tag{6.36}$$

wird die Bedingung $n = \sum_{j=1}^{m} n_j$ erfüllt.

Diesen Ansatz kann man für das Beispiel der Abb. 6.1 anwenden, wenn man aus den einzelnen Ereignissen eine Likelihood-Funktion ('unbinned likelihood') konstruieren will (statt aus den Histogrammeinträgen, wie vorher behandelt): die Anpassung soll dann n_S Signalereignisse und n_B Untergrundereignisse mit der Bedingung für die Gesamtzahl $n = n_S + n_B$ ergeben.

6.3.4 Freiheitsgrade und Zwangsbedingungen

Eine Anpassung einer Hypothese an eine Stichprobe kann nur gemacht werden, wenn die Anzahl der Parameter m höchstens gleich der Anzahl der Messwerte n ist. Die Anzahl der Freiheitsgrade ergeben sich dann zu:

$$n_F = n - m. \tag{6.37}$$

Jede unabhängige Zwangsbedingung trägt wie ein zusätzlicher Messwert bei, so dass sich für k_c Bedingungen ergibt:

$$n_F = n - m + k_c. \tag{6.38}$$

Ein positiver Wert von n_F erlaubt eine Verbesserung der Messung durch Ausgleich zwischen den Messwerten. Bei kinematischen Anpassungen spricht man von n_FC-Fit

[1] Diese Normierung ist zum Beispiel bei der in [Z. Phys. C16 (1982) 13] dargestellten Analyse benutzt worden.

('constrained fit'). Ein 4C-Fit ('four-C fit') ergibt sich zum Beispiel, wenn man $3n$ Impulskomponenten eines n-Teilchensystems gemessen hat, das System 4 Zwangsbedingungen durch die Viererimpuls-Erhaltung unterliegt und die $3n$ Impulskomponenten als Parameter des Systems angepasst werden. Es wäre nur ein 1C-Fit, wenn ein Teilchen nicht beobachtet würde (das man aber dann wegen der Zwangsbedingungen rekonstruieren kann).

6.4 Fehlerbestimmung für ML-Schätzungen

Die Fehler oder Unsicherheiten in der Parameterbestimmung mit der ML-Methode lassen sich nur in speziellen Fällen explizit angeben, zum Beispiel wenn die Likelihood-Funktion normalverteilt in den Parametern ist (siehe unten). Andererseits ist eine Parameterbestimmung ohne Aussagekraft, wenn man nicht einen Fehler oder ein Vertrauensniveau angeben kann. Im allgemeinen wird die vollständige Kovarianzmatrix benötigt, wenn man ML-Ergebnisse für die weitere Auswertung braucht.

6.4.1 Allgemeine Methoden der Varianzabschätzung

Direkte Methode: Die direkte Methode gibt die Streuung der Schätzwerte $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ an, wenn man viele Messungen mit Stichproben $(x_1, \ldots, x_n)$ macht:

$$V_{ij}(\theta) = \int (\hat{\theta}_i - \theta_i)\,(\hat{\theta}_j - \theta_j)\, L(x_1, \ldots, x_n|\theta)\, dx_1 \ldots dx_n \tag{6.39}$$

Hier ist also $\theta = (\theta_1, \ldots, \theta_m)$ der 'wahre' Parametersatz und $\hat{\theta}(x_1, \ldots, x_n)$ sind die Schätzungen, die man jeweils für eine Stichprobe erhält. Die Stichproben, über die integriert wird, folgen der Wahrscheinlichkeitsdichte $L(x_1, \ldots, x_n)$.

Bei dieser Varianzbestimmung wird die Kenntnis des wahren Parametersatzes θ und der Verlauf von L als Funktion der x_i vorausgesetzt. Bei einer Messung weiss man in der Regel weder das eine noch das andere. Man kann diese Methode aber zum Beispiel zur Planung von Experimenten benutzen, um die zu erwartenden Fehler beim Testen eines Modells mit bestimmten Parametern auszuloten. Die Auswertung wird dann in der Regel mit Simulationen der Stichproben gemacht. Auch für experimentelle Messungen kann man diese Bestimmung der Varianzen benutzen. Für den geschätzten Parametersatz $\hat{\theta}$ simuliert man den Verlauf der Likelihood-Funktion durch die Simulation vieler Messungen, die man in der Praxis nicht durchführen könnte.

Praktische Methode: In der Praxis wird meistens $L(x_1, \ldots, x_n|\theta)$ bei fester Stichprobe $(x_1, \ldots, x_n)$ als Wahrscheinlichkeitsdichte für θ angenommen. Dann erhält man für die Varianzmatrix:

$$V_{ij}(\theta) = \frac{\int (\theta_i - \hat{\theta}_i)\,(\theta_j - \hat{\theta}_j)\, L(x_1, \ldots, x_n|\theta)\, d\theta_1 \ldots d\theta_m}{\int L(x_1, \ldots, x_n|\theta)\, d\theta_1 \ldots d\theta_m} \tag{6.40}$$

Hier ist $\hat{\theta}$ die ML-Schätzung, die aus der einen gemessenen Stichprobe $(x_1, \ldots, x_n)$ bestimmt wurde. In der Formel (6.40) ist berücksichtigt, dass L nicht auf den θ-Bereich normiert ist, wie bereits oben erwähnt wurde.

In der Regel werden die Integrationen numerisch durch Abtasten der Likelihood-Funktion für verschiedene Parameter θ durchgeführt.

6.4.2 Varianzabschätzung durch Entwicklung um das Maximum

Wenn die Likelihood-Funktion gewisse günstige Eigenschaften hat, insbesondere wenn der Verlauf um den optimalen Parametersatz als Funktion der Parameter ein ausgeprägtes Maximum hat und nach beiden Seiten monoton abfällt, kann man eine Entwicklung um das Maximum versuchen. Aus Gründen, die wir gleich verstehen werden, entwickeln wir die Log-Likelihood-Funktion:

$$\mathcal{L}(x_1,\ldots,x_n|\theta) = \mathcal{L}(x_1,\ldots,x_n|\hat{\theta}) + (\theta-\hat{\theta})\left.\frac{\partial\mathcal{L}}{\partial\theta}\right|_{\theta=\hat{\theta}} + \frac{1}{2}(\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)\left.\frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j}\right|_{\theta=\hat{\theta}} + \ldots \tag{6.41}$$

Wegen der Maximumbedingung verschwindet die erste Ableitung. Die zweiten Ableitungen werden zusammengefasst:

$$V_{ij}^{-1} = -\left.\frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j}\right|_{\theta=\hat{\theta}} \tag{6.42}$$

Damit ergibt sich in der Umgebung des Maximums:

$$\mathcal{L}((x_1,\ldots,x_n|\theta) \approx \mathcal{L}_{max} - \frac{1}{2}(\theta-\hat{\theta})^T V^{-1}(\theta-\hat{\theta}) \tag{6.43}$$

und für die Likelihood-Funktion L folgt:

$$L((x_1,\ldots,x_n|\theta) \approx L_{max}\, e^{-\frac{1}{2}(\theta-\hat{\theta})^T V^{-1}(\theta-\hat{\theta})} \tag{6.44}$$

Das heisst, wenn die Likelihood-Funktion als Funktion der Parameter ein annähernd gaussisches Verhalten zeigt, kann die Varianz durch die zweiten Ableitungen entsprechend (6.42) abgeschätzt werden. In der Praxis wird häufig angenommen, dass die Likelihood-Funktion einer (Multi)-Normalverteilung folgt.

Wenn die Parameter unkorreliert sind, ist V^{-1} diagonal und die Varianz der Parameter ist:

$$\sigma_i^2 = \frac{1}{V_{ii}^{-1}} = \left(-\left.\frac{\partial^2\mathcal{L}}{\partial{\theta_i}^2}\right|_{\theta=\hat{\theta}}\right)^{-1} \tag{6.45}$$

6.4.3 Vertrauensintervalle und Likelihood-Kontouren

Die Fehler der Parameter werden häufig als die Wurzeln aus den Varianzen, wie sie im vorigen Abschnitt bestimmt wurden, angegeben. Wenn man genauer sein will, kann man Likelihood-Kontouren angeben. Das sind im allgemeinen Fall Hyperflächen im Parameterraum, die durch

$$L((x_1,\ldots,x_n|\theta) = const \tag{6.46}$$

festgelegt sind und einen bestimmten Wahrscheinlichkeitsinhalt η, entprechend einem Vertrauensniveau, haben. Bei zwei Parametern (θ_i,θ_j) ergibt sich zum Beispiel

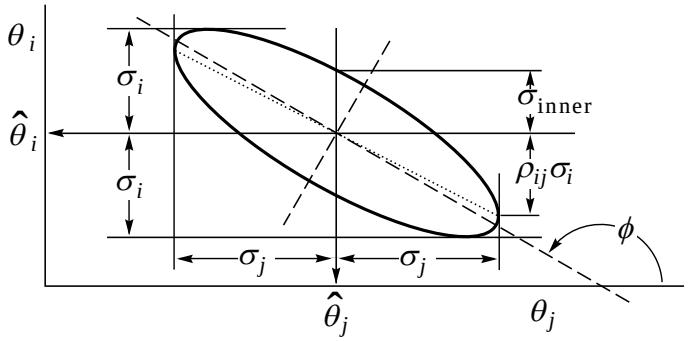


Abbildung 6.2: Standard-Fehlerellipse für die Schätzwerte $\hat{\theta}_i$ und $\hat{\theta}_j$.

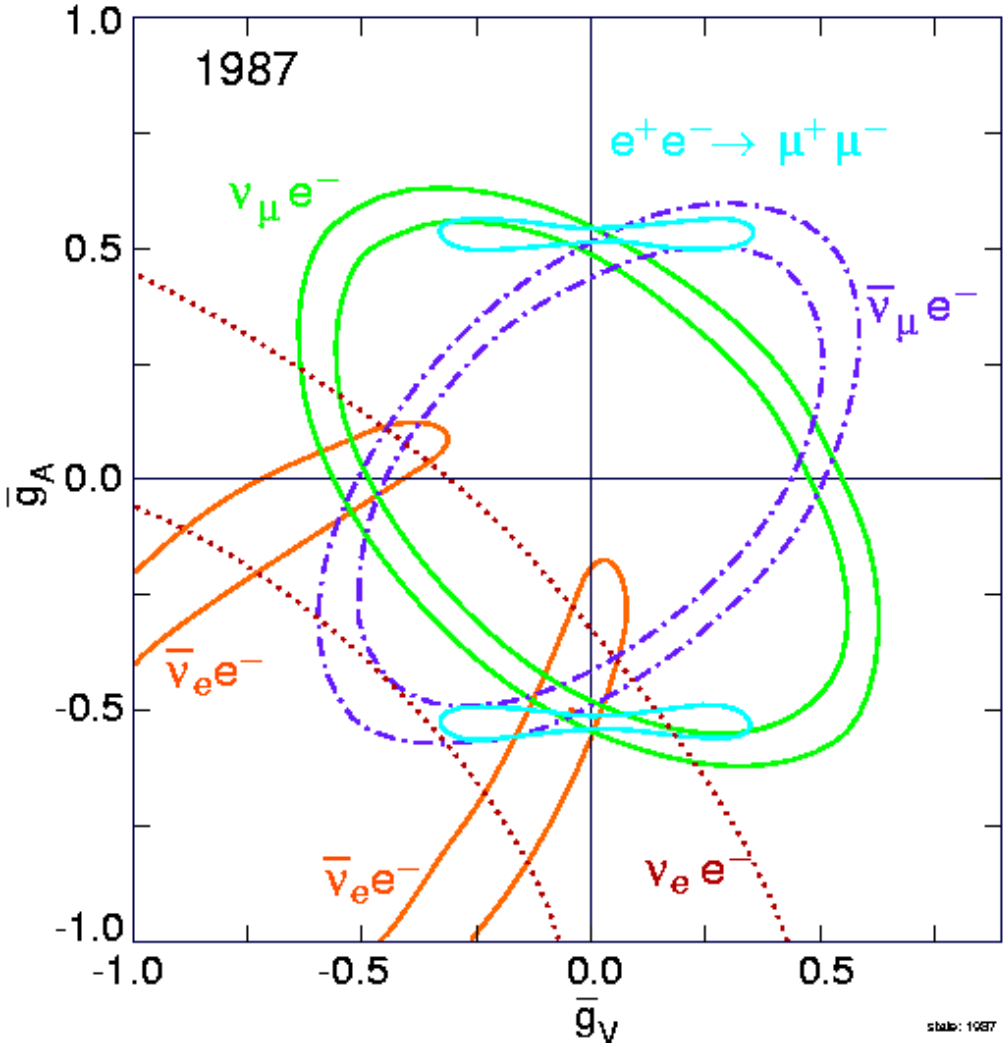


Abbildung 6.3: Beispiel für Likelihood-Kontouren: Die zwei Konstanten g_V und g_A (Kopplung von Leptonen an das Z^0-Boson) werden in verschiedenen Teilchenreaktionen gemessen, die sehr unterschiedliche Likelihood-Kontouren liefern. Die besten Schätzwerte liegen innerhalb der Kontouren, die ringförmige und zum Teil auch nicht zusammenhängende Gebiete beschreiben. Der einzige Bereich, den alle Likelihood-Kontouren umschreiben, ist nahe $g_V = 0$, $g_A = -0.5$. Für die genaue Analyse müssen alle Likelihood-Funktionen kombiniert werden.

in der Regel eine geschlossene, zwei-dimensionale Raumkurve um die Schätzwerte $(\hat{\theta}_i, \hat{\theta}_j)$ der Parameter (Abb. 6.2). Im allgemeinen können die Hyperflächen beliebige Volumina im Parameterraum einschliessen, zum Beispiel brauchen diese Volumina auch nicht zusammenzuhängen (ein Beispiel ist in Abb. 6.3 gezeigt).

Als Vertrauensniveau können Werte wie 68%, 90%, 95% usw. angegeben werden. Im allgemeinen müssen die Likelihood-Kontouren dafür numerisch integriert werden. In dem speziellen Fall, dass die Likelihood-Funktion durch eine Normalverteilung entsprechend (6.44) beschrieben werden kann, folgt

$$2\,\Delta\mathcal{L} \;=\; 2\,[\mathcal{L}_{max} \;-\; \mathcal{L}((x_1,\ldots,x_n|\theta)] \;=\; (\theta-\hat{\theta})^T\,V^{-1}\,(\theta-\hat{\theta}) \tag{6.47}$$

einer χ^2-Verteilung mit m Freiheitsgraden (m = Anzahl der Parameter). In diesem Fall ergibt $2\,\Delta\mathcal{L} = 1$ die Kovarianzen der Parameter. Die Kontouren zu einem Vertrauensniveau η ergeben sich aus den Kurven in Abb. 4.3 durch $2\,\Delta\mathcal{L} = \chi^2 = const$ für $n_F = m$ und mit $\eta = 1-\alpha$. Die Kontouren sind im Zweidimensionalen Ellipsen und im allgemeinen m-dimensionale Ellipsoide.

Zum Beispiel enthält die Kontour mit $m = 2$, $2\,\Delta\mathcal{L} = 1$ (das ist die Ellipse, die die $\pm 1\sigma$-Linien schneidet, siehe Abb. 6.2) nur 39.4% Wahrscheinlichkeit, während das für $m = 1$ bekanntlich 68.3% sind.

6.5 Eigenschaften von ML-Schätzungen

Die Likelihood-Schätzung der Parameter hat in vieler Hinsicht optimale Eigenschaften. Im Rahmen dieser Vorlesung ist allerdings nicht ausreichend Zeit, in die Details und die mathematischen Beweise zu schauen. Einige dieser Eigenschaften sollen hier nur kurz erwähnt werden:

1. Invarianz gegenüber Parametertransformationen: Im allgemeinen ist die Schätzung unabhängig davon, wie die Parameter dargestellt werden. Für eine Transformation

   $$\theta \rightarrow \phi \tag{6.48}$$

   ergibt sich:

   $$\hat{\phi} = \phi(\hat{\theta}) \tag{6.49}$$

   Zum Beispiel kann man für die Schätzung einer mittleren Lebensdauer τ auch die Zerfallswahrscheinlichkeit $\lambda = 1/\tau$ benutzen, denn aus

   $$\frac{\partial L}{\partial \lambda}(\hat{\lambda}) = 0 \tag{6.50}$$

   folgt

   $$\frac{\partial L}{\partial \tau}\frac{\partial \tau}{\partial \lambda}(\hat{\lambda}) = 0 \;\Rightarrow\; \frac{\partial L}{\partial \tau}(\tau(\hat{\lambda})) = 0 \qquad (\frac{\partial \tau}{\partial \lambda} \neq 0) \tag{6.51}$$

2. Konsistenz: Für große Stichproben geht der Schätzwert in den tatsächlichen Wert über:

   $$\lim_{n\rightarrow\infty} \hat{\theta} = \theta \tag{6.52}$$

3. Verzerrung: Wir hatten am Beispiel der Schätzung der Varianz einer Gauss-Verteilung gesehen (siehe (6.19)), dass die ML-Schätzung nicht unbedingt verzerrungsfrei ist, d. h. es gilt nicht $E(\hat{\theta}) = \theta$ für alle n. Allgemein gilt allerdings, dass die ML-Schätzung asymptotisch verzerrungsfrei ist:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta \tag{6.53}$$

4. Effizienz: In den meisten Fällen ist eine ML-Schätzung effizient, das heisst, die geschätzten Parameter haben minimale Varianz. Jedenfalls gilt das im Fall großer Stichproben: die ML-Schätzung ist asymptotisch effizient.

Schwieriger ist die Beurteilung der Fehler und Vertrauensintervalle einer Schätzung. Das Problem tritt dann auf, wenn man die Likelihood-Funktion als Wahrscheinlichkeitsdichte der Parameter interpretiert und benutzt. Zur Fehlerabschätzung braucht man eigentlich den Verlauf der gesamten Likelihood-Funktion. Wir hatten bereits darauf hingewiesen, dass die Likelihood-Funktion in Abhängigkeit von den Parametern nicht normiert ist. Um richtig normieren zu können, müsste man eigentlich den möglichen Bereich der Parameter genau kennen und auch, ob alle Parameter gleich wahrscheinlich sind oder was die 'a priori' Wahrscheinlichkeiten der Parameter sind.

Nach dem Bayes-Theorem (1.13) würde man bei einer gegebenen Stichprobe $\vec{x}$ und für diskrete Hypothesen θ_i folgende 'a posteriori' Wahrscheinlichkeit, dass die Hypothese θ_i wahr ist, erhalten:

$$P(\theta_i|\vec{x}) = \frac{P(\vec{x}|\theta_i) \cdot P(\theta_i)}{\sum_j P(\vec{x}|\theta_j) \cdot P(\theta_j)} \tag{6.54}$$

Hier entspricht $P(\vec{x}|\theta_i)$ der Likelihood-Funktion $L(\vec{x}|\theta_i)$ und $P(\theta_i)$ ist die 'a priori' Wahrscheinlichkeit der Hypothese θ_i. Der Nenner normiert auf alle möglichen Hypothesen (für kontinuierliche Hypothesen-Parameter ergibt sich ein Normierungsintegral).

Beispiel: In Teilchenexperimenten möchte man häufig die gemessenen langlebigen Teilchen identifizieren, typischerweise die 5 Teilchensorten i, $i = p,\ K,\ \pi,\ e,\ \mu$. Aus den Informationen verschiedener Detektoren, die uns hier nicht im Detail interessieren, kann man eine Masse m des Teilchens bestimmen (zum Beispiel aus der Messung von Impuls und Geschwindigkeit) und damit eine Wahrscheinlichkeit für eine Teilchenhypothese i:

$$P(i|m) = \frac{P(m|i) \cdot P(i)}{\sum_j P(m|j) \cdot P(j)} \tag{6.55}$$

Die Wahrscheinlichkeit $P(m|i)$, bei Vorliegen des Teilchens i eine Masse m zu messen, bestimmt man in der Regel experimentell mit bekannten Teilchenstrahlen. Die 'a priori' Wahrscheinlichkeit $P(i)$ für das Auftreten der Teilchensorte i entnimmt man dem gleichen Experiment, weil die Teilchenhäufigkeiten abhängig von der Energie der Reaktion (und eventuell noch anderen Parametern) sind. Die Teilchenhäufigkeiten sind im allgemeinen sehr unterschiedlich,

mit starker Dominanz der Pionen. Wenn es zum Beispiel einen Faktor 10 mehr Pionen als Kaonen gibt, muss $P(m|K) > 10 \cdot P(m|\pi)$ sein, damit es als Kaon identifiziert wird. Die Kenntnis der 'a priori' Wahrscheinlichkeit einer Teilchensorte ist also in diesem Fall besonders wichtig.

In vielen Fällen kennt man die 'a priori' Wahrscheinlichkeiten für die Hypothesen nicht und nimmt dann an, dass sie konstant sind. Dass das problematisch ist, sieht man auch daran, dass die Vertrauensintervalle nicht invariant gegen Transformationen der Parameter sind. Für die Transformation

$$\theta \to \phi(\theta) \tag{6.56}$$

ergibt sich für die Berechnung eines Vertrauensintervalls:

$$\int_{\theta_1}^{\theta_2} L(\vec{x}|\theta)\, d\theta = \int_{\phi(\theta_1)}^{\phi(\theta_2)} L(\vec{x}|\phi(\theta)) \left|\frac{\partial\theta}{\partial\phi}\right| d\phi \neq \int_{\phi_1}^{\phi_2} L(\vec{x}|\phi)\, d\phi. \tag{6.57}$$

Das rechte Integral hätte man ja erhalten, wenn man von vornherein ϕ als Parameter gewählt hätte.

Kapitel 7

Methode der kleinsten Quadrate

Im Folgenden wird die Methode der kleinsten Quadrate (LS = 'least square'), die auf dem χ^2-Test beruht, für die Anpassung von parametrisierten Funktionen an normalverteilte (oder annähernd normalverteilte) Messwerte eingeführt. Im vorigen Kapitel hatten wir bereits darauf hingewiesen, dass diese Methode der Maximum-Likelihood-Methode im Falle normalverteilter Wahrscheinlichkeiten entspricht.

7.1 Prinzip der Methode der kleinsten Quadrate

Gegeben sei eine Stichprobe mit folgenden Messwerten und der parametrisierten Beschreibung der Messwerte:

y_i: Messwerte an den (ohne Fehler) bekannten Punkten x_i (unabhängige Variable, kann auch ein Vektor sein, $i = 1, \ldots, n$);

σ_i: Fehler von y_i, Standardabweichung;

η_i: $\eta_i = f(x_i|\theta)$ ist der Erwartungswert von y_i, wenn die Abhängigkeit von x_i durch $f(x|\theta)$ beschrieben wird;

θ_j: Parameter der Funktion f, die so optimiert werden sollen, dass $f(x_i|\theta) = \eta_i$ die Messwerte y_i möglichst gut beschreibt ($j = 1, \ldots, m$).

Das LS-Prinzip lautet: Bestimme die Schätzwerte $\hat{\theta}$ der Parameter $\theta = (\theta_1, \ldots, \theta_m)$ durch Minimierung der Summe der Quadrate der auf die Fehler normierten Abweichungen:

$$S = \sum_{i=1}^{n} \frac{(y_i - \eta_i)^2}{\sigma_i^2} = \sum_{i=1}^{n} \frac{(y_i - f(x_i|\theta))^2}{\sigma_i^2} \tag{7.1}$$

Wenn die Messwerte korreliert sind, $\mathrm{cov}(y_i, y_j) \neq 0$, muss man die gesamte Kovarianzmatrix $V_{ij}(y)$ der y-Werte benutzen:

$$S = \sum_{i=1}^{n} \sum_{j=1}^{n} (y_i - \eta_i)\ V_{ij}^{-1}(y)\ (y_j - \eta_j) \tag{7.2}$$

Wenn die Messwerte y_i einer Normalverteilung mit einer Breite σ_i um den wahren Wert $\eta_i = f(x_i|\theta)$ folgen, dann folgt die LS-Funktion S einer χ^2-Verteilung mit

$n_F = n - m$ Freiheitsgraden (Anzahl der Messungen minus Anzahl der aus den Messungen bestimmten Parametern). Da der Erwartungswert von $E(\chi^2) = n_F$ ist, ist die Erwartung für die Verminderung von χ^2 bei Hinzunahme eines Parameters $E(\Delta\chi^2) \geq 1$. Das heisst χ^2 vermindert sich im Mittel um 1, selbst wenn der zusätzliche Parameter nicht notwendig ist. Die Signifikanz für die Notwendigkeit eines Parameters ergibt sich aus $\Delta\chi^2$.

Für den betrachteten Fall normalverteilter Messwerte ergibt sich die Likelihood-Funktion:

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}}\, e^{-\frac{(y_i-\eta_i)^2}{2\sigma_i^2}} \tag{7.3}$$

Daraus berechnet sich die Log-Likelihood-Funktion:

$$\mathcal{L} = -\sum_{i=1}^{n} \frac{(y_i-\eta_i)^2}{2\sigma_i^2} - \sum_{i=1}^{n} \ln\sqrt{(2\pi\sigma_i^2)} = -\frac{1}{2}\, S + const. \tag{7.4}$$

In diesem Fall entspricht also die Parameteroptimierung durch Maximierung von $\mathcal{L}$ genau der Optimierung durch Minimierung der LS-Funktion S, das heisst die ML- und LS-Methoden sind für normalverteilte Messwerte äquivalent. Das LS-Prinzip wird allerdings häufig auch für andere Verteilungen der Messwerte benutzt, weil die formelmässige Behandlung des Problems in der Regel einfacher ist.

7.2 Lineare Anpassung

In der Praxis kommt häufig der Fall vor, dass die Anpassungsfunktion $f(x|\theta)$ eine lineare Funktion der Parameter $\theta = (\theta_1, \ldots, \theta_m)$ ist:

$$f(x|\theta) = \theta_1\, f_1(x) + \ldots + \theta_m\, f_m(x) \tag{7.5}$$

Die f_j können beliebige (also auch nicht-lineare) Funktionen von x sein.

7.2.1 Anpassung der Messwerte an eine Gerade

Für die Hypothese, dass die Messwerte auf einer Geraden liegen sollen, ergibt sich die Anpassungsfunktion ($f_1(x) = 1,\ f_2(x) = x$):

$$f(x|\theta) = \theta_1 \,+\, x\,\theta_2 \tag{7.6}$$

Die Messungen ergeben die n Tripel (x_i, y_i, σ_i) (Abb. 7.1). Wenn die y_i unabhängig sind erhält man die LS-Funktion:

$$S = \sum_{i=1}^{n} \frac{(y_i-\eta_i)^2}{\sigma_i^2} = \sum_{i=1}^{n} \frac{(y_i - \theta_1 \,-\, x_i\,\theta_2)^2}{\sigma_i^2} \tag{7.7}$$

Die Minimierung von S als Funktion der Parameter fordert:

$$\begin{array}{rcccl} \frac{\partial S}{\partial\theta_1} & = & \sum \frac{-2}{\sigma_i^2}(y_i - \theta_1 \,-\, x_i\,\theta_2) & = & 0 \\ \frac{\partial S}{\partial\theta_2} & = & \sum \frac{-2x_i}{\sigma_i^2}(y_i - \theta_1 \,-\, x_i\,\theta_2) & = & 0 \end{array} \tag{7.8}$$

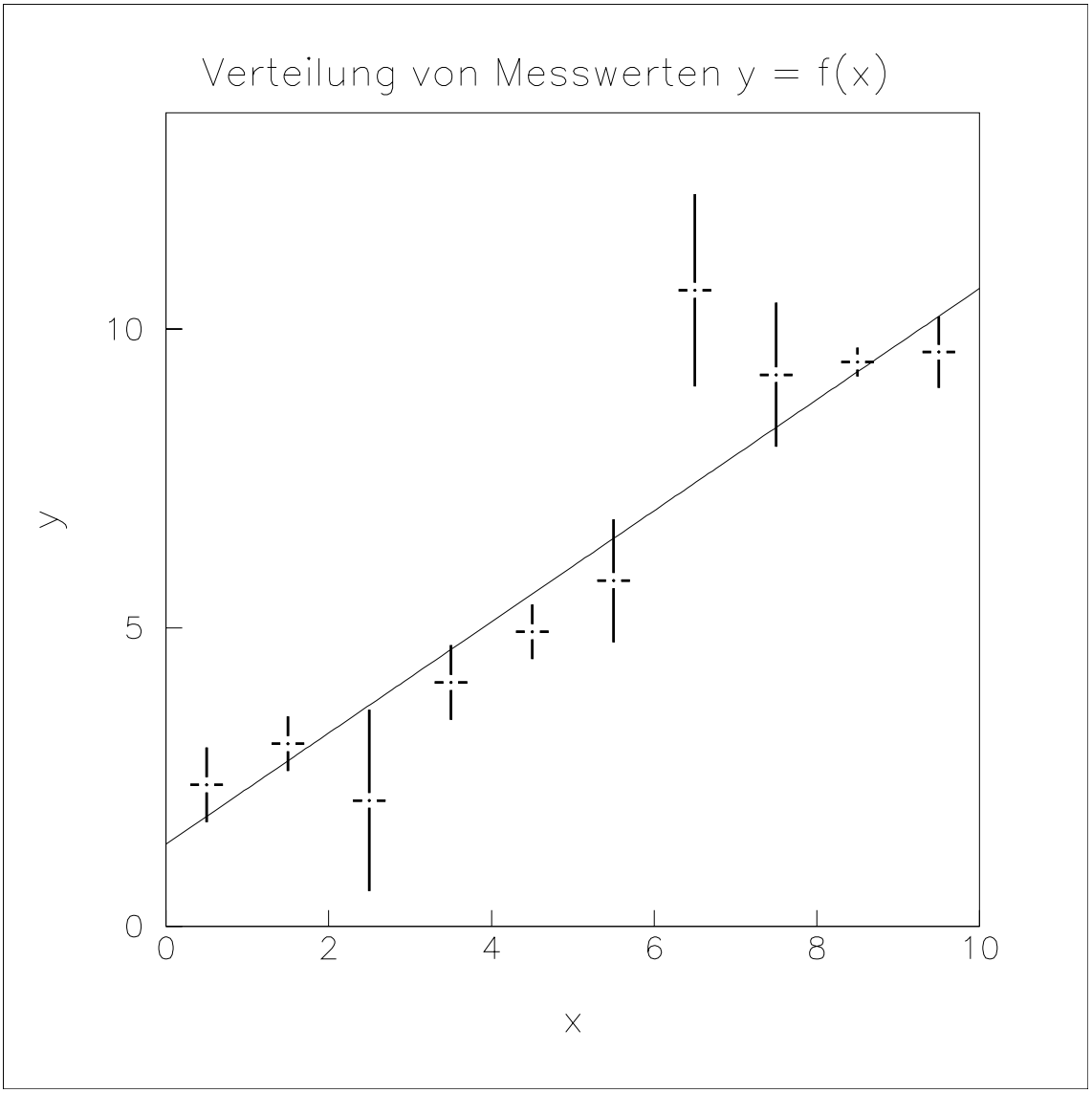


Abbildung 7.1: Messwerte y_i als Funktion von x mit normalverteilten Fehlern. Die Anpassung einer Geraden an die 10 Datenpunkte liefert für Achsenabschnitt und Steigung: $\theta_1 = 1.37 \pm 0.36$, $\theta_2 = 0.93 \pm 0.05$ und $\chi^2 = 11.4$ bei 8 Freiheitsgraden, entsprechend einem Vertrauensniveau von etwa 20%. Die Anpassung wurde mit dem CERN-Programm MINUIT durchgeführt.

Aus der Minimierungsbedingung ergibt sich ein lineares inhomogenes Gleichungssystem für die θ_i. Zur weiteren Behandlung bilden wir folgende Summen, die zum Beispiel auch in entsprechenden Computer-Programmen gebildet werden:

$$\begin{aligned} S_1 &= \sum \tfrac{1}{\sigma_i^2} \\ S_x &= \sum \tfrac{x_i}{\sigma_i^2} \\ S_y &= \sum \tfrac{y_i}{\sigma_i^2} \\ S_{xx} &= \sum \tfrac{x_i^2}{\sigma_i^2} \\ S_{xy} &= \sum \tfrac{x_i y_i}{\sigma_i^2} \end{aligned} \tag{7.9}$$

Damit folgt aus (7.8) für die LS-Schätzung $\hat{\theta}$:

$$\begin{aligned} S_1 \cdot \hat{\theta}_1 &+ S_x \cdot \hat{\theta}_2 &= S_y \\ S_x \cdot \hat{\theta}_1 &+ S_{xx} \cdot \hat{\theta}_2 &= S_{xy} \end{aligned} \tag{7.10}$$

Mit der Determinante der Koeffizientenmatrix

$$D = S_1 S_{xx} - S_x^2 \tag{7.11}$$

ergeben sich durch Auflösung von (7.10) die **LS-Schätzwerte der Parameter**:

$$\begin{aligned} \hat{\theta}_1 &= \tfrac{1}{D}(S_{xx}\, S_y - S_x\, S_{xy}) \\ \hat{\theta}_2 &= \tfrac{1}{D}(S_1\, S_{xy} - S_x\, S_y) \end{aligned} \tag{7.12}$$

Kovarianzmatrix der Parameter: Die Fehler der Parameter ergeben sich aus der Relation (6.42):

$$V_{ij}^{-1} = -\frac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j}\bigg|_{\theta=\hat{\theta}} = +\frac{1}{2}\frac{\partial^2 S}{\partial\theta_i \partial\theta_j}\bigg|_{\theta=\hat{\theta}} \tag{7.13}$$

Die einzelnen Matrixelemente sind:

$$\begin{aligned} \tfrac{1}{2}\tfrac{\partial^2 S}{\partial\theta_1^2} &= \sum \tfrac{1}{\sigma_i^2} &= S_1 \\ \tfrac{1}{2}\tfrac{\partial^2 S}{\partial\theta_1 \partial\theta_2} &= \sum \tfrac{x_i}{\sigma_i^2} &= S_x \\ \tfrac{1}{2}\tfrac{\partial^2 S}{\partial\theta_2^2} &= \sum \tfrac{x_i^2}{\sigma_i^2} &= S_{xx} \end{aligned} \tag{7.14}$$

Die inverse Kovarianzmatrix ist also:

$$V^{-1}(\theta) = \begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \tag{7.15}$$

Die Kovarianzmatrix erhält man aus der Inversion:

$$V(\theta) = \frac{1}{D}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} \tag{7.16}$$

Extrapolationsfehler: Damit lässt sich der y-Wert zu jedem beliebigen x-Wert berechnen:

$$y = \hat{\theta}_1 \, + \, x\,\hat{\theta}_2 \tag{7.17}$$

Der Fehler von y ergibt sich durch Fehlerfortpflanzung:

$$\sigma^2(y) = V_{11} + x^2\,V_{22} + 2\,x\,V_{12} = \frac{1}{D}(S_{xx} + x^2\,S_1 - 2\,x\,S_x) \tag{7.18}$$

Güte der Anpassung: Die Größe

$$\chi^2 = S_{min} = S(\hat{\theta}) \tag{7.19}$$

folgt einer χ^2-Verteilung mit $n_F = n - m = n - 2$ Freiheitsgraden (Anzahl der Messungen minus Anzahl der Parameter) mit dem Erwartungswert

$$E(\chi^2) = n_F. \tag{7.20}$$

Für das Ergebnis der Anpassung (oder des 'Fits') kann man dann den p-Wert wie in Abschnitt 4.3 (Gl. 4.36 und Abb. 4.2) bestimmen, wenn die Messwerte normalverteilt sind. Zum Beispiel ist bei 12 Messungen $n_F = 10$ und man liest folgende p-Werte für $\chi^2 = S_{min}$ ab:

S_{min}	p [%]
8	62.9
10	44.0
12	28.5
16	10.0

Mit dem folgenden Python-Skript kann diese Tabelle reproduziert werden:

```
from scipy import *
for x in [8.,10.,12.,16] :
  print  x, stats.chi2.sf(x,10.)
```

Geringe Vertrauensniveaus können die gleichen Gründe haben, wie in Abschnitt 4.3 angeführt (falsches Modell, falsche Fehler, Untergrund). Wenn das Gauss-Modell nicht zutrifft, kann die Variation von χ^2 um das Minimum immer noch ein gutes Mass für die Bestimmung der Parameter sein. Wie in Abschnitt 6.4.3 ausgeführt, erhält man eine **Schätzung der Standardabweichung eines Parameters**, wenn man diesen Parameter so variiert (die anderen Parameter bleiben fest), dass sich χ^2 um

$$\Delta\chi^2 = 1 \tag{7.21}$$

ändert.

7.2.2 Anpassung einer allgemeinen linearen Funktion der Parameter

Wir wollen jetzt die LS-Anpassung einer allgemeinen linearen Funktion von m Parametern betrachten:

$$f(x|\theta) = \theta_1\,f_1(x) + \ldots + \theta_m\,f_m(x) \tag{7.22}$$

Die LS-Anpassung an n Messwerte y_i an den Punkten x_i hat $n_F = n - m$ Freiheitsgrade. Es wird zugelassen, dass die Messwerte nicht unabhängig sind, dass also die Kovarianzmatrix $V(y)$ nicht-verschwindende $\mathrm{cov}(y_i, y_j)$-Terme hat.

Die Erwartungswerte für die n Messwerte y_i sind dann:

$$\eta_i = \theta_1 f_1(x_i) + \ldots + \theta_m f_m(x_i) = \sum_{j=1}^{m} \theta_j f_j(x_i) \tag{7.23}$$

Um eine kompakte Schreibweise zu erhalten, definieren wir die $n \times m$-Matrix H:

$$H_{ij} = f_j(x_i) \tag{7.24}$$

Damit wird (7.23):

$$\eta_i = \sum_{j=1}^{m} H_{ij}\, \theta_j \;\Rightarrow\; \vec{\eta} = H\, \theta \tag{7.25}$$

Mit der Kovarianzmatrix $V(y)$ der Messwerte ergibt sich dann die LS-Funktion (zur Abkürzung soll im Folgenden $V(y) = V$ gesetzt werden; die Kovarianzmatrix der Parameter wird dann $V(\theta)$ genannt):

$$S = (\vec{y} - H\,\theta)^T\, V^{-1}\, (\vec{y} - H\,\theta) \tag{7.26}$$

Die **Minimierungsbedingung** fordert, dass der Gradient von S bezüglich der Parameter verschwindet:

$$\vec{\nabla}_\theta S = -2\, H^T\, V^{-1}\, (\vec{y} - H\,\theta) = 0 \tag{7.27}$$

Daraus ergibt sich ein lineares Gleichungssystem für θ:

$$H^T\, V^{-1}\, H\, \theta = H^T\, V^{-1}\, \vec{y} \tag{7.28}$$

Wenn $H^T\, V^{-1}\, H$ nicht singulär und damit invertierbar ist, ist die Lösung:

$$\hat{\theta} = (H^T\, V^{-1}\, H)^{-1}\, H^T\, V^{-1}\, \vec{y} \tag{7.29}$$

Durch Matrixinversionen lassen sich die Lösungen im Prinzip exakt bestimmen. Allerdings wird man bei $m > 3$ auf numerische Methoden für die Matrixinversionen zurückgreifen müssen.

Kovarianzmatrix der Parameter: Nach (7.29) ergeben sich die Parameter θ aus einer linearen Transformation der Messwerte:

$$\hat{\theta} = (H^T\, V^{-1}\, H)^{-1}\, H^T\, V^{-1}\, \vec{y} = A\, \vec{y} \tag{7.30}$$

Dann ergibt sich nach (3.63) die Kovarianzmatrix der Parameter θ durch Fehlerfortpflanzung als lineare Transformation der Kovarianzmatrix der Messwerte $\vec{y}$:

$$V(\theta) = A \cdot V(y) \cdot A^T \tag{7.31}$$

Nach Einsetzen von A erhält man:

$$V(\theta) = A \cdot V(y) \cdot A^T = (H^T\, V^{-1}\, H)^{-1}\, H^T\, V^{-1}\, V\, \left[(H^T\, V^{-1}\, H)^{-1}\, H^T\, V^{-1}\right]^T = (H^T\, V^{-1}\, H)^{-1}$$

$$V(\theta) = (H^T\, V^{-1}\, H)^{-1} \tag{7.32}$$

Der Ausdruck $(H^T\, V^{-1}\, H)^{-1}$ ist bereits zur Lösung der Gleichung (7.29) für die Parameter berechnet worden.

Zusammenfassung der Formeln für die lineare Anpassung: Die beste Anpassung ergibt sich fuer die Parameter nach (7.29):

$$\hat{\theta} = (H^T V^{-1} H)^{-1} H^T V^{-1} \vec{y} \tag{7.33}$$

Die Parameter haben die Kovarianzmatrix (7.32)

$$V(\theta) = (H^T V^{-1} H)^{-1} \tag{7.34}$$

Der χ^2-Wert der Anpassung ist:

$$\chi^2_{min} = S(\hat{\theta})) = (\vec{y} - H\,\hat{\theta})^T\, V^{-1}\, (\vec{y} - H\,\hat{\theta}) \tag{7.35}$$

In MATLAB (oder mit Python) lassen sich diese Formeln mit den Matrixoperationen sehr einfach programmieren. Ein Beispiel ist in Abb. 7.2 gezeigt.

Beispiel: Wir betrachten den Fall, den wir im vorigen Abschnitt 7.2.1 bereits speziell behandelt haben: Geradengleichung ($m = 2$), unabhängige Messungen y_i:

$$H = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{pmatrix}, \qquad V^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & 0 & 0 \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ 0 & 0 & 0 & 0 & \frac{1}{\sigma_n^2} \end{pmatrix} \tag{7.36}$$

Die benötigten Produkte dieser Matrizen sind:

$$V^{-1} H = \begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{x_1}{\sigma_1^2} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \frac{1}{\sigma_n^2} & \frac{x_n}{\sigma_n^2} \end{pmatrix} = \left(H^T V^{-1}\right)^T \;\Rightarrow\; H^T V^{-1} \vec{y} = \begin{pmatrix} \sum \frac{y_i}{\sigma_i^2} \\ \sum \frac{x_i y_i}{\sigma_i^2} \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix} \tag{7.37}$$

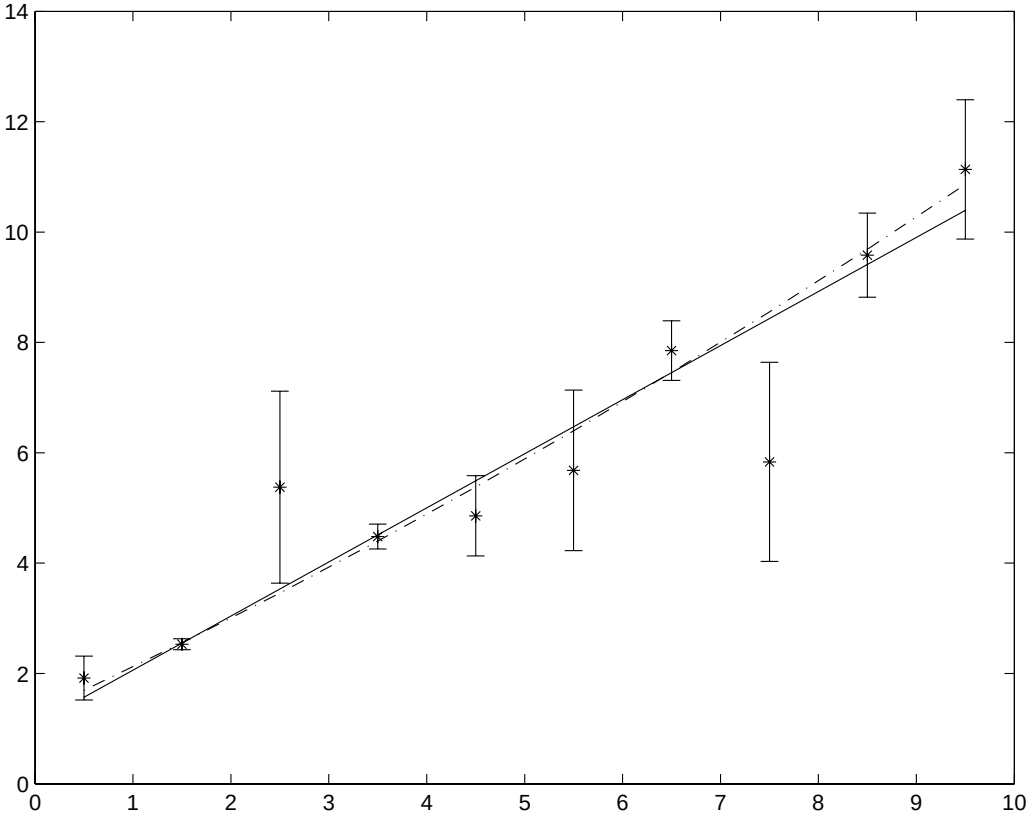
$$H^T V^{-1} H = \begin{pmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{pmatrix} = \begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \tag{7.38}$$

Damit wird also die Gleichung (7.10) reproduziert:

$$\begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix} \tag{7.39}$$

Anpassung an ein orthogonales Funktionensystem: Wenn die Messwerte y_i unabhängig sind, also ihre Kovarianzmatrix diagonal,

$$V_{ij}(y) = \sigma_i^2(y) \cdot \delta_{ij}, \tag{7.40}$$




```
Fit mit 2 Parametern y= a0 + a1*x
---------------------------------
Parameter mit Fehlern:                Kovarianzmatrix:

    1.0787 +/-   0.1482                  0.0220   -0.0071
    0.9805 +/-   0.0589                 -0.0071    0.0035

chi2 = 6.0240      chi2/nf = 0.7530


Fit mit 3 Parametern y= a0 + a1*x +a2*x^2
---------------------------------
Parameter mit Fehlern:                Kovarianzmatrix:

    1.2780 +/-   0.2883                  0.0831   -0.0552    0.0060
    0.8239 +/-   0.2032                 -0.0552    0.0413   -0.0047
    0.0195 +/-   0.0242                  0.0060   -0.0047    0.0006

chi2 = 5.3751      chi2/nf = 0.7679
```

Abbildung 7.2: Beispiel für eine lineare Anpassung (mit einem MATLAB-Programm): Ein Polynom 1. Grades (durchgezogene Linie) oder 2. Grades (gestrichelt) wird an 10 Messwerte mit normalverteilten Fehlern angepasst. Die Messwerte sind ursprünglich entlang einer Geraden erzeugt worden. Man sieht an dem Fehler des Koeffizienten a2 des quadratischen Terms, dass dieser Beitrag nicht signifikant ist.

ergibt sich aus (7.32) für die inverse Kovarianzmatrix der Parameter:

$$V_{ij}^{-1}(\theta) = \sum_{k=1}^{n}\sum_{l=1}^{n} H_{ki}\, V_{kl}^{-1}\, H_{lj} = \sum_{k=1}^{n}\sum_{l=1}^{n} f_i(x_k)\,\frac{1}{\sigma_k^2(y)}\delta_{kl} f_j(x_l) = \sum_{k=1}^{n}\frac{f_i(x_k)\,f_j(x_k)}{\sigma_k^2(y)} \tag{7.41}$$

Die Parameter sind unkorreliert, wenn die Nicht-Diagonalelemente von $V^{-1}(\theta)$ Null sind:

$$V_{ij}^{-1}(\theta) = \sum_{k=1}^{n}\frac{f_i(x_k)\,f_j(x_k)}{\sigma_k^2(y)} = \frac{1}{\sigma_i^2(\theta)}\,\delta_{ij}. \tag{7.42}$$

Wenn die Fehler der Messwerte alle gleich sind, $\sigma_k^2(y) = \sigma^2(y)$, folgt aus (7.42) die Orthogonalität der Funktionen f_i in Bezug auf die Messwerte:

$$\sum_{k=1}^{n} f_i(x_k)\,f_j(x_k) = \frac{\sigma^2(y)}{\sigma_i^2(\theta)}\,\delta_{ij}. \tag{7.43}$$

Im Grenzfall einer unendlich großen Stichprobe geht die Summe in (7.43) in ein Integral über den Definitionsbereich Ω der f_i über:

$$\int_\Omega f_i(x)\,f_j(x)\,dx \sim \delta_{ij}. \tag{7.44}$$

Dieses Integral definiert ein Skalarprodukt in dem Raum der Funktionen f_i und (7.44) bedeutet, dass die f_i orthogonale Basisvektoren sind. Eine Anpassung mit orthogonalen Funktionen, erlaubt die sukzessive Hinzunahme weiterer Terme, ohne die bisher bestimmten Parameter wesentlich zu verändern. Das ist zum Beispiel wichtig für die Beurteilung der Signifikanz des Beitrags eines f_i-Terms. Orthogonale Funktionen sind zum Beispiel die sin- und cos-Funktionen einer Fourier-Zerlegung, die Legendre-Polynome, die Kugelflächenfunktionen usw.

Beispiel: Für eine Geradengleichung mit $f_1 = 1;\ f_2 = x$ ergibt sich:

$$\sum_k f_1(x_k)\,f_2(x_k) = \sum_k x_k = n\,\bar{x} \tag{7.45}$$

Mit der Transformation

$$f_2 \to f_2' = x - \bar{x} \tag{7.46}$$

ergibt sich:

$$\sum_k f_1(x_k)\,f_2'(x_k) = \sum_k (x_k - \bar{x}) = n\,\bar{x} - n\,\bar{x} = 0 \tag{7.47}$$

Daraus folgt, dass man den Ursprung der x-Koordinate am günstigsten in den Schwerpunkt $\bar{x}$ zwischen den Messwerten legt (siehe Übungsaufgabe).

Extrapolationsfehler: Mit den Anpassungfunktionen kann man nun y für beliebige x-Werte berechnen:

$$y = \sum_{j=1}^{m} \hat{\theta}_j\, f_j(x) \tag{7.48}$$

Der Fehler in y ergibt sich durch Fehlerfortpflanzung:

$$\sigma^2(y) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\partial y}{\partial\theta_i}\frac{\partial y}{\partial\theta_j}\,V_{ij}(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} f_i(x)\,f_j(x)\,V_{ij}(\theta) \tag{7.49}$$

Güte der Anpassung: Die Güte der Anpassung wird wieder über das minimale $\chi^2 = S_{min}$, wie in im vorigen Abschnitt 7.2.1 besprochen, abgeschätzt.

7.3 Anpassung nicht-linearer Funktionen der Parameter

Wir betrachten jetzt die Anpassung einer beliebigen Funktion $f(x|\theta)$ an die n Messwerte y_i. Die LS-Funktion lautet wie in (7.2):

$$S = \sum_{i=1}^{n} \sum_{j=1}^{n} (y_i - \eta_i)\ V_{ij}^{-1}(y)\ (y_j - \eta_j) \tag{7.50}$$

Diese Funktion soll wieder als Funktion der Parameter minimalisiert werden. Im allgemeinen muss die Lösung $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$, die S minimiert, mit numerischen Methoden iterativ gesucht werden.

Iterationsverfahren: Es sei im ν-ten Iterationsschritt eine Näherung von $\hat{\theta}$ gefunden:

$$\theta^{\nu} = (\theta_1^{\nu}, \ldots, \theta_m^{\nu}). \tag{7.51}$$

Gesucht ist ein Inkrement $\Delta\theta^{\nu}$, das zu der nächsten Näherung für die Parameter führt,

$$\theta^{\nu+1} = \theta^{\nu} + \Delta\theta^{\nu}, \tag{7.52}$$

und das die Näherung verbessert:

$$S(\theta^{\nu+1}) < S(\theta^{\nu}) \tag{7.53}$$

Das Verfahren wird abgebrochen, wenn Konvergenz erreicht ist. Als **Konvergenzkriterium** verlangt man in der Regel, dass S sich von einem Schritt zum nächsten um weniger als einen kleinen Betrag ϵ ändert:

$$\left|S(\theta^{\nu+1}) - S(\theta^{\nu})\right| < \epsilon \tag{7.54}$$

Es gibt verschiedenen Verfahren, die Inkremente $\Delta\theta^{\nu}$ zu bestimmen, um das Minimum von S zu finden. Bei vielen Parametern und etwas komplexer strukturierten LS-Funktionen können solche multi-dimensionalen Optimierungsprobleme zu einer mathematischen Herausforderung werden. In der Teilchenphysik wird sehr viel das beim CERN entwickelte Programm MINUIT benutzt, das verschiedene Verfahren zur Auswahl anbietet (Abb. 7.4). Bei komplexen Problemen ist es notwendig, dass der Benutzer die verschiedenen Möglichkeiten kennt und steuernd eingreift. Wichtig sind gute Startwerte θ^0, die man häufig nur durch ein gutes Gespür erhält, um eventuelle Nebenminima im Parameterraum zu vermeiden (Abb. 7.3). Man muss deshalb immer überprüfen, ob die Lösung von den Startwerten abhängt.

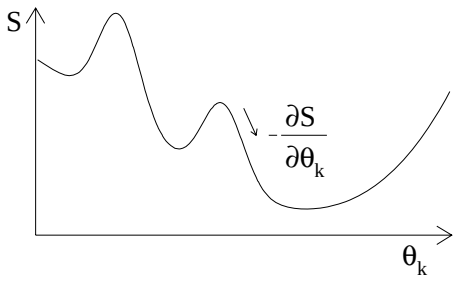


Abbildung 7.3: Beispiel für den Verlauf einer LS-Funktion im Parameterraum.

Gradientenverfahren: Eine naheliegende Möglichkeit, Extremwerte einer Funktion zu finden, ist das Gradientenverfahren: Man geht mit einer vorgewählten Schrittweite $\Delta\theta$ in Richtung des Gradienten der Funktion, im Fall der Minimierung in Richtung des negativen Gradienten (Abb. 7.3). Häufig wird die Schrittweite proportional dem Gradienten gewählt:

$$\Delta\theta^{\nu+1} = -\eta \left(\vec{\nabla}_\theta S\right)_{|\theta^\nu} \tag{7.55}$$

Die Wahl der Schrittweite proportional zum Gradienten von S scheint vernünfig zu sein, weil im Minimum von S Konvergenz erreicht wird und die Schrittweite dann tatsächlich gegen Null geht. Häufig wird der Schrittparameter η aber auch dynamisch angepasst, um zum Beispiel nicht zu lange in Gebieten mit flachem Funktionsverlauf zu verweilen (große Schritte) oder in Bereichen steiler Gradienten auch das Minimum finden zu können (kleine Schritte). Wenn sich in einem Iterationsschritt das Vorzeichen des Gradienten ändert, das Extremum also überschritten wurde, sollte man die Schrittweite verkleinern.

Linearisierung der Anpassungsfunktion: Durch Entwicklung der Anpassungsfunktion nach den Parametern bis zu den linearen Termen, kann man das Problem auf lineare Anpassungen mit Iterationen zurückführen:

$$\eta_i(\theta) = \eta_i(\theta^\nu) + \left(\vec{\nabla}_\theta \eta_i\right)_{|\theta=\theta^\nu} \Delta\theta^\nu + \ldots \tag{7.56}$$

In der ν-ten Iteration sind die Abweichungen der Messwerte von dem Anpassungswert ('Residuen'):

$$\Delta y_i^\nu = y_i - \eta_i(\theta^\nu) \tag{7.57}$$

Mit der Definition der Matrix H

$$H_{ij} = \frac{\partial \eta_i}{\partial \theta_j} \tag{7.58}$$

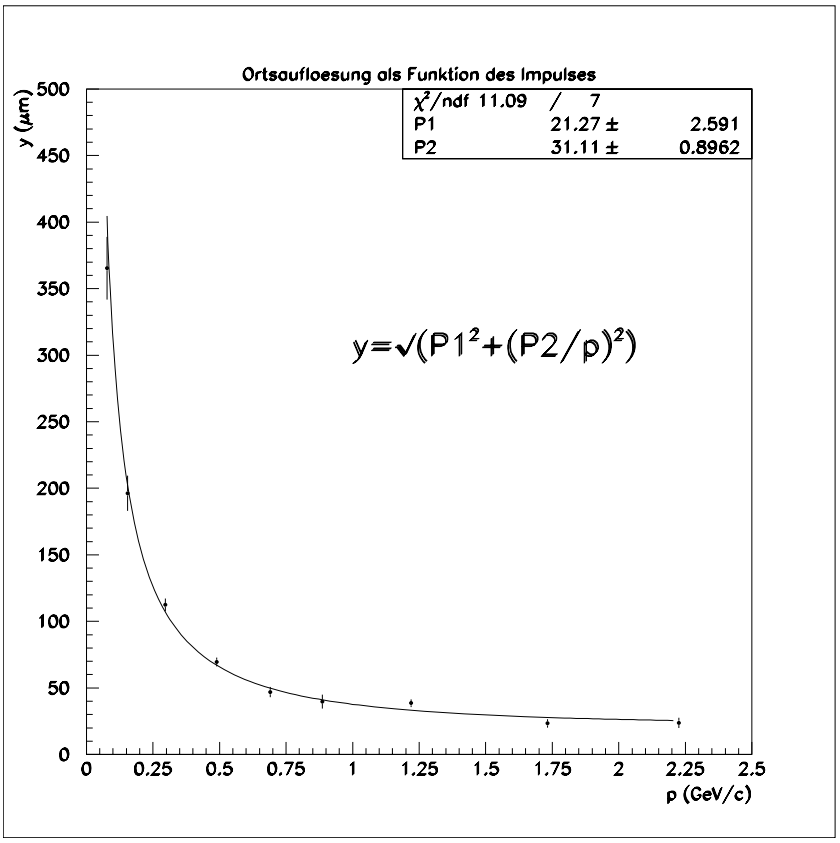
ergibt sich dann die LS-Funktion in der ν-ten Iteration:

$$S^\nu = (\Delta\vec{y}^\nu - H\Delta\theta^\nu)^T V^{-1} (\Delta\vec{y}^\nu - H\Delta\theta^\nu) \tag{7.59}$$

Diese LS-Funktion entspricht völlig derjenigen für die lineare Anpassung (7.26), wenn man die Ersetzung

$$\vec{y} \to \Delta\vec{y}^\nu; \qquad \theta \to \Delta\theta^\nu \tag{7.60}$$

macht.




```
    ********************************************
    *                                          *
    * Function minimization by SUBROUTINE HFITV  *
    * Variable-metric method                   *
    * ID =          0  CHOPT = S               *
    *                                          *
    ********************************************
 Convergence when estimated distance to minimum (EDM) .LT.  0.10E+01

 FCN=   11.08658     FROM MIGRAD    STATUS=CONVERGED     37 CALLS    38 TOTAL
                     EDM=  0.32E-05    STRATEGY= 1      ERROR MATRIX ACCURATE

  EXT PARAMETER                                   STEP         FIRST
  NO.   NAME        VALUE          ERROR          SIZE      DERIVATIVE
   1      P1        21.272        2.5912       0.28605       0.10326E-02
   2      P2        31.111       0.89618       0.99125E-01   0.37551E-03

 CHISQUARE = 0.1584E+01  NPFIT =     9
```

Abbildung 7.4: Beispiel für die Anwendung des Programmes MINUIT. Unter der graphischen Darstellung ist der Ausdruck des Programmes MINUIT gezeigt. Eine nicht-lineare Funktion der Parameter, angegeben in der Graphik, wird an Messwerte angepasst.

Kapitel 8

Signifikanzanalysen

8.1 Einführung

In den vorhergehenden Kapiteln haben wir Methoden kennengelernt, um aus Messungen Hypothesen abzuleiten, die mit den Daten verträglich sind. Es kann sich dabei um diskrete Hypothesen handeln oder auch um Funktionen, deren Parameter so bestimmt werden, dass die Funktion die beste Anpassung an die Daten darstellt. Die Bestimmung der Güte der Anpassung und der Signifikanz der Richtigkeit einer Hypothese haben wir für spezielle Fälle schon mehrfach angesprochen. Im Folgenden wollen wir allgemeiner statistische Tests zur Bestimmung der Signifikanz von Hypothesen besprechen, einerseits für die Signifikanz einer einzelnen Hypothese oder für die Entscheidung zwischen mehreren Hypothesen.

Wir nehmen an, es liegen Messwerte $(x_1, \ldots, x_n)$ vor, von denen eine Hypothese H_0 ('Null-Hypothese') abgeleitet wird, die zu prüfen ist. Zum Beispiel würde bei einer ML-Anpassung einer Funktion $f(x|\theta)$ die Funktion mit dem Parametersatz θ_0, der die Likelihood-Funktion maximiert, der Null-Hypothese entsprechen. Zur Beurteilung der Signifikanz der Hypothese definiert man eine Testgröße t als eine Abbildung der Messdaten auf eine Größe, die möglichst die gesamte Information der Messung in komprimierter Form zusammenfasst:

$$(x_1, \ldots, x_n) \to t(x_1, \ldots, x_n | f, \theta_0) \tag{8.1}$$

Die Testgröße ('test statistic') hängt von den Messungen und der Hypothese H_0 ab, die hier durch die Anpassungsfunktion mit den Parametern θ_0 gegeben ist. Ein uns bereits bekanntes Beispiel für eine Testfunktion ist die χ^2-Funktion. Die Testfunktion ist abhängig von der speziellen Stichprobe $(x_1, \ldots, x_n)$ und ist damit ebenfalls eine Zufallsvariable, die einer Wahrscheinlichkeitsverteilung $g(t)$ folgen soll. Dabei ist zu beachten, dass $g(t) = g(t|H_0)$ die Wahrscheinlichkeitsverteilung von t für eine feste Hypothese ist und damit von den Messwerten abhängt. Wir werden also keine Wahrscheinlichkeit für die Hypothese formulieren können, sondern nur die Wahrscheinlichkeit für die spezielle Messung bei einer gegebenen Hypothese erhalten.

Als Maß für das Vertrauen in eine Hypothese oder die Güte einer Parameteranpassung bilden wir den p-Wert:

$$p = \int_{t_{mess}}^{\infty} g(t|H_0) dt. \tag{8.2}$$

Der p-Wert (auch 'Signifikanz') ist also die Wahrscheinlichkeit bei Wiederholung der Messungen Ergebnisse zu erhalten, die so gut oder schlechter wie die betrachtete Messung mit der Hypothese verträglich sind. Eine Hypothese wird akzeptiert, wenn der p-Wert größer als ein vorgegebenes Signifikanzniveau α (gleich dem früher eingeführten Konfidenzniveau) ist. Man beachte, dass der p-Wert für eine bestimmte Messung bestimmt wird, während das Signifikanz- oder Vertrauensniveau eine vorgegebene Größe ist (zum Beispiel $\alpha = 5\%$ oder 10%). Weiterhin ist zu beachten, dass alle p-Werte gleich wahrscheinlich sind, wenn die Messungen tatsächlich den Verteilungen entsprechend der Hypothese folgen.

8.2 Prüfung von Hypothesen

In diesem Abschnitt sollen einige spezielle Hypothesentests behandelt werden.

8.2.1 χ^2-Test

Der χ^2-Test, der bereits in Abschnitt 4.3 besprochen wurde, wird benutzt, um Messwerte y_i, $i = 1, \ldots, n$, an den Punkten x_i mit Erwartungswerten η_i zu vergleichen. Wenn $\eta_i = \eta(x_i|\theta_0)$ die Erwartungswerte von Verteilungen mit Varianzen σ_i^2 sind, ist die Testfunktion:

$$t = \chi^2 = \sum_{i=1}^{n} \frac{(y_i(x_i) - \eta_i)^2}{\sigma_i^2}. \tag{8.3}$$

Wenn die y_i Stichprobenwerte aus Normalverteilungen sind, folgt t einer χ^2-Verteilung (4.23) mit $n_F = n - m$ Freiheitsgraden, wobei m die Anzahl der bestimmten Parameter ist. Der χ^2-Test wird auch häufig für nur näherungsweise normalverteilte Messwerte benutzt. Ein häufig vorkommendes Beispiel ist die Beschreibung poisson-verteilter Histogrammeinträge n_i durch Erwartungswerte $\nu_i = \nu(i|\theta_0)$ mit Varianzen $\sigma_i^2 = \nu_i$ (also die Varianzen von den Erwartungswerten und nicht von den Messwerten abgeleitet):

$$t = \chi^2 = \sum_{i=1}^{n} \frac{(n_i(x_i) - \nu_i)^2}{\nu_i}. \tag{8.4}$$

Der p-Wert zu einem χ^2-Wert χ_m^2 einer Messung mit n_F Freiheitsgraden ist in den Abbildungen 4.3 und 4.4 in Abschnitt 4.3 abzulesen.

8.2.2 Studentsche t-Verteilung

Die Fragestellung ist, ob der Mittelwert $\bar{x} = \sum_i x_i/n$ einer Stichprobe x_i, $i = 1, \ldots, n$, mit einem theoretischen Mittelwert μ vereinbar ist. Die Varianz des Mittelwertes wird mit der Varianz der Stichprobe s^2 entsprechend (4.11) zu s^2/n abgeschätzt. Die Wahrscheinlichkeitsdichte für die Testgröße

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \tag{8.5}$$

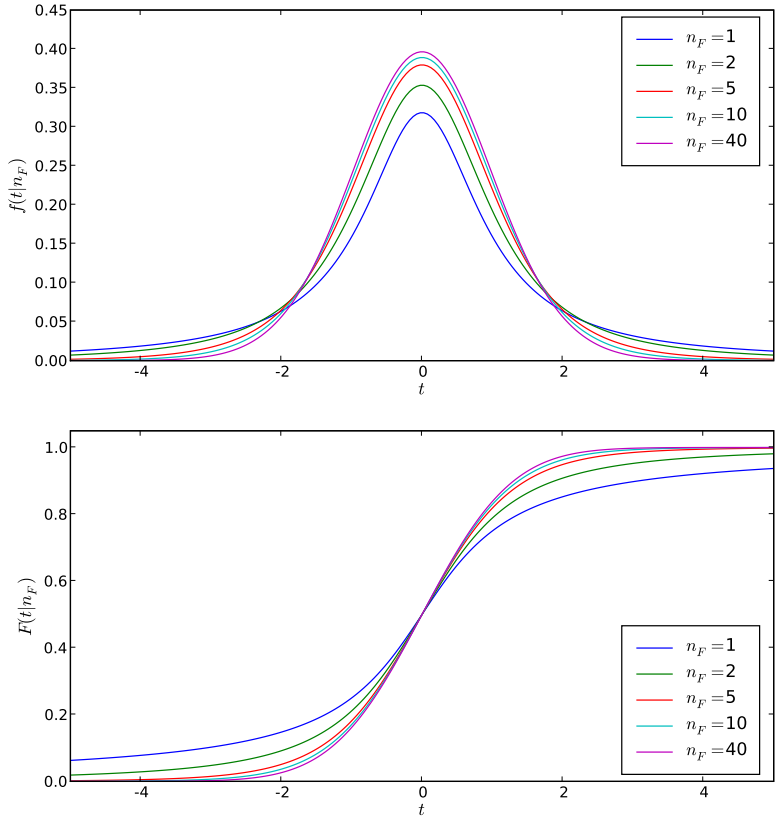


Abbildung 8.1: Oben: Die Studentsche t-Verteilung für verschiedene Freiheitsgrade k. Unten: kumulative Verteilungsfunktion der t-Verteilung.

folgt einer t-Verteilung,

$$f(t|n_F) = \frac{1}{\sqrt{n_F \pi}} \frac{\Gamma\left(\frac{n_F+1}{2}\right)}{\Gamma\left(\frac{n_F}{2}\right)} \left(1 + \frac{t^2}{n_F}\right)^{-\frac{n_F+1}{2}} \qquad (-\infty < t < +\infty). \tag{8.6}$$

Die Verteilung ist symmetrisch um $t = 0$, ist für $n_F = 1$ eine Cauchy-Verteilung $\sim 1/(1+t^2)$ und nähert sich für große n_F einer Gauss-Verteilung an (Abb. 8.1). Die t-Verteilung und deren kumulative Verteilungsfunktion findet man tabelliert in der entsprechenden Literatur. Das Python-Skript

```
from scipy import *
for t in [0.,0.5,1.0,1.5,2.0] :
  print  t, stats.t.sf(t,10.)
```

berechnet folgende Tabelle für die p-Werte zu den angegebenen Werten von t und $n_F = 10$:

t	p [%]
0.0	50
0.5	31
1.0	17
1.5	8.2
2.0	3.7

Beispiel: Es seien drei Messungen ($x_1 = -1,\ x_2 = 0,\ x_3 = 1$) mit dem Mittelwert $\bar{x} = 0$ gegeben. Was ist der p-Wert, wenn der wahre Mittelwert $\mu = -1$ ist (Beispiel aus [3])? Mit den berechneten Zahlenwerten

$$s^2 = \frac{1}{2}(1+0+1) = 1.0, \qquad t = \frac{\bar{x}-\mu}{\sqrt{s^2/n}} = \sqrt{3} = 1.732, \qquad n_F = n-1 = 2$$

ergibt das obige Python-Skript einen p-Wert von 11%, Das heißt, bei einem vorgegebenen Signifikanzniveau von zum Beispiel 5% oder 10% wäre die Hypothese zu akzeptieren.

8.2.3 F-Verteilung

Vergleich von Streuungen zweier Stichproben des Umfangs n_1 und n_2 mit gleichem Erwartungswert. Frage: haben beide Grundgesamtheiten die Gleiche Varianz. Die Fragestellung tritt zum Beispiel auf, wenn eine Größe mit zwei verschiedenen Apparaturen gemessen wird und zu klären ist, ob beide Apparaturen die gleiche Auflösung haben.

Dazu werden die empirischen Varianzen $s_1^2 = \chi_1^2/(n_1-1)$ und $s_2^2 = \chi_2^2/(n_2-1)$ nach (4.11) bestimmt. Die Testgröße ist der Quotient

$$F = \frac{s_1^2}{s_2^2} \tag{8.7}$$

Die Wahrscheinlichkeitsverteilung lässt sich aus mit Hilfe der χ^2-Verteilungen zu den Freiheitsgraden $\nu_1 = n_1 - 1$ und $\nu_2 = n_2 - 1$ ableiten (Abb. 8.2), wenn die Stichproben normalverteilt sind:

$$f(F|\nu_1,\nu_2) = \nu_1^{\frac{\nu_1}{2}}\nu_2^{\frac{\nu_2}{2}} \cdot \frac{\Gamma(\frac{\nu_1}{2}+\frac{\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \cdot \frac{F^{\frac{\nu_1}{2}-1}}{(\nu_1 F+\nu_2)^{\frac{\nu_1+\nu_2}{2}}} \qquad (0 \le F \le +\infty) \tag{8.8}$$

Die Formel wird zum Beispiel in [1] abgeleitet. Der Erwartungswert der Verteilung ist

$$E(F) = \frac{\nu_2}{\nu_2 - 2} \qquad \text{für } \nu_2 \gg 2. \tag{8.9}$$

Wegen des Quotienten in der Verteilung gilt

$$f(F_{12}|\nu_1,\nu_2) = f(\frac{1}{F_{12}}|\nu_2,\nu_1), \tag{8.10}$$

wobei jeweils ein F-Wert größer als 1 und der andere kleiner als 1 ist. Für einen Signifikanztest benutzt man üblicherweise den größeren der beiden Werte und verlangt

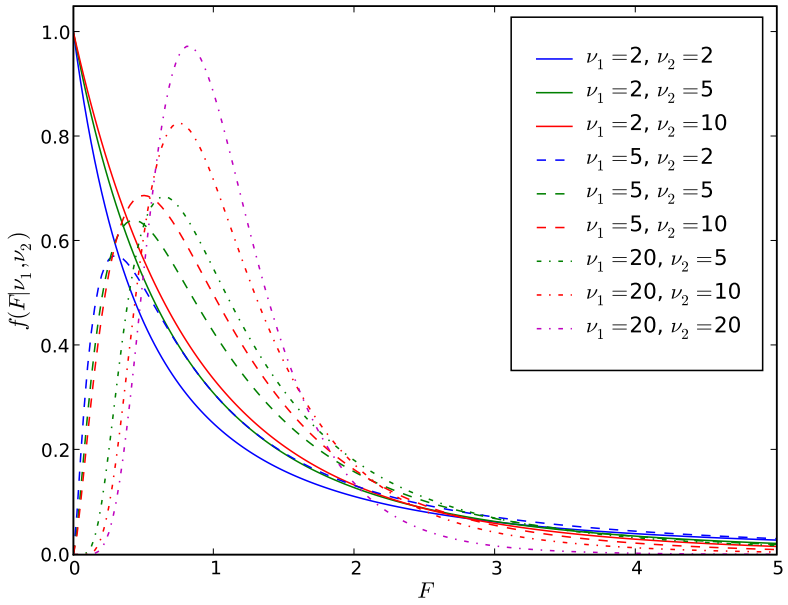


Abbildung 8.2: Wahrscheinlichkeitsdichte der F-Verteilung für verschiedene Freiheitsgrade der beiden beteiligten Stichproben.

wie auch bei den anderen Tests, dass das die Wahrscheinlichkeit, einen F-Wert größer als den gemessenen zu erhalten, ein vorgegebenes Signifikanzniveau übersteigt. Es ist allerdings zu beachten, dass mit der Einschänkung $F \geq 1$ die Normierung der F-Verteilung um eine Faktor 2 gegenüber der tabellierten Funktionen skaliert werden muss.

Man kann F-Werte und ihre Signifikanzen in Tabellen finden oder zum Beispiel mit Python berechnen:

```
>>> from scipy import *
>>> for F in [1.,2.,3.] : print F, 2.*stats.f.sf(F,10.,10.)
1.0 1.0
2.0 0.289691612051
3.0 0.0978546142578
```

Für vorgegebene Signifikanzniveaus kann man andererseits den dazugehörigen F-Wert berechnen:

```
>>> from scipy import *
>>> for p in [0.10,0.05,0.01] : print p,  stats.f.isf(p/2.,10.,10.)
0.1 2.97823701608
0.05 3.7167918646
0.01 5.84667842506
```

Beispiel: Mit zwei Messapperaturen wird jeweils eine Messung gemacht. Die Ergebnisse sind: $n_1 = 10,\ s_1^2 = 3.7;\ n_2 = 7,\ s_2^2 = 6.5$ (aus [1]). Daraus ergibt sich $F = 6.5/3.7 = 1.8$ mit einem p-Wert von 41%, mit dem die Hypothese sicherlich akzeptiert wird.

```
>>> from scipy import *
>>> F=1.8
>>> print F,  2.*stats.f.sf(F,6.,9.)
1.8 0.410775850533
```

8.2.4 Kolmogorov-Smirnov-Test

Es geprüft werden, ob eine Stichprobe $(x_1, \ldots, x_n)$ einer Gesamtheit mit Wahrscheinlichkeitsdichte $f(x)$ entnommen ist. Dazu könnte man die Daten in x-Intervalle einteilen und mit einem χ^2-Test die Hypothese überprüfen. Problematisch wird dieser Test bei kleinen Anzahlen in den Bins. Auch ist der χ^2-Test nicht sehr sensitiv auf tendentielle Abweichungen nach oben oder unten in begrenzten x-Bereichen. Durch Einteilung der Daten in Überschuß- und Unterschussbereiche könnte man solche Tendenzen sichtbar machen. Aber wie bestimmt man dann die p-Werte, da ja eine solche Neueinteilung auf einer subjektiven Einschätzung beruht?

Mit dem Kolmogorov-Smirnov-Test kann man die Verträglichkeit der Stichprobe mit einer Wahrscheinlichkeitsdichte ohne Intervalleinteilung prüfen. Dazu wird die Verteilungsfunktion

$$F(x) = \int_{-\infty}^{x} f(\xi) d\xi \tag{8.11}$$

verglichen mit der Schätzung dieses Integrals mit Hilfe der Stichprobe:

$$F_n(x) = \frac{\text{Anzahl der } x_i\text{−Werte} \leq x}{n}. \tag{8.12}$$

Die Testgröße ist proportional zu der größten Differenz zwischen den beiden kumulativen Verteilungen:

$$t = \max |F_n(x) - F(x)|. \tag{8.13}$$

Werte von t zu vorgegebenen Signifikanzniveaus sind für verschiedene Freiheitsgrade in Tabelle 8.1 aufgelistet. Für große n_F-Werte ist der p-Wert durch eine unendliche Reihe gegeben:

$$p = 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2 n_F t^2) \tag{8.14}$$

Mit dem Python-Skript

```
>>> from scipy import *
>>> t=0.447
>>> print t,  stats.ksone.sf(t,5.)
0.447 0.099980005201
```

wird $p = 0.1$ für $t = 0.447$ bei $n_F = 5$ in der Tabelle 8.1 reproduziert. Andererseits lässt sich mit der inversen Funktion stats.ksone.isf bei vorgegebenem p-Wert oder Signifikanzniveau und Freiheitsgrad der dazugehörige t-Wert bestimmen:

```
>>> from scipy import *
>>> p=0.1
>>> print p,  stats.ksone.isf(p,5.)
0.1 0.446980061221
```

Tabelle 8.1: Kolmogorov-Smirnov-Test (einseitig): Tabelle der Werte t_0 zu einem Signifikanzniveau für verschiedene Freiheitsgrade.

	Signifikanzniveau				
n_F	0.1	0.05	0.025	0.01	0.005
1	0.9000	0.9500	0.9750	0.9900	0.9950
2	0.6838	0.7764	0.8419	0.9000	0.9293
3	0.5648	0.6360	0.7076	0.7846	0.8290
4	0.4927	0.5652	0.6239	0.6889	0.7342
5	0.4470	0.5094	0.5633	0.6272	0.6685
6	0.4104	0.4680	0.5193	0.5774	0.6166
7	0.3815	0.4361	0.4834	0.5384	0.5758
8	0.3583	0.4096	0.4543	0.5065	0.5418
9	0.3391	0.3875	0.4300	0.4796	0.5133
10	0.3226	0.3687	0.4092	0.4566	0.4889
11	0.3083	0.3524	0.3912	0.4367	0.4677
12	0.2958	0.3382	0.3754	0.4192	0.4490
13	0.2847	0.3255	0.3614	0.4036	0.4325
14	0.2748	0.3142	0.3489	0.3897	0.4176
15	0.2659	0.3040	0.3376	0.3771	0.4042
16	0.2578	0.2947	0.3273	0.3657	0.3920
17	0.2504	0.2863	0.3180	0.3553	0.3809
18	0.2436	0.2785	0.3094	0.3457	0.3706
19	0.2373	0.2714	0.3014	0.3369	0.3612
20	0.2316	0.2647	0.2941	0.3287	0.3524
21	0.2262	0.2586	0.2872	0.3210	0.3443
22	0.2212	0.2528	0.2809	0.3139	0.3367
23	0.2165	0.2475	0.2749	0.3073	0.3295
24	0.2120	0.2424	0.2693	0.3010	0.3229
25	0.2079	0.2377	0.2640	0.2952	0.3166
26	0.2040	0.2332	0.2591	0.2896	0.3106
27	0.2003	0.2290	0.2544	0.2844	0.3050
28	0.1968	0.2250	0.2499	0.2794	0.2997
29	0.1935	0.2212	0.2457	0.2747	0.2947
30	0.1903	0.2176	0.2417	0.2702	0.2899
31	0.1873	0.2141	0.2379	0.2660	0.2853
32	0.1844	0.2108	0.2342	0.2619	0.2809
33	0.1817	0.2077	0.2308	0.2580	0.2768
34	0.1791	0.2047	0.2274	0.2543	0.2728
35	0.1766	0.2018	0.2242	0.2507	0.2690
36	0.1742	0.1991	0.2212	0.2473	0.2653
37	0.1719	0.1965	0.2183	0.2440	0.2618
38	0.1697	0.1939	0.2154	0.2409	0.2584
39	0.1675	0.1915	0.2127	0.2379	0.2552
40	0.1655	0.1891	0.2101	0.2349	0.2521
> 40	$1.07/\sqrt{n_F}$	$1.22/\sqrt{n_F}$	$1.36/\sqrt{n_F}$	$1.52/\sqrt{n_F}$	$1.63/\sqrt{n_F}$

Quelle: http://www.york.ac.uk/depts/maths/tables

8.3 Vertrauensintervalle

Die Angabe von Vertrauensintervallen im Parameterraum, das ist der Bereich in dem der gesuchte Satz von Parametern mit einer vorgegebenen Wahrscheinlichkeit liegt, ist problematisch, weil meistens die Wahrscheinlichkeiten für Parameter nicht bekannt sind. Deshalb entbrennen auf diesem Feld auch die heftigsten Kämpfe zwischen Bayes-Anhängern und Frequentisten. Im PDG Review [15] werden beide Sichtweisen diskutiert und weiterführende Literatur angegeben.

8.3.1 Bayes-Vertrauensintervalle

Die Wahrscheinlichkeitsdichte für die Parameter θ bei einem gegebenen Satz von Messung x ist nach dem Bayes-Theorem:

$$p(\theta|x) = \frac{L(x|\theta)p(\theta)}{\int L(x|\theta')p(\theta')d\theta'}. \tag{8.15}$$

Das Problem ist das die 'A-Priori-Wahrscheinlichkeit' $p(\theta)$ im allgemeinen nicht bekannt ist und Annahmen gemacht werden müssen (die einfachste Annahme wäre eine Gleichverteilung). Vorteilhaft ist diese Formulierung für den Ausschluss unphysikalischer Bereiche, in denen man $p(\theta) = 0$ setzen kann (zum Beispiel, damit eine Zählrate nicht negativ wird).

Das Intervall $[\theta_u, \theta_o]$, mit dem das gesuchte θ mit eine (Posterior-)Wahrscheinlichkeit von $1 - \alpha$ liegt, wird bestimmt zu:

$$1 - \alpha = \int_{\theta_u}^{\theta_o} p(\theta|x)d\theta \tag{8.16}$$

Das vorgegebene Vertrauensniveau $1-\alpha$ kann mit verschiedenen Intervallgrenzen erreicht werden. Naheliegend ist eine Auswahl, so dass jeweils unterhalb und oberhalb des Intervalls die Wahrscheinlichkeiten $\alpha/2$ sind. Man kann auch den Vertrauensbereich so festzulegen, dass $p(\theta|x)$ in dem Bereich immer größer ist als außerhalb. Wenn man obere oder untere Ausschliessungsgrenzen zu einem Vertrauensniveau $1 - \alpha$ geben will, kann man in (8.16) $\theta_u = 0$ beziehungsweise $\theta_o = \infty$ setzen.

8.3.2 'Klassische' Vertrauensintervalle

'Frequentisten' benutzen die Neyman-Konstruktion der Vertrauensintervalle wie in Abb. 8.3 gezeigt. Statt die Wahrscheinlichkeitsdichte für die Parameter bestimmt man die Wahrscheinlichkeitsdichte $f(x|\theta)$ der Messwerte x bei festen Parametern θ. Für verschiedene Parameter θ werden nun die Grenzen x_1 und x_2 bestimmt, in denen mit einer Wahrscheinlichkeit $1 - \alpha$ die Messwerte liegen:

$$P(x_1 < x < x_2|\theta) = 1 - \alpha = \int_{x_1}^{x_2} f(x|\theta)dx. \tag{8.17}$$

Diese Intervalle werden nun kontinuierlich als Funktion von θ bestimmt, so dass man das Band ('confidence belt') wie in Abb. 8.3 erhält. Diese Konstruktion kann, beziehungsweise sollte, vor der Messung gemacht werden. Wenn das Messergebnis

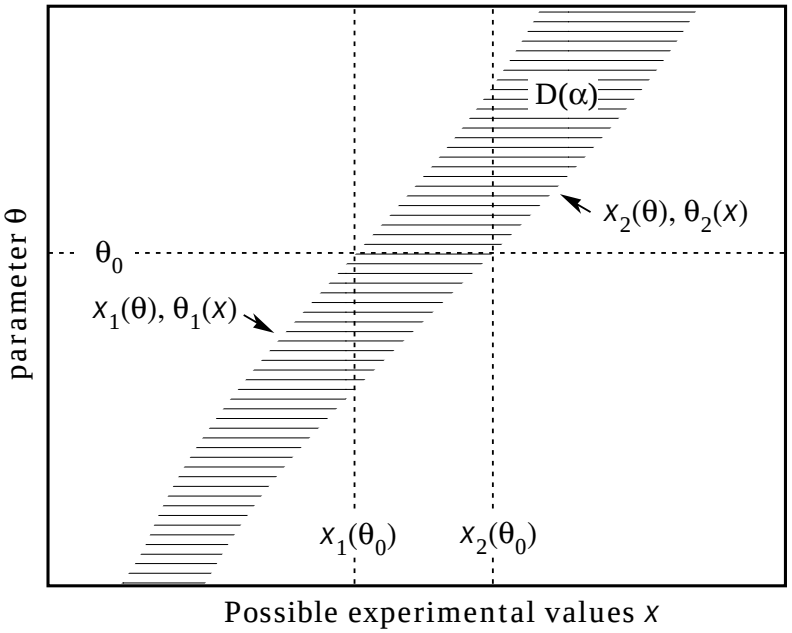


Abbildung 8.3: Konstruktion des Vertrauensbandes (siehe Text); aus [15].

dann x_0 ist, ergeben sich die unteren und oberen Grenzen θ_1, θ_2 als die Schnittpunkte der vertikalen Linie $x = x_0$ mit dem unteren beziehungsweise oberen Bandrand. Die Bandränder werden in Abb. 8.3 als Funktionen $\theta_1(x)$ und $\theta_2(x)$ bezeichnet.

Auch hier ist die Lage des Vertrauensintervals zunächst nicht festgelegt. Feldmann und Cousins[1] haben eine Anordnung nach Likelihood-Verhältnissen vorgeschlagen. Bei der Bestimmung des Vertrauensintervalles x_1, x_2) zu festem θ (horizontal) wird zu jedem x-Wert der Parameter θ_{best} gesucht (entlang der Vertikalen), für den die Likelihood-Funktion an dieser x-Position maximal ist:

$$L(x|\theta_{best}) \geq L(x|\theta) \quad \forall\, \theta \text{ bei festem } x. \tag{8.18}$$

Das Verhältnis

$$\lambda = \frac{L(x|\theta)}{L(x|\theta_{best})} \tag{8.19}$$

wird als Funktion von x bei festem θ (also in der Horizontalen) bestimmt und es werden die x-Werte nach der Größe von λ geordnet, wobei dem Punkt mit dem größten λ der Rang 1 zugeordnet wird. Das Vertrauensinterval wird nun sukzessive durch Hinzunahme von x-Werten entsprechend ihrer Rangfolge so aufgebaut, bis das vorgegebene Vertrauensniveau $1-\alpha$ erreicht ist. Dazu werden bei diskreten Verteilungen die Wahrscheinlichkeiten summiert und bei kontinuierlichen Verteilungen wird das entsprechende Integral in diskreten Schritten approximiert.

Die Feldmann-Cousins-Konstruktion stellt unter anderem sicher, dass die beste Parameteranpassung in dem Vertrauensinterval jedenfalls enthalten ist. Zudem liefert das Verfahren ein Rezept, wann als Ergebnis ein zentrales Vertrauensinterval und wann eine obere oder untere Grenze angegeben werden sollen. Eine Grenze wird angegeben, wenn das Band für eine x-Messung die untere oder obere Grenze des er-

[1] G.J. Feldman and R.D. Cousins, Phys. Rev. D57, 3873 (1998).

Tabelle 8.2: Konstruktion der Vertrauensintervalle für ein Signal μ, wenn n Ereignisse gemessen werden und der Untergrund $b = 3.0$ ist. Das Beispiel in der Tabelle zeigt die Berechnung für $\mu = 0.5$.

n	$L(n\|\mu)$	μ_{best}	$L(n\|\mu_{best})$	λ	Rang
0	0.030	0.0	0.050	0.607	6
1	0.106	0.0	0.149	0.708	5
2	0.185	0.0	0.224	0.826	3
3	0.216	0.0	0.224	0.963	2
4	0.189	1.0	0.195	0.966	1
5	0.132	2.0	0.175	0.753	4
6	0.077	3.0	0.161	0.480	7
7	0.039	4.0	0.149	0.259	
8	0.017	5.0	0.140	0.121	
9	0.007	6.0	0.132	0.050	
10	0.002	7.0	0.125	0.018	
11	0.001	8.0	0.119	0.006	

laubten θ-Bereiches erreicht. Das ist am besten in folgendem Beispiel zu sehen, das aus der Veröffentlichung von Feldmann und Cousins stammt:

Beispiel: In einem Experiment soll eine bestimmte Reaktion untersucht werden. Als Kandidaten für die Reaktion werden n Ereignisse gezählt (die vorher benutzte Variable x ist jetzt also die diskrete Variable n), die sich aus einem Signalanteil s und einem Untergrundanteil b zusammensetzen. Der Erwartungswert des Untergrundes sei zu $b = 3.0$ bestimmt. Für verschiedene Messergebnisse n sollen 90%-Vertrauensintervalle für den Erwartungswert μ des Signals ermittelt werden. Die Rate n folgt einer Poisson-Verteilung,

$$L(n|\mu) = \frac{(\mu + b)^n}{n!} \mathrm{e}^{-(\mu+b)} \tag{8.20}$$

Für die Konstruktion des Vertrauensbandes nimmt man sich in diskreten Schritten jeweils einen festen Wert $\mu \geq 0$ vor. Dann bildet man zu jedem möglichen Messergebnis n das Verhältnis

$$\lambda = \frac{L(n|\mu)}{L(n|\mu_{best})}, \tag{8.21}$$

wobei μ_{best} die beste μ-Schätzung für dieses n ist. Als Beispiel ist in Tab. 8.2 für $\mu = 0.5$ die Bestimmung der Likelihood-Ordnung gezeigt. Um ein 90%-Intervall zu erhalten addiert man die Wahrscheinlichkeiten der Ränge 1 bis 7, entsprechend $n = 0 - 6$, was 93.5% ergibt. Da die Summe bis Rang 6 nur 85.8% ergibt, entscheidet man sich für die 'konservativere' Lösung.

Wenn man diese Prozedur für den gesamten abzudeckenden μ-Bereich wiederholt hat, erhält man schliesslich die Darstellung des Vertrauensbandes in Abb. 8.4. Bei gemessenen Raten bis $n = 4$ wird das Band durch $\mu = 0$ begrenzt; deshalb würde man bei einem Messergebnis $n \leq 4$ eine obere Grenze für das Signal angeben.

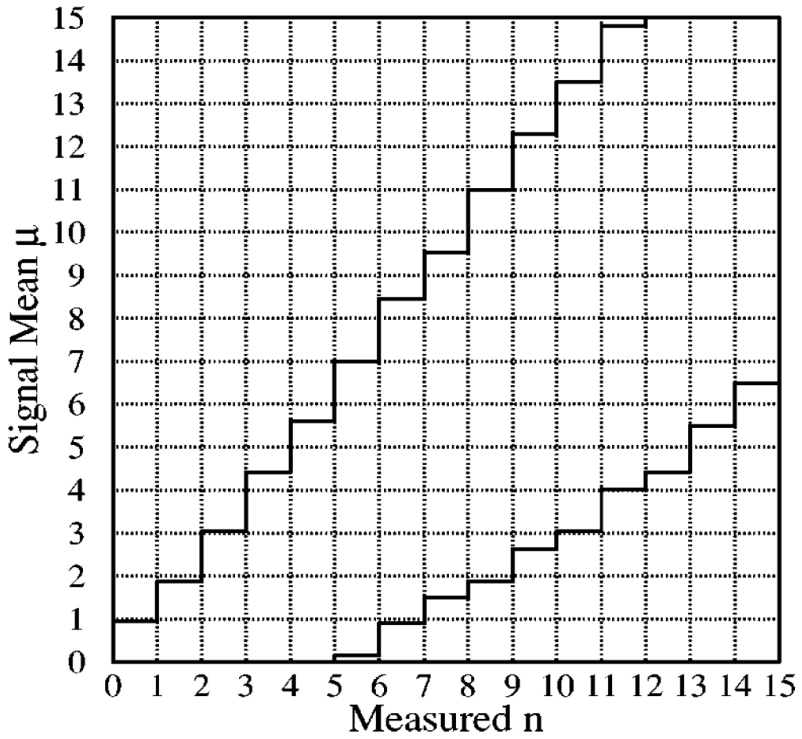


Abbildung 8.4: Vertrauensband zu 90% Vertrauensniveau für die Bestimmung einer Signalrate μ bei einem bekannten Untergrund von $b = 3.0$ (nach Feldman-Cousins).

Sensitivität: Experimentell bestimmte Ausschließungsgrenzen können wegen statistischer Fluktuationen, bei ansonsten gleichen Bedingungen, für verschiedene Experimente unterschiedlich ausfallen. Zur Beurteilung der Leistungsfähigkeit eines Experimentes ist es üblich, die 'Sensitivität' eines Experimentes auf eine Messgröße anzugeben, indem man die entsprechenden Vertrauensintervalle oder Grenzen für die Erwartungswerte angibt.

Beispiel: In dem obigen Beispiel ist die Sensitiviät für die Hypothese $\mu = 0$ bei einem Untergrund $b = 3.0$ mit 90% Vertrauensniveau durch den Erwartungswert $E(\mu_{90\%}) = \langle\mu_{90\%}\rangle$ gegeben. Wenn man den Erwartungswert von $\mu_{90\%}$ durch

$$\langle\mu_{90\%}\rangle \approx \mu_{90\%}(E(n)|\mu = 0, b = 3) = \mu_{90\%}(n = 3) \tag{8.22}$$

nähert, entnimmt man aus der Abb. 8.4 eine obere Grenze von etwa $\langle\mu_{90\%}\rangle = 4.3$.

Bei einem Beschleunigerexperiment muss man $\langle\mu_{90\%}\rangle$ durch die integrierte Luminosität L und die Akzeptanz ϵ dividieren, um den Wirkungsquerschnitt zu erhalten, den man im Mittel mit 90% Vertrauensniveau ausschliessen kann:

$$\langle\sigma_{90\%}\rangle = \frac{\langle\mu_{90\%}\rangle}{\epsilon\, L}. \tag{8.23}$$

Bei kosmischer Strahlung muss man entsprechend durch die effektive Detektorfläche A und die Zeitspanne der Datennahme T teilen, um die Sensitivität

auf einen Teilchenfluß zu bestimmen:

$$\langle \phi_{90\%} \rangle = \frac{\langle \mu_{90\%} \rangle}{AT}. \tag{8.24}$$

Kapitel 9

Klassifikation und statistisches Lernen

9.1 Einführung

In diesem Kapitel soll die Fragestellung behandelt werden, wie Ereignisse einer Stichprobe optimal in Klassen eingeteilt werden können. Beispiele für Klassifizierungsprobleme sind die Unterscheidung von Signal und Untergrund in einem Teilchenphysikexperiment (Trigger, Datenselektion), die Zuordnung von Treffern in einem Detektor zu verschiedenen Spuren, die Zuordnung von Pixeln eines Bildes zu einem Buchstaben oder einem Gesicht, die Zuordnung zu 'arm' oder 'reich' ('gesund' oder 'krank') in einer Bevölkerungsstichprobe oder die Zuordnung SPAM oder Nicht-SPAM bei E-Mails.

Formal betrachten wir Ereignisse, die gewisse Eigenschaften oder Merkmale (englisch 'features') haben, nach denen sie klassifiziert werden sollen und die wir in einem Merkmalvektor $\vec{x} = (x_1, x_2, \ldots, x_m)$ zusammenfassen. Die Klasseneinteilung wird im Allgemeinen schwieriger mit wachsender Dimension m des Merkmalraums (deshalb versucht man häufig als ersten Schritt, wenig aussagekräftige oder redundante Variable zu eliminieren). Weitere Erschwernisse ergeben sich, wenn die Ereignisklassen im Merkmalraum überlappen oder sich auf unzusammenhängende Gebiete verteilen. Häufig ist in dem Merkmalraum nicht von vornherein ein 'Abstand' zwischen verschiedenen Merkmalen definiert, so dass man zunächst eine sinnvolle Abstandsmetrik zu definieren hat, um die Merkmale vergleichbar zu machen. In der Regel werden die Merkmale zunächst aufgearbeitet, um die Klassifikation zu erleichtern. Mögliche Maßnahmen sind:

- Normierung der einzelnen Merkmale x_j auf eine Varianz 1 oder ein festes Intervall, zum Beispiel $[0, 1]$;

- Diagonalisieren der Kovarianzmatrix der Merkmale, so dass die transformierten Merkmale (Linearkombinationen der ursprünglichen) unkorreliert sind (Hauptkomponenten-Analyse, 'principle component analysis (PCA)');

- als Verallgemeinerung von PCA die Suche nach Merkmalskombinationen, die besonders signifikante Aussagen machen (Faktorenanalyse);

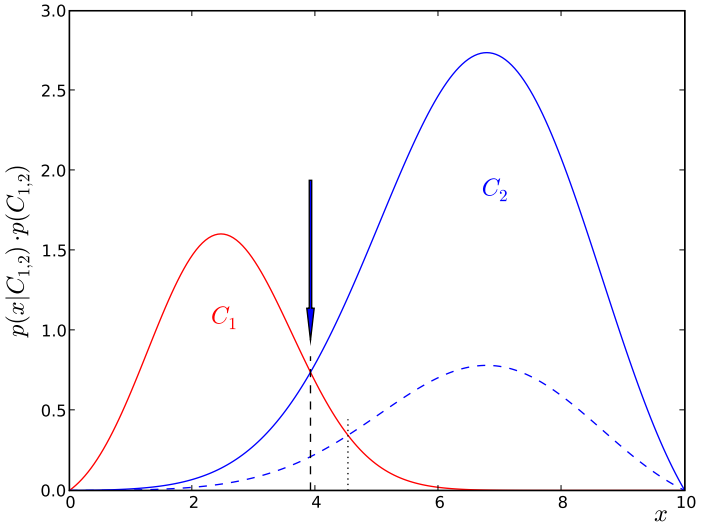


Abbildung 9.1: Wahrscheinlichkeitsdichte für das Merkmal x für zwei Klassen mit unterschiedlichen A-Priori-Wahrscheinlichkeiten. Im Fall $p(C_1) < p(C_2)$ (durchgezogenen Linien) ist die optimale Trennung bei niedrigerem x-Wert als im Fall $p(C_1) > p(C_2)$ (gestrichelte Linie für C_2).)

- Reduktion der Dimensionalität des Merkmalraumes durch Beseitigung redundanter oder unsignifikanter Information (zum Bespiel die Merkmalskombinationen mit den kleinsten Eigenwerten bei PCA).

Bayes-Diskriminante: Ein naheliegendes Klassifizierungsschema ist die Zuordnung eines Ereignisses e_i zu einer Klasse k, wenn die Wahrscheinlichkeit für die Klasse C_k (entsprechend einer 'Hypothese' im vorigen Kapitel) bei gegebenen Merkmalen $\vec{x}_i$ größer ist als für alle anderen Klassen:

$$e_i \to C_k \iff p(C_k|\vec{x}_i) > p(C_j|\vec{x}_i) \quad \forall\, j \neq k. \tag{9.1}$$

Die Wahrscheinlichkeit für eine Klasse ergibt sich wieder aus dem Bayes-Theorem (1.18):

$$p(C_k|\vec{x}_i) = \frac{p(\vec{x}_i|C_k) \cdot p(C_k)}{\sum_{j=1}^{n} p(\vec{x}_i|C_j) \cdot p(C_j)} \tag{9.2}$$

Das Klassifizierungsschema ist anschaulich in Abb. 9.1 anhand nur eines Merkmals x dargestellt: das Merkmal tritt in den zwei betrachteten Klassen normalverteilt mit unterschiedlichen Mittelwerten und Breiten auf. Die Normierungen entsprechen den A-priori-Wahrscheinlichkeiten für die Klassen ($p(C_1)$, $p(C_2)$), die in der Abbildung mit zwei unterschiedlichen Verhältnissen angenommen sind. Die Trennung der beiden Klassen nach (9.1) ergibt sich, wo sich die beiden Kurven schneiden.

Das ist natürlich ein besonders einfaches Beispiel, insbesondere wollen wir im Folgenden multi-dimensionale Merkmalsräume betrachten ('multivariate analysis'). In multi-dimensionalen Räumen werden die Klassen durch Hyperflächen getrennt, die durch (9.1) festgelegt werden. Im einfachsten Fall ist die Fläche eine lineare

Funktion, im allgemeinen eine komplizierte Funktion der Merkmale, eventuelle auch nicht zusammenhängend.

Training: Im Allgemeinen werden die Wahrscheinlichkeitsdichten (9.2), auf deren Basis die Klassentrennung erfolgt, nicht bekannt sein. Mit wachsender Dimensionalität wird es auch immer schwieriger, diese Wahrscheinlichkeitsdichten aus Simulationen zu konstruieren, weil zunehmend weniger Ereignisse in ein diskretes Bin fallen. Es sind deshalb Algorithmen entwickelt worden, die Klassentrennung mit Hilfe von Trainigsdatensätzen "lernen" können. Trainiert wird mit simulierten oder auch realen Daten auf eine Ausgabegröße des Algorithmus, die ein Maß für die Zugehörigkeit zu einer Klasse ist. Zum Beispiel kann bei zwei disjunkten Klassen die Ausgabegröße 0 oder 1 sein je nachdem, ob der Merkmalvektor in die Klasse 1 oder 2 gehört. Bei sich überlappenden Verteilungen kann die Ausgabe eine kontinuierliche Zahl sein, die ein Maß für die Wahrscheinlichkeit für eine Klassenzugehörigkeit ist. Das Trainingsergebnis wird mit einem unabhängigen Datensatz getestet, um damit Effizienz und Reinheit der Klassenzuordnung zu bestimmen.

9.2 Schätzung von Wahrscheinlichkeitsdichten

Das Trennungskriterium (9.1) kann man direkt anwenden, wenn man die Wahrscheinlichkeiten $p(C_k|\vec{x})$ in (9.2) als Funktion der Merkmale $\vec{x}$ numerisch zur Verfügung hat. Häufig muss man sich die Wahrscheinlichkeiten aus Simulationen beschaffen. Dazu simuliert man Ereignisse entsprechend der Wahrscheinlichkeitsdichte $p(\vec{x}_i|C_k)$ für jede Klasse C_k und wendet dann das Bayes-Theorem mit den relativen Häufigkeiten der C_k an, um $p(C_k|\vec{x})$ zu bestimmen.

Es gibt verschiedene Methoden aus simulierten, diskreten Ereignissen die Wahrscheinlichdichte zu schätzen:

- Falls eine parametrisierte Modellfunktion bekannt ist, können mit den MC-Ereignissen die Parameter, zum Beispiel durch ML-Anpassung, bestimmt werden.

- Als Modellfunktion kann man auch eine Linearkombination von orthogonalen Funktionen benutzen, zum Beispiel Wavelets.

- Die Dichte wird an jedem Punkt durch Mittelung der Ereignisse über ein Nachbarschaftsvolumen mit vorgebbarer Größe bestimmt.

- Bei der Mittelung kann man die nahen Ereignisse mehr wichten als die weiter entfernten, zum Beispiel durch eine Gauss-Funktion. Die Wichtungsfunktion nennt man 'Kernfunktion' ('kernel funktion') und die Methode 'Kernel Probability Density Estimation (kernel PDE)'.

Im Folgenden wird beispielhaft nur die letzte Methode besprochen.

'Kernel Probability Density Estimation': Gegeben sei eine Stichprobe $\vec{x}_i$ $(i = 1, ..., N)$. Die Wahrscheinlichkeitsdichte an einem Punkt $\vec{x}$ wird abgeschätzt durch:

$$\hat{p}(\vec{x}) = \frac{1}{Nh^m} \sum_{i=1}^{N} K\left(\frac{\vec{x} - \vec{x}_i}{h}\right). \tag{9.3}$$

Dabei ist K die Kern-Funktion, h ein Parameter, der die Reichweite der Mittelung bestimmt, und m ist die Dimension von $\vec{x}$. Der Reichweiteparameter h muss so gewählt werden, dass genügend Ereignisse in der Nachbarschaft liegen. Als mögliche Wahl findet sich zum Beispiel in der Literatur $h = N^{-1/(m+4)}$ (man beachte, dass $V \cdot N^{-1/m}$ der mittlere Abstand zwischen zwei Ereignissen in dem m-dimensionalen Volumen V ist).

Gauss-Kern: Wenn die Kern-Funktion eine Gauss-Funktion ist, kann man auch mögliche Korrelationen der Merkmale mit deren Kovarianzmatrix V einbeziehen, wobei V aus der Simulation geschätzt wird, entweder global für den ganzen Datensatz oder lokal um $\vec{x}$ für die Ereignisse, die wesentlich zu $\hat{p}(\vec{x})$ beitragen. Die Formel für die geschätzte Wahrscheinlichkeitsdichte lautet für den Gauss-Kern:

$$\hat{p}(\vec{x}) = \frac{1}{N\sqrt{2\pi \det V}\, h^m} \sum_{i=1}^{N} \exp\left(\frac{(\vec{x} - \vec{x}_i)^T V^{-1} (\vec{x} - \vec{x}_i)}{2h^2}\right). \tag{9.4}$$

9.3 Lineare Diskriminanten

9.3.1 Klassentrennung durch Hyperebenen

Ein Trennungskriterium wie (9.1) definiert Hyperflächen im Merkmalsraum, die in die verschiedenen Klassen aufteilen. Im einfachsten Fall sind diese Flächen Hyperebenen, die zwei Klassen trennen. Die Hessesche Normalform einer Ebene ist:

$$\vec{n}(\vec{x} - \vec{x}_0) = 0, \tag{9.5}$$

wobei $\vec{n}$ der Normalenvektor auf der Ebene, $\vec{x}$ einen beliebigen Punkt und $\vec{x}_0$ einen bestimmten Punkt auf der Ebene beschreibt (der Differenzvektor $\vec{x} - \vec{x}_0$ liegt in der Ebene, siehe Abb. 9.17).

Wenn der Punkt mit dem Ortsvektor $\vec{x}$ nicht auf der Ebene liegt, ist die Gleichung (9.5) nicht erfüllt und es ergibt sich:

$$\vec{n}(\vec{x} - \vec{x}_0) = d \quad \text{mit} \quad d > 0 \text{ oder } d < 0, \tag{9.6}$$

wobei d der Abstand des durch $\vec{x}$ gegebenen Punktes von der Ebene ist und das Vorzeichen die beiden Hemisphären kennzeichnet. Insbesondere ergibt sich für $\vec{x} = 0$ aus $\vec{n}\vec{x}_0 = -d_0$ der Abstand der Ebene vom Ursprung (mit dem durch die Ebenenorientierung festgelegten Vorzeichen).

Im Folgenden wird ein Festlegung der Ebene eingeführt, die eine optimale Trennung zwischen zwei Klassen ergeben, wenn sich deren Verteilungen im Merkmalsraum annähernd durch Normalverteilungen beschreiben lassen.

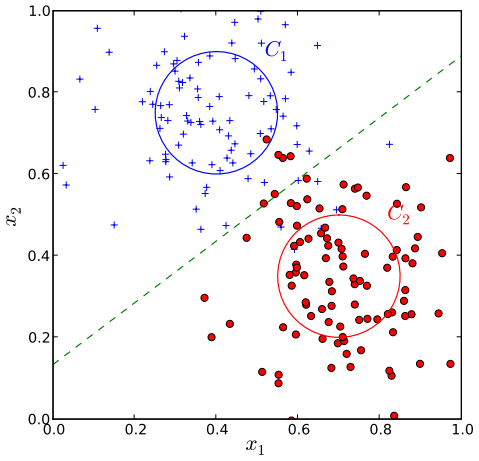


Abbildung 9.2: Stichprobe von Ereignissen mit Merkmalen (x_1, x_2), die aus zwei Klassen gezogen wurden (Kreuze und Kreise). Die Klassenzuordnung kennt man nur für die Trainings- und Testdatensätze. Die Linie zwischen den beiden Anhäufungen ist die Fisher-Diskriminante, die beide Klassen optimal trennt.

9.3.2 Fisher-Diskriminante

Gegeben sei eine Stichprobe von Ereignissen, die zwei Klassen C_1 und C_2 entnommen sind und jeweils durch einen Merkmalvektor $\vec{x}$ gekennzeichnet sind (Abb. 9.2). Die Wahrscheinlichkeitsdichten der Merkmale seinen $f(\vec{x}|C_1)$ und $f(\vec{x}|C_2)$. Wir bilden nun aus einer Linearkombination der Komponenten von $\vec{x}$ eine Testfunktion:

$$t(\vec{x}) = \sum_{j=1}^{m} a_j x_j = \vec{a}^T \vec{x} \tag{9.7}$$

Diese Testfunktion hat unterschiedliche Wahrscheinlichkeitsdichten für die beiden Klassen, die sich durch die Projektion der Ereignisse auf eine Achse senkrecht zur Ebene ergeben (das ist die t-Achse):

$$g(t|C_k), \;\; k = 1, 2. \tag{9.8}$$

Der Koeffizientenvektors $\vec{a}$ soll nun so bestimmt werden, dass die beiden Wahrscheinlichkeitsdichten (9.8) möglichst optimal getrennt sind. Man kann die Testfunktion so interpretieren, dass der Vektor $\vec{a}$ die Orientierung einer Ebene definiert und für den Ortsvektor $\vec{x}_0$ eines Punktes in der Ebene gibt $t(\vec{x}_0)$ den Abstand vom Ursprung an (siehe (9.6)). Durch Anpassung des Koeffizientenvektors $\vec{a}$ und durch Festlegung eines kritischen Wertes t_c der Testfunktion soll nun eine optimale Trennung zwischen zwei Klassen C_1 und C_2 erreicht werden.

Dazu berechnen wir die Erwartungswerte der $\vec{x}$ und die Kovarianzmatrizen für beide Klassen getrennt:

$$\vec{\mu}^{(k)} = \int \vec{x} f(\vec{x}|C_k)\, dx_1 \ldots dx_m, \quad k = 1, 2; \tag{9.9}$$

$$V_{ij}^{(k)} = \int (x_i - \mu_i^{(k)})(x_j - \mu_j^{(k)}) f(\vec{x}|C_k)\, dx_1 \ldots dx_m, \quad k = 1, 2; \;\; i, j = 1, \ldots, m \tag{9.10}$$

In der Regel werden diese Erwartungswerte mit Hilfe von simulierten Datensätzen für beide Klassen geschätzt ('gelernt').

Wegen der linearen Abhängigkeit von t von den Merkmalen, sind die Erwartungswerte von t und deren Varianzen für die beiden Klassen einfach zu berechnen:

$$t_k = \int t\, g(t|C_k)\, dt = \vec{a}^T \vec{\mu}^{(k)} \tag{9.11}$$

$$\sigma_k = \int (t - t_k)^2\, g(t|C_k)\, dt = \vec{a}^T V^{(k)} \vec{a} \tag{9.12}$$

Die Trennungsebene soll jetzt durch Wahl von $\vec{a}$ so gelegt werden, dass der Abstand $|t_1 - t_2|$ möglichst groß wird und die t-Werte möglichst dicht um die Erwartungswerte konzentriert sind, was durch die Varianzen der t_k gegeben ist. Durch Maximierung des χ^2-artigen Ausdrucks

$$J(\vec{a}) = \frac{(t_1 - t_2)^2}{\sigma_1^2 + \sigma_2^2} = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}} \tag{9.13}$$

in Bezug auf $\vec{a}$ ergibt sich die optimale Trennung. Die Matrix B auf der rechten Seite von (9.13) ist die Kovarianzmatrix von $\vec{\mu}^{(1)} - \vec{\mu}^{(2)}$,

$$(t_1 - t_2)^2 = \sum_{i,j=1}^{m} a_i a_j (\mu^{(1)} - \mu^{(2)})_i (\mu^{(1)} - \mu^{(2)})_j = \sum_{i,j=1}^{m} a_i a_j B_{ij} = \vec{a}^T B \vec{a}, \tag{9.14}$$

und die Matrix $W = V^{(1)} + V^{(2)}$, die Summe der Kovarianzmatrizen der beiden Klassen, ergibt sich aus

$$\sigma_1^2 + \sigma_2^2 = \sum_{i,j=1}^{m} a_i a_j (V^{(1)} + V^{(2)})_{ij} = \vec{a}^T W \vec{a}. \tag{9.15}$$

Die Maximierung von $J(\vec{a})$ legt $\vec{a}$ bis auf einen Skalenfaktor fest:

$$\vec{a} \sim W^{-1}(\vec{\mu}^{(1)} - \vec{\mu}^{(2)}) \tag{9.16}$$

Die rechte Seite der Gleichung kann aus Simulationen bestimmt werden. Für die Trennung der beiden Klassen muß noch ein kritischer Wert t_c der Testfunktion festgelegt werden, so dass die Klassenzugehörigkeit nach $t < t_c$ oder $t > t_c$ entschieden wird. Das Kriterium für die Wahl von t_c sind Effizienz und Reinheit der klassifizierten Ereignisse.

9.4 Neuronale Netze zur Datenklassifikation

9.4.1 Einleitung: Neuronale Modelle

Die Entwicklung der Neuroinformatik hat seit Beginn der 80er Jahre einen großen Aufschwung erfahren. Der wesentliche Grund dafür ist sicherlich die große Leistungssteigerung bei den Computern. Damit wurden Computersimulationen von komplexeren Gehirnmodellen und künstlichen neuronalen Netzen (KNN) erst möglich. Dagegen gehen die ersten aussagekräftigen Theorien über die Informationsverarbeitung im Gehirn und den Nervenzellen bis in die 40er Jahre zurück.

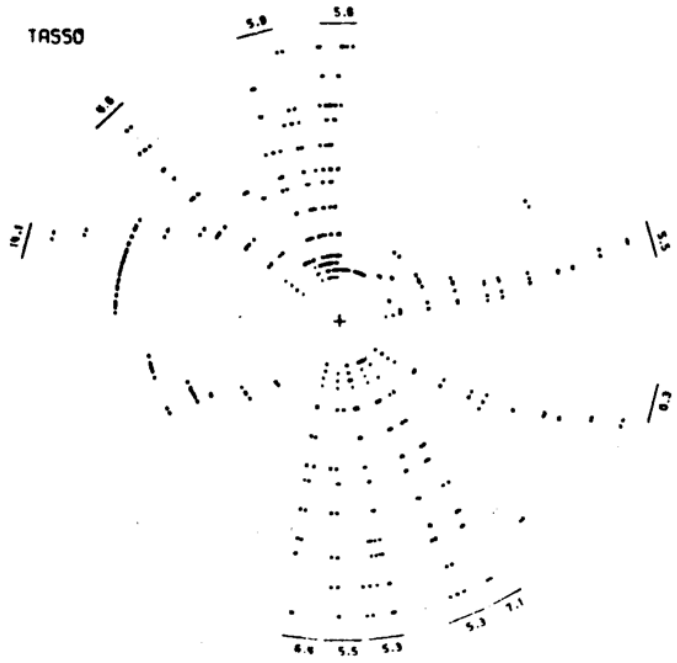


Abbildung 9.3: Hit-Muster, die von Teilchenspuren in einer Driftkammer (TASSO-Experiment) hinterlassen wurden.

Es ist offensichtlich, dass von-Neumann-Computer bei kognitiven Aufgaben (Hören, Sehen, Mustererkennen, etc.) und bei unvollständiger, inkonsistenter oder verrauschter Information im Vergleich zum Gehirn versagen. Das Hit-Muster, das zum Beispiel Teilchenspuren in einer Driftkammer hinterlassen (Abb. 9.3), hat unser Auge 'momentan', innerhalb O(0.1s), als stetig aufeinanderfolgende Punkte erkannt und miteinander verbunden. Der Zeitbedarf eines Computers ist nur dank seiner sehr viel größeren Geschwindigkeit pro einzelnem Rechenschritt vergleichbar. Mit künstlichen neuronalen Netzen könnte dieselbe Leistung innerhalb von O(μs) erzielt werden.

Gehirn-Architektur: Die charakteristischen Merkmale der Datenverarbeitung im Gehirn machen den Unterschied zu dem heutigen Standard für Computerarchitekturen klar:

- sehr viele parallele Prozessoren, O(10^{11}), insgesamt kompakt, geringer Energieverbrauch;
- langsame Einzelschritte, O(ms);
- massiv parallele Verarbeitung (O(10^{13}) Synapsen);
- keine Hardware-Software-, Algorithmen-Daten-Trennung;
- lernfähig:
  - evolutionäres, dynamisches Lernen gibt hohe Flexibilität für die Informationsverarbeitung,

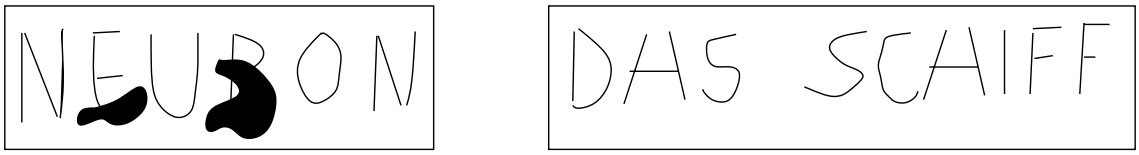


Abbildung 9.4: Beispiele für Fehlertoleranz und Ausgleich von Ungenauigkeiten im Gehirn: auf der linken Seite ist die Information verstümmelt; rechts wird exakt das gleiche Symbol einmal als 'A' und dann als 'H' im Zusammenhang richtig erkannt.

  - evolutionäre Selbstorganisation gibt dem Netz eine gewisse Plastizität zur Anpassung an Neues;

- fehlertolerant (Abb. 9.4), Information kann bis zu einem gewissen Grade
  - unvollständig,
  - inkonsistent,
  - verrauscht sein;
- Stärke: schnelle Erfassung komplexer Zusammenhänge, kognitive Aufgaben, Mustererkennung, assoziative Verknüpfungen.

Literatur zu Neuronalen Netzen: Einführende Literatur zu neuronalen Netzen findet man unter [5, 6, 7, 8, 9, 10, 11, 12]. Siehe auch Spektrum der Wissenschaft, Nov. 79 und Nov. 92, beide Hefte sind dem Gehirn gewidmet [13, 14].

9.4.2 Natürliche neuronale Netze

Die intellektuellen Leistungen werden in der Hirnrinde (Neokortex) erzielt (Fläche etwa 0.2 m^2, Dicke 2-3 mm). Die Hirnrinde ist in Felder für verschiedene Teilaufgaben organisiert (zum Beispiel visuelle, motorische, somatosensorische, Assoziations-Felder).

Ein Schnitt durch die Hirnrinde zeigt ein vertikal ausgerichtetes Netz von Neuronen (Nervenzellen) mit ihren Verzweigungen (Abb. 9.5). In einer vertikalen Säule von 1 mm^2 befinden sich etwa 10^5 Neuronen, insgesamt gibt es etwa 10^{11} Neuronen im Gehirn.

Aufbau und Funktion der Neuronen:

Es gibt viele unterschiedliche Neuron-Typen. Um die uns interessierenden wesentlichen Eigenschaften von Neuronen zu beleuchten, konzentrieren wir uns auf die schematische Darstellung eines typischen Neurons in Abb. 9.6. Solch ein Neuron besteht aus

- dem Zellkörper, Durchmesser 5-80 μm,
- den Dendriten, die sich zu Dendritenbäumen mit einer Reichweite von 0.01-3 mm verzweigen,

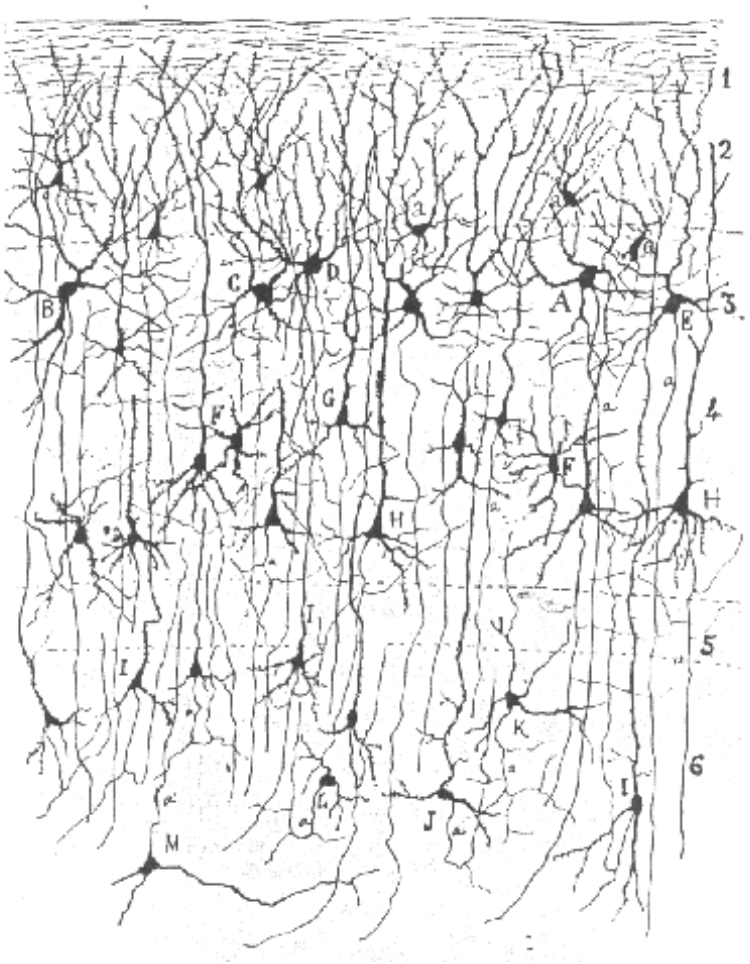


Abbildung 9.5: Vertikaler Schnitt durch die Hirnrinde. Die Dichte der Neuronen ist um einen Faktor 100 untersetzt

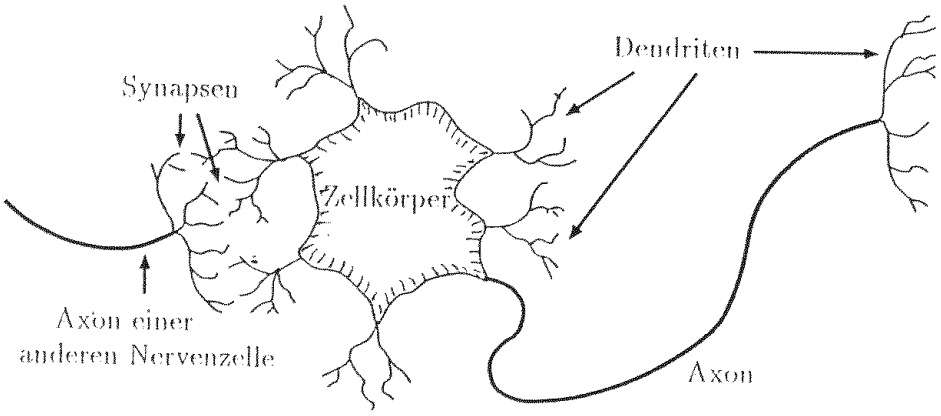


Abbildung 9.6: Schematische Darstellung eines Neurons.

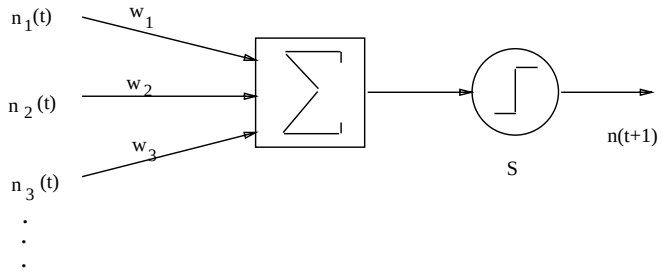


Abbildung 9.7: Neuron als logisches Schaltelement

- den Axons, die bis zu 1 m lang sein können.

Funktionsweise eines Neurons:

- Die Dendriten sammeln in einer Umgebung bis zu etwa 400 μm Signale von benachbarten Neuronen oder von den Axonen weiter entfernter Neuronen.
- Die Signalübertragung auf die Dendriten oder direkt auf den Zellkörper erfolgt über chemische Kontakte (Neurotransmitter) an den Synapsen innerhalb von O(1 ms). In der Hirnrinde hat jedes Neuron O(10^3) Synapsen (allgemein im Gehirn O(1) bis O(10^5)). Die Zeitskala für die Übertragung ist 1 ms, d.h. dass zum Beispiel die visuelle Erkennung eines Bildes mit nicht mehr als O(10) seriellen Schritten erfolgen muß.
- Das Summensignal aller Dendriten verändert das elektrische Potential des Zellkörpers.
- Bei Überschreiten einer Schwelle erzeugt diese Potentialänderung einen Nadelpuls (Spike) auf dem Axon (Signalgeschwindigkeit etwa 10 m/s).

Einfaches Modell: das McCulloch-Pitts-Neuron: Abbildung 9.7 zeigt das McCulloch-Pitts-Neuron, das einem logischen Schaltelement entspricht. Die binären Eingangssignale n_i erzeugen ein binäres Ausgangssignal n (n_i, $n = 0$ oder 1) nach der Vorschrift:

$$n(t+1) \;=\; \Theta\left(\sum_j w_j n_j(t) - s\right) \tag{9.17}$$

Dabei ist t eine diskrete Zeitvariable. Die Heaviside-Funktion ist definiert als:

$$\Theta(x) = \begin{array}{ll} 1 & x \geq 0 \\ 0 & sonst \end{array}$$

Die Gewichte w_i entsprechen den Synapsenstärken, s ist der Schwellenwert. Das Neuron 'feuert' also, wenn die gewichtete Summe der Eingangssignale die Schwelle s überschreitet. Die Gewichte können > 0 (erregend) oder < 0 (hemmend) sein, wie es auch tatsächlich für Synapsen beobachtet wird.

Neuronale Vernetzung: Wesentlich für die Funktion des Gehirns ist das kollektive Verhalten eines Systems von nichtlinear gekoppelten Neuronen. Im Beispiel Abb. 9.8 werden die Eingangsreize x_i (zum Beispiel visuelle Signale) in Ausgangssignale y_i (zum Beispiel zur Bewegung eines Muskels) transformiert.

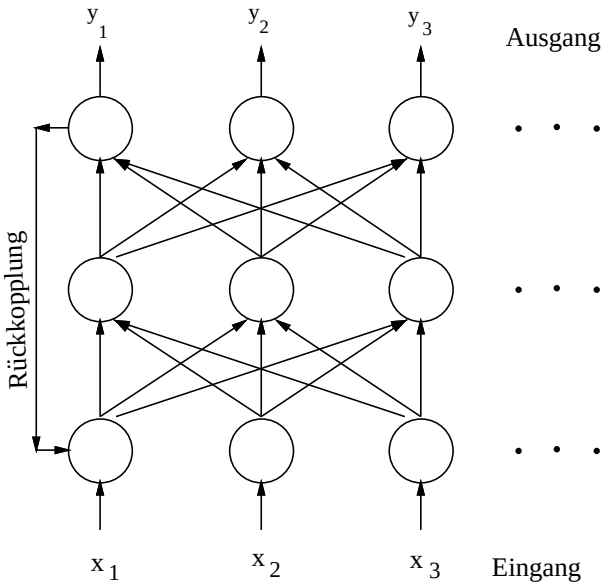


Abbildung 9.8: Beispiel für ein neuronales Netz.

Lernen und Selbstorganisation:

Aus eigener Erfahrung wissen wir, dass das Gedächtnis auf unterschiedlichen Zeitskalen arbeitet. Manches ist bereits nach Sekunden verpflogen, wie die dauernd einwirkenden sensorischen Reize, anderes behalten wir für Minuten oder Tage oder Jahre. Das Behalten im Gedächtnis ist also ähnlich einem evolutionärem Prozess. Generell scheint zu gelten, dass die Stärke und Häufigkeit eines Reizes das Lernen wesentlich beeinflußt. Man beachte, dass wir zum Lernen offensichtlich in der Regel nicht zu wissen brauchen, ob das Gelernte richtig ist ('Lernen ohne Lehrer').

Auf diese Beobachtungen ist die Lernregel von Hebb begründet: Die Synapsenstärke ändert sich proportional zu der Korrelation zwischen prä- und postsynaptischem Signal:

$$\Delta w_i = \eta \cdot y(x_i) \cdot x_i, \text{ mit } 0 < \eta < 1 \tag{9.18}$$

Der Lernparameter η legt die Lerngeschwingigkeit fest. Es ist ein besonders empfindlicher Parameter: einerseits möchte man schnell lernen, andererseits birgt zu schnelles Lernen die Gefahr, dass zuviel Unsinn abgespeichert wird.

Strukturbildung: Mit den etwa 10^{13} Synapsen ergeben sich etwa $10^{10^{14}}$ mögliche Konfigurationen des Gehirns. Das kann nicht alles genetisch festgelegt sein! Genetisch kodiert sind wahrscheinlich nur Organisationsschemata und ein Strukturbildungsmechanismus. Die Verbindungen zwischen den Neuronen werden zum Teil evolutionär aufgrund sensorischer Reize gebildet und können meistens auch später noch verändert werden.

Topographische Abbildungen: Der Lernvorgang führt offensichtlich zu Strukturen im Gehirn, die vorgegebene topographische Zusammenhänge bei den einlaufenden Sinnesreizen intakt lassen. Beispielsweise wird im somatosensorischen Kortex

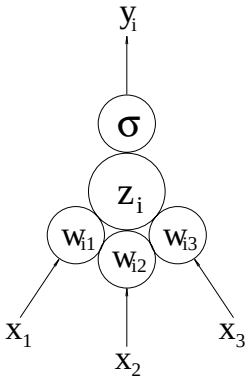


Abbildung 9.9: Struktur eines künstlichen Neurons

der Tastsinn der Hautoberfläche so abgebildet, dass benachbarte Körperbereiche benachbart bleiben. Eine wesentliche Eigenschaft der Abbildung ist die Anpassung der Größe der Bildbereiche entsprechend der Wichtigkeit und das jeweils benötigte Auflösungsvermögen.

9.4.3 Künstliche neuronale Netze (KNN)

Künstliche neuronale Netze und neuronale Algorithmen sind in den letzten Jahren intensiv theoretisch untersucht, auf Computern simuliert und – seltener – als Hardware realisiert worden. Bei der Entwicklung von NN-Modellen wird man sich natürlich von den biologischen Befunden inspirieren lassen. Für die Anwendung ist es aber nicht wichtig, ob ein Modell tatsächlich in der Natur realisiert wird. Hier ist der praktische Erfolg ausschlaggebend.

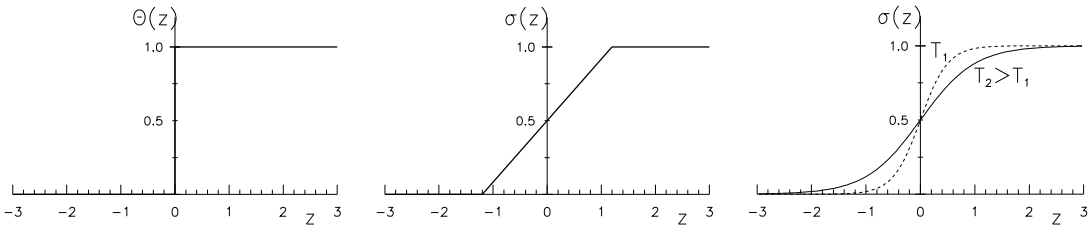
Ausgehend von den im vorigen Abschnitt entwickelten Vorstellungen über natürliche neuronale Netze definieren wir im folgenden, welches die gemeinsamen Elemente der KNN-Modelle sein sollen. Diese Aufstellung ist nicht strikt, sondern soll eher eine Orientierung sein.

- Prozessorelement: (formales) Neuron, Netzwerk-Knoten (Abb. 9.9).
- Eingabeaktivitäten x_j (Signale auf den Dendriten) sind reelle Zahlen (oder Spannungen, Ströme), eventuell binär (-1,1) oder (0,1).
- Gewichte (entspricht den Synapsen) w_{ij}, > 0 (erregend), < 0 (hemmend)
- Aktivitätsfunktion, zum Beispiel:

$$z_i = \sum_j w_{ij} x_j - s_i$$

- Ausgabefunktion (oder Transferfunktion) g:

$$y_i = g(z_i)$$



Stufe linear sigmoid

Abbildung 9.10: Beispiele von Schwellenfunktionen

I.a. liegt y_i im Intervall [-1,1] oder [0,1] und hat häufig ein Schwellwertverhalten mit Sättigung an den Intervallgrenzen. Neben der Θ-Funktion werden häufig folgende 'sigmoide' Funktionen gewählt (Abb. 9.10):

$$\sigma(z) = \frac{1}{1+e^{-z/T}} \tag{9.19}$$

$$\sigma(z) = \tanh(z/T) \tag{9.20}$$

$$\sigma(z) = 1/2(1+\tanh(z/T)) \tag{9.21}$$

Die Funktionen (9.19) und (9.21) haben Werte im Intervall [0,1] und die Funktion (9.20) im Intervall [-1,1]. Sigmoide Funktionen haben den Vorteil im Bereich der Schwelle differenzierbar zu sein. Die 'Temperatur' T bestimmt den Bereich mit variabler Verstärkung:

Für T→ 0 geht σ in die Θ-Funktion über (binäres Neuron).

T groß: weiche Entscheidung.

- Netzwerk-Architektur: Netzwerk mit Knoten und Verbindungen
  - 'jeder mit jedem'
  - Nachbarschaftsverknüpfung
  - uni- oder bi-direktional
  - Schicht-Struktur mit hierarchischer Anordnung (zum Beispiel feed-forward)
  - mit oder ohne Rückkopplung
  - ...
- Lernen:
  - Anpassung der Gewichte
  - Anpassung der Architektur: Erzeugen und Löschen von Neuronen und Verbindungen
- Lernregel:
  - selbständig (ohne Lehrer, unsupervised), zum Beispiel Hebb-Regel

  - angeleitet (mit Lehrer, supervised) Vergleich des Netzwerk-Outputs mit der (vom Lehrer vorgegebenen) Erwartung, Anpassung durch Fehlerminimierung (zum Beispiel Backpropagation- Algorithmus).

- Update-Regel: Neubestimmung eines Netzzustandes kann synchron, sequentiell oder iterativ (wegen nichtlinearer Kopplungen) gemacht werden.
- Netzwerk-Phasen:
  - Trainingsphase (Verwendung eines Trainings-Datensatzes)
  - Generalisierungsphase (Anwendung auf unbekannte Daten)

Feed-Forward-Netzwerke

In dieser Vorlesung wollen wir uns auf sogenannte Feed-Forward-Netzwerke beschränken, in denen die Neuronen geschichtet angeordnet sind und die Verbindungen streng nur in eine Richtung, jeweils zur nächsthöheren Schicht, von der Eingabeschicht bis zur Ausgabeschicht laufen (Abb. 9.8, ohne Rückkopplung). Feed-Forward-Netze (FFN) werden häufig zur

- Lösung von Klassifikationsaufgaben,
- Mustererkennung und
- Funktionsapproximation

benutzt. Für praktische Anwendungen sind sie wahrscheinlich der wichtigste Netzwerktyp. Ihre Bedeutung haben FFN wohl durch die von herkömmlichen Computern gut ausführbaren, im Prinzip sequentiellen, Algorithmen und insbesondere die Backpropagation-Lernvorschrift erhalten.

Das einfachste Beispiel ist das (einfache) Perzeptron mit nur einer Eingangsschicht und einer Ausgangsschicht. Mit Computersimulationen konnte gezeigt werden, dass ein Perzeptron 'intelligenter' Leistungen fähig ist: Es kann angebotene Muster unterscheiden und kann diese Musterklassifizierung mit Hilfe eines Lehrers lernen (supervised learning).

9.4.4 Das einfache Perzeptron

Definition und Eigenschaften des Perzeptrons:

Abbildung 9.11 zeigt das einfache Perzeptron mit einer Eingangsschicht (oder -lage) und einer Ausgangsschicht (wir ordnen den Eingängen eine Schicht zu, ist manchmal auch anders definiert). Jeder der k Eingänge ist mit jedem der l Ausgänge verbunden, den Verbindungen werden die Gewichte w_{ij} $(i = 1, \ldots, k;\ j = 1, \ldots, l)$ zugeordnet. Die Eingänge $x_1,\ x_2, \ldots, x_k$ lassen sich in einem 'Mustervektor' $\vec{x}$ zusammenfassen, der einen Punkt im 'Musterraum' (pattern space) darstellt. Die einzelnen Komponenten sind 'Merkmale' (features). Über die folgende Vorschrift wird einem Mustervektor $\vec{x}$ ein Ausgabevektor $\vec{y}$ zugeordnet:

$$y_i = g\left(\sum_j w_{ij} x_j\right) = g(\vec{w}_i \vec{x}) \tag{9.22}$$

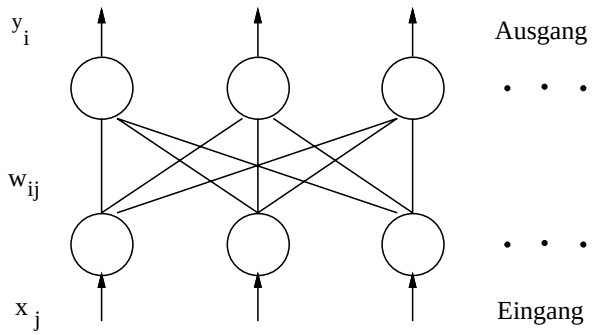


Abbildung 9.11: Perzeptron-Netzwerk

Im letzten Teil wurden die Gewichte zu einem Ausgangsknoten i zu einem Vektor zusammengefaßt. Die Transferfunktion g ist gewöhnlich eine sigmoide Funktion (ursprünglich beim Perzeptron einfach die Θ-Funktion, wir wollen uns hier nicht darauf beschränken). In Gl. (9.22) kommen keine expliziten Schwellen s_i vor wie in der Formel (9.17) für das McCulloch-Pitts-Neuron. Schwellen können durch eine zusätzliche konstante Eingabe $x_0 = 1$ und die Gewichte $w_{i0} = -s_i$ berücksichtigt werden.

Beispiel: Darstellung der Boolschen Funktionen AND und OR: Wir wollen hier binäre Ein-und Ausgabegrößen betrachten mit Werten 0 und 1. Dann muß die Transferfunktion die Θ-Funktion sein, $g = \Theta$. Im folgenden wird gezeigt, dass sich die Funktionen AND und OR entsprechend der Wahrheitstafel in Abb. 9.12 durch ein Netz mit den 2 Eingängen x_1 und x_2 und einem Ausgang y realisieren lassen ('Ja-Nein-Maschine').

Wir wollen an dieser Stelle zunächst nicht der Frage nachgehen, wie das Netz die richtigen Antworten lernt; das wird dann allgemeiner für mehrschichtige FFN gezeigt (siehe Abschnitt 9.4.6). Man kann sich aber leicht davon überzeugen, dass die Gewichte

$$\text{AND}: \ (\mathrm{w}_0, \mathrm{w}_1, \mathrm{w}_2) \ = \ (-1.5, 1.0, 1.0)$$

$$\text{OR} \ : \ (\mathrm{w}_0, \mathrm{w}_1, \mathrm{w}_2) \ = \ (-0.5, 1.0, 1.0)$$

das Problem lösen (Abb. 9.12). Die Bedeutung dieses Resultates ist sehr anschaulich: Nach Gl. (9.22) wird der Raum der Muster (x_1, x_2) in 2 Klassen geteilt, die der Bedingung

$$\vec{w}\vec{x} \ < \ 0 \text{ bzw. } \vec{w}\vec{x} \ < \ 0$$

genügen. Die Trennung zwischen beiden Klassen

$$\vec{w}\vec{x} \ = \ 0$$

definiert eine Hyperebene im Musterraum, auf der der Vektor $\vec{w}$ senkrecht steht. In unserem Fall sind die Hyperebenen Geraden, die folgenden Gleichungen genügen:

$$\text{AND}: x_1 + x_2 = 1.5$$

$$\text{OR} \ : x_1 + x_2 = 0.5$$

x_1	x_2	$y(AND)$	$y(OR)$	$w_1x_1 + w_2x_2$
0	0	0	0	0
1	0	0	1	1
0	1	0	1	1
1	1	1	1	2

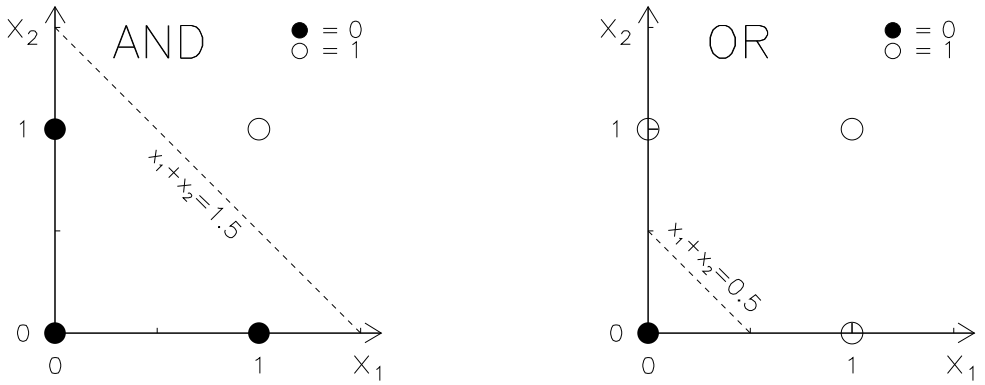


Abbildung 9.12: Oben: Wahrheitstafel für die Boolschen Funktionen AND und OR zusammen mit der Summe der gewichteten Eingänge wie vom Perzeptron berechnet. Unten: Klasseneinteilung im Musterraum für das AND- und OR-Problem. Die gestrichelten Geraden geben die von dem Perzeptron jeweils gefundene Klassentrennung an.

x_1	x_2	$y(XOR)$	$\vec{w}\vec{x}$
0	0	0	$w_0 < 0$
1	0	1	$w_0 + w_1 > 0$
0	1	1	$w_0 + w_2 > 0$
1	1	0	$w_0 + w_1 + w_2 < 0$

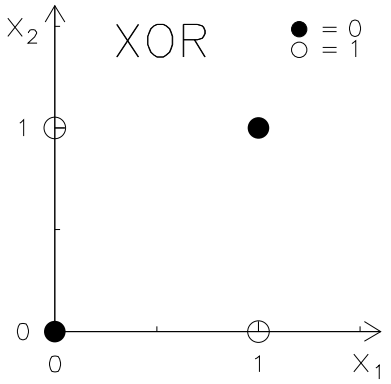


Abbildung 9.13: Links: Wahrheitstafel für die Boolschen Funktionen XOR zusammen mit den Bedingungen an die Gewichte. Rechts: Klasseneinteilung im Musterraum für das XOR-Problem.

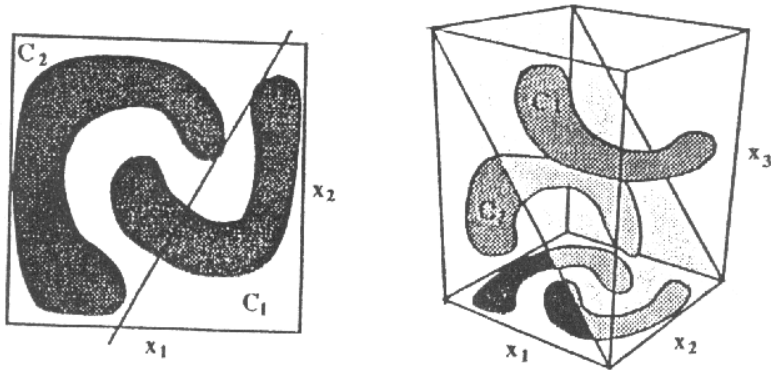


Abbildung 9.14: Lineare Separierbarkeit: a) in 2 Dimensionen nicht separierbar, b) in 3 Dimensionen separierbar.

Abbildung 9.12 zeigt die Lage der Geraden in dem Musterraum.

Allgemein gilt, dass durch Gl. (9.22) für jeden Ausgabeknoten eines Perzeptrons eine Hyperebene definiert wird, die jeweils den Musterraum in zwei Klassen einteilt. Die Trennung ist scharf für $g = \Theta$, was für eine Klasse $y = 0$ und für die andere $y = 1$ liefert. Bei einer sigmoiden Funktion ist die Ausgangsaktivität y ein (im Allgemeinen nichtlineares) Maß für den Abstand von der Hyperebene, solange man sich noch so nahe an der Hyperebene befindet, dass g noch nicht in Sättigung ist.

Limitierung des einfachen Perzeptrons:

Aus der vorangehenden Diskussion ergibt sich sofort, dass ein Perzeptron nur dann Muster in Klassen einteilen kann, wenn diese durch eine Hyperebene zu trennen sind. Man sagt in diesem Fall: die Klassen sind 'linear separierbar'; die Hyperebenen werden 'lineare Diskriminanten' genannt (siehe Abschnitt 9.3). Ein bekanntes, einfaches Beispiel, bei dem das einfache Perzeptron keine Lösung findet, ist die XOR-Funktion (Exclusive-OR) definiert in der Tabelle in Abb. 9.13. Man erkennt sofort, dass die Bedingungen an die Gewichte nicht gleichzeitig erfüllt werden können. Das entspricht der Tatsache, dass in Abb. 9.13 keine Gerade gefunden werden kann, die die $(y = 0)$- von der $(y = 1)$-Klasse trennt.

Ein anderes Beispiel von nicht linear separierbaren Punktemengen ist in Abb. 9.14a gezeigt. In solchen Fällen kann man eventuell doch noch eine Perzeptron-Lösung finden, wenn man ein weiteres Merkmal findet, dass die Klassen diskriminiert. Die trennende Hyperebene läge dann in einem um eine Dimension erweiterten Raum (Abb. 9.14b). Das Problem ließe sich auch mit Hilfe komplizierterer Transferfunktionen lösen, was aber dem grundlegenden Konzept für neuronale Netze (möglichst einfache Einzelschritte) widerspräche.

Eine allgemein anwendbare Lösung findet man durch Erweiterung des Perzeptron-Modells auf mehrschichtige Netze.

9.4.5 Das Mehrlagen-Perzeptron

Lösung des XOR-Problems:

Wir haben gesehen, dass ein einfaches Perzeptron durch

$$\vec{w}\vec{x} \;=\; 0 \tag{9.23}$$

Hyperebenen im Musterraum definiert, die den Raum in die beiden Klassen

$$\begin{array}{lll} \vec{w}\vec{x} & < \; 0 & \text{Klasse 1} \\ \vec{w}\vec{x} & > \; 0 & \text{Klasse 2} \end{array} \tag{9.24}$$

unterteilt. Mit der Kombination von Hyperebenen lassen sich offensichtlich Volumina im Musterraum definieren. Eine solche Kombination gelingt tatsächlich durch die Erweiterung des einfachen Perzeptrons um eine (oder mehrere) Lagen. Dieses Mehrlagen-Perzeptron hat dann neben den Eingangs- und Ausgangslagen auch versteckte Lagen (hidden layers).

Bei dem XOR-Problem (Abb. 9.13) sehen wir, dass die 1-Klasse zwischen den beiden für das AND und das OR gefundenen Hyperebenen (Abb. 9.12) liegt. Das liegt natürlich daran, dass sich das XOR aus einer entsprechenden AND-OR-Kombination ergibt:

$$y(XOR) \;=\; \overline{y(AND)} \;\wedge\; y(OR).$$

Wir definieren also ein dreilagiges Netz mit 2 Knoten in der Eingangslage, 2 Knoten in der versteckten Lage, 1 Knoten in der Ausgangslage (Netz-Konfiguration: 2 - 2 - 1). Die Aktivitäten der Knoten und die Gewichte sind:

$\vec{x}$: Eingangsaktivitäten,

$\vec{x}'$: Aktivitäten der versteckten Knoten,

y : Ausgangsaktivität (im Allgemeinen auch ein Vektor),

$\vec{w}_i$: Gewichte für die Eingänge ($i = 1, 2$ ist der Index der versteckten Knoten),

$\vec{w}'$: Gewichte für die Ausgänge $\vec{x}'$ der versteckten Knoten.

In Abb. 9.15 sind an die Netz-Verbindungen die Gewichte w_{i1}, w_{i2} bzw. w'_1, w'_2 und an die Knoten die Schwellen $-w_{i0}$ bzw. $-w'_0$ geschrieben. Mit der Tabelle sieht man, dass in diesem Netz die beiden versteckte Knoten jeweils das AND und OR realisieren und die Ausgangslage die logische Verknüpfung von beiden. Die 1-Klasse des Netzes liegt also zwischen den beiden Geraden in Abb. 9.15b, die 0-Klasse außerhalb.

Für das Anlernen von Netzen ist es wichtig zu sehen, dass die Lösungen für die Klassenseparation nicht eindeutig sind. In unserem Beispiel gibt es eine unendliche Schar von Hyperebenen, die kontinuierlich durch Translationen und Rotationen auseinanderhervorgehen und die, solange sie nicht einen der Musterpunkte überspringen, dasselbe leisten. Problematischer für die Kontrolle des Lernens ist allerdings, dass es auch Lösungen geben kann, die nicht kontinuierlich zusammenhängen. Für das XOR-Problem finden wir zum Beispiel die in Abb. 9.16 angegebene Lösung, bei der die zwei Hyperebenen diesmal die 0-Klasse einschließen, während die 1-Klasse außerhalb liegt.

x_1	x_2	x'_1	x'_2	$w'_1x'_1 + w'_2x'_2$	y
0	0	0	0	0	0
1	0	0	1	1	1
0	1	0	1	1	1
1	1	1	1	0	0

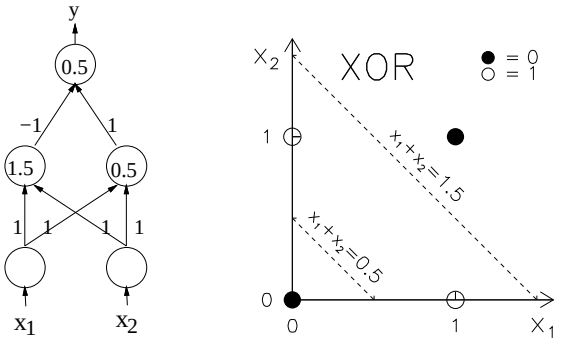


Abbildung 9.15: Links: Wahrheitstafel für das XOR-Netz auf der rechten Seite. Mitte: Netzwerk mit Gewichten und Schwellen zur Lösung des XOR-Problems. Rechts: Musterraum des XOR-Problems mit den durch das Netz bestimmten Hyperebenen.

x_1	x_2	x'_1	x'_2	$w'_1x'_1 + w'_2x'_2$	y
0	0	0	1	-1	0
1	0	0	0	0	1
0	1	1	1	0	1
1	1	0	1	-1	0

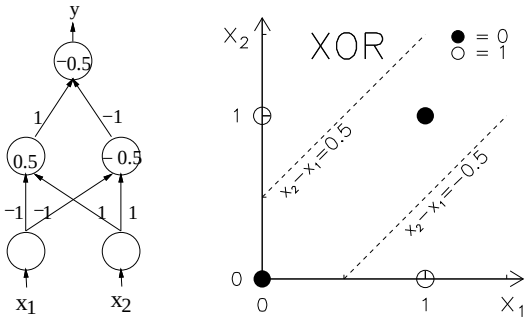


Abbildung 9.16: Links: Wahrheitstafel für das XOR-Netz auf der rechten Seite. Mitte: Netzwerk mit Gewichten und Schwellen zur Lösung des XOR-Problems (alternativ zu Abb. 9.15). Rechts: Musterraum des XOR-Problems mit den durch das Netz bestimmten Hyperebenen.

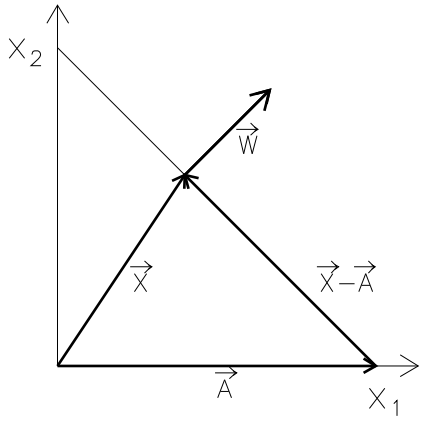


Abbildung 9.17: Zur Darstellung der Hesseschen Normalform der Geradengleichung.

Die Hessesche Normalform für die Hyperebenen:

Die Gleichung einer Hyperebene, $\vec{w}\vec{x} = 0$, ist offensichtlich invariant gegenüber einer Transformation

$$\vec{w} \rightarrow -\vec{w} \tag{9.25}$$

Dasselbe gilt aber nicht für die Klasseneinteilung durch $\vec{w}\vec{x} < 0$ und $\vec{w}\vec{x} > 0$, weil durch (9.25) die Klassen gerade vertauscht werden. Wir wollen uns deshalb die Bedeutung der Orientierung von $\vec{w}$ genauer klar machen.

Für die folgenden Überlegungen wollen wir die Gewichte und Vektoren für einen 2-dimensionalen Musterraum betrachten:

$$\begin{aligned} \vec{X} &= (x_1, x_2) \\ \vec{W} &= (w_1, w_2) \end{aligned}$$

(die großen Buchstaben sollen von den Vektoren $\vec{x}$ und $\vec{w}$ unterscheiden, die ja mit den 0-Komponenten die Schwellen enthalten). Dann ist die Gleichung der Hyperebene:

$$\vec{W}\vec{X} = -w_0,$$

so dass auch für einen festen Ortsvektor $\vec{A}$ eines Punktes auf der Geraden gilt:

$$\vec{W}\vec{A} = -w_0$$

und damit:

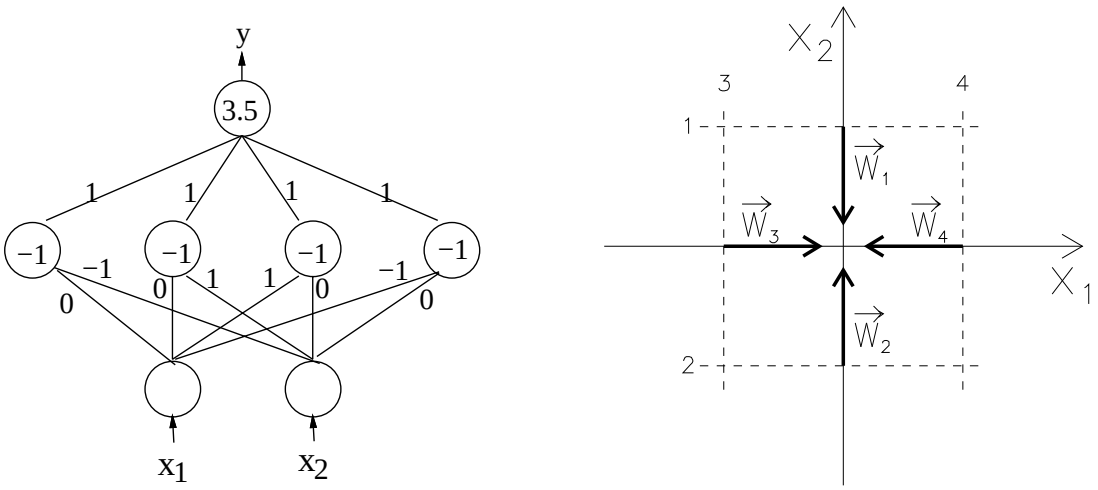
$$\vec{W}(\vec{X} - \vec{A}) = 0 \tag{9.26}$$

Das heißt, $\vec{W}$ steht senkrecht auf $\vec{X} - \vec{A}$ und damit senkrecht auf der Geraden, weil $\vec{X} - \vec{A}$ die Richtung der Geraden hat (Abb. 9.17). Durch die Wahl des Vorzeichens der Gewichte wird damit eine Orientierung der Normalen auf der Hyperebene festgelegt. Gleichung (9.26) ist die Hessesche Normalform der Geradengleichung (wobei genau genommen $\vec{W}$ zu normieren wäre).

Musterklassifizierung mit einem Dreilagen-Perzeptron:

Die Punkte in dem Quadrat $[-1 < x < +1; -1 < y < +1]$ sollen zur Musterklasse A gehören (Abb. 9.18). Um diese Klasse zu separieren, sind dann 4 verdeckte Knoten notwendig, die jeweils eine Begrenzungsgerade festlegen (siehe Tabelle in Abb. 9.18). Wenn man die Vorzeichen so wählt, dass die Gewichtsvektoren alle in das Volumeninnere zeigen (Abb. 9.18), dann lassen sich die Ausgänge der verdeckten Knoten alle mit positiven Gewichten kombinieren, um die Klasse A zu selektieren.

Θ-Funktion als Übertragungsfunktion: Benutzt man die Θ-Funktion als Übertragungsfunktion dann wird mit den Gewichten und Schwellen in Abb. 9.18 das Quadrat exakt herausgeschnitten.



i	Geraden-Gl.	w_{i0}	w_{i1}	w_{i2}	w'_i
1	$-x_2 + 1 = 0$	1	0	-1	1
2	$x_2 + 1 = 0$	1	0	1	1
3	$x_1 + 1 = 0$	1	1	0	1
4	$-x_1 + 1 = 0$	1	-1	0	1

Abbildung 9.18: Oben: a) Netzwerk mit Gewichten und Schwellen zur Selektion der Punkte innerhalb des in b) gezeigten Quadrates. Unten: Definition der Geraden und Gewichtsvektoren für das Netzwerk in der Abbildung. Der Index i steht sowohl für einen versteckten Knoten als auch für die zu diesem Knoten gehörige Gerade.

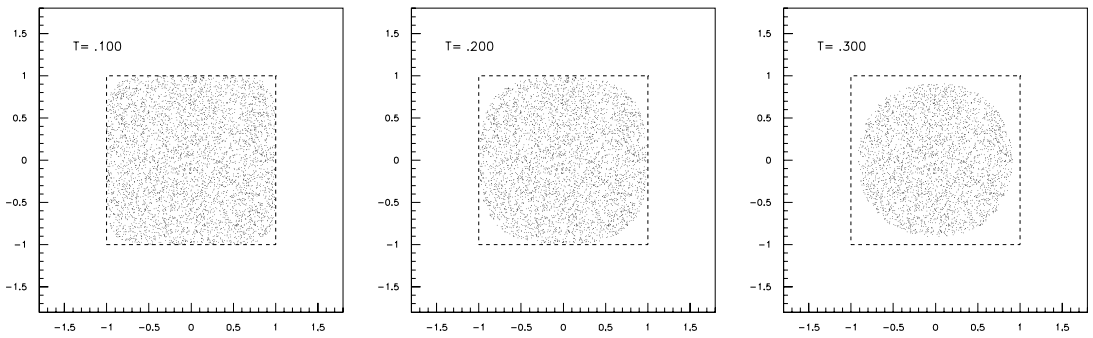


Abbildung 9.19: Durch das Netz in Abb. 9.18 selektierte Punktmenge bei Benutzung einer sigmoiden Schwellenfunktion mit Temperaturparameter a) $T = 0.1$, b) $T = 0.2$, c) $T = 0.3$.

Sigmoide Übertragungsfunktion: Bei Verwendung von sigmoiden Funktionen als Übertragungsfunktion werden in der ersten verdeckten Lage die trennenden Hyperebenen immer noch scharf definiert. Im Gegensatz zu der 0-1-Entscheidung ('links' oder 'rechts' von der Hyperebene) der Θ-Funktion erhält man hier jedoch ein kontinuierliches Maß für den Abstand von der Hyperebene. Erst bei der gewichteten Summe dieser Abstände in der nächsten Stufe spielt die relative Größe der Abstände eine Rolle. In dieser Summe kann nämlich ein kleiner Abstand von einer Hyperebene einen großen Abstand von einer anderen Ebene kompensieren. Das führt zu Abrundungen von Ecken bei der Klassifikation und erlaubt im Allgemeinen die Konturen des Klassenvolumens besser zu approximieren.

In Abb. 9.19 wird gezeigt, wie sich die Kontur der selektierten Punktmenge verändert, wenn man im obigen Beispiel des Quadrates statt der Θ-Funktion die 'logistische Funktion' (9.19) mit dem Temparaturparameter $T = 1$ benutzt.

An diesem Beispiel läßt sich der Einfluß des Parameters T gut verdeutlichen: Für $T \to 0$ nähert man sich der Θ- Funktion an und damit nähert sich das ausgeschnittene Volumen mehr dem Quadrat; für $T \to \infty$ wird das Volumen abgerundeter. Trotz dieses starken Einflusses ist ein variabler T-Parameter eigentlich überflüssig: die Wirkung von T kann durch geeignete Normierung der Gewichte ebenso erreicht werden (große Gewichte ergeben scharfe Grenzen und umgekehrt). In der Lernphase kann es sich andererseits als nützlich erweisen, mit einem T-Parameter das Lernverhalten zu steuern.

9.4.6 Lernen

Die Lernstrategie:

Für Feed-Forward-Netze sind Lernstrategien entwickelt worden, bei denen das Netz mit Hilfe eines Trainingsdatensatzes lernt, die richtige Antwort zu geben. Während des Trainings kann das Netz seine Antwort mit der richtigen vergleichen; das ist also die Situation 'Lernen mit Lehrer' (supervised learning). Wenn wir Muster in Klassen einteilen wollen, erwarten wir für einen Mustervektor $\vec{x}$ folgende Antworten y_j:

$$\vec{x} \;\to\; y_j \quad \begin{array}{ll} = & 1 \quad \text{wenn } \vec{x} \text{ in Klasse } j \\ = & 0 \quad \text{sonst} \end{array}$$

Dieses Lernziel ist sofort einsichtig, wenn die Klassen disjunkt sind. Wir wollen es aber auch beibehalten, wenn die Klassen sich überlappen wie im Fall der beiden Gauß-Verteilungen in Abb. 9.20. Wenn die Fläche unter den Kurven ein Maß für die Häufigkeit des Auftretens von Mustern der jeweiligen Klasse ist, dann ist die optimale Trennung dort, wo beide Wahrscheinlichkeiten gleich sind, d.h. der Schnittpunkt beider Kurven ('Bayes-Diskriminante'). Wir werden sehen, dass ein wohl-trainiertes Netz diesen optimalen Grenzfall erreichen kann.

Wie gut das Netz gelernt hat, wird mit einem dem Netz unbekannten Datensatz getestet, d.h. man prüft, ob das Netz das Gelernte auf unbekannte Daten übertragen, ob es 'generalisieren' kann.

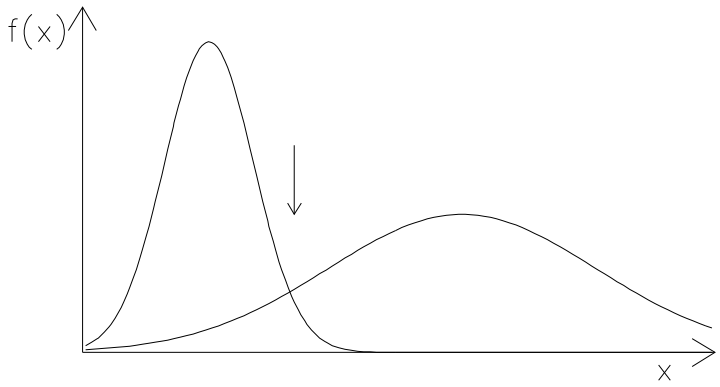


Abbildung 9.20: Beispiel für überlappende Verteilungen im Musterraum.

Lernalgorithmen:

Wir betrachten ein Feed-Forward-Netz mit n Lagen, die Ausgangsaktivitäten der k-ten Lage seien durch den Vektor $\vec{x}^k$ gegeben, die Gewichte zwischen der k-ten Lage und dem i-ten Knoten in der k+1-ten Lage sei $\vec{w}_i^k$. Das Netz hat dann folgende Struktur:

$$
\begin{array}{lccccccc}
 & \bigcirc & \bigcirc & \bigcirc & \bigcirc \ldots & x_i^n & = & g(\sum_j w_{ij}^{n-1} x_j^{n-1}) = y_i \\
w_{ij}^{n-1} & & & & & & & \\
 & \bigcirc & \bigcirc & \bigcirc & \bigcirc \ldots & x_i^{n-1} & = & g(\sum_j w_{ij}^{n-2} x_j^{n-2}) \\
 & \cdot & \cdot & \cdot & \cdot \quad \cdot & \cdot & \cdot & \cdot \\
 & \cdot & \cdot & \cdot & \cdot \quad \cdot & \cdot & \cdot & \cdot \\
 & \cdot & \cdot & \cdot & \cdot \quad \cdot & \cdot & \cdot & \cdot \\
 & \bigcirc & \bigcirc & \bigcirc & \bigcirc \ldots & x_i^2 & = & g(\sum_j w_{ij}^1 x_j^1) \\
w_{ij}^1 & & & & & & & \\
 & \bigcirc & \bigcirc & \bigcirc & \bigcirc \ldots & x_i^1 & = & \text{Eingabe}
\end{array}
$$

Der Trainingsdatensatz enthalte N Mustervektoren, für jedes Muster p ($p = 1, \ldots, N$) und für jeden Ausgangsknoten i sei die richtige Antwort $\hat{y}_i^{(p)}$ bekannt, die mit der Antwort $y_i^{(p)}$ des Netzes verglichen werden kann. Als Maß für die Optimierung des Netzwerkes definieren wir die Fehlerfunktion (l ist die Zahl der Ausgangsknoten)

$$E = \frac{1}{2} \sum_{p=1}^{N} \sum_{i=1}^{l} (y_i^{(p)} - \hat{y}_i^{(p)})^2 \tag{9.27}$$

Die Fehlerfunktion soll durch Variation der Gewichte w_{ij}^k minimiert werden, es muß also gelten:

$$\frac{\partial E}{\partial w_{ij}^k} = 0 \qquad k = 1, \ldots n-1 \tag{9.28}$$

Da E nicht-linear von den Gewichten abhängt, kann das Gleichungssystem (9.28) im allgemeinen nur iterativ gelöst werden. Wir wählen das für solche Optimierungsprobleme geläufige Gradientenabstiegs-Verfahren (Abb. 9.21) um das (globale) Minimum zu suchen. Es sei hier bemerkt, dass es bei multi-dimensionalen Problemen

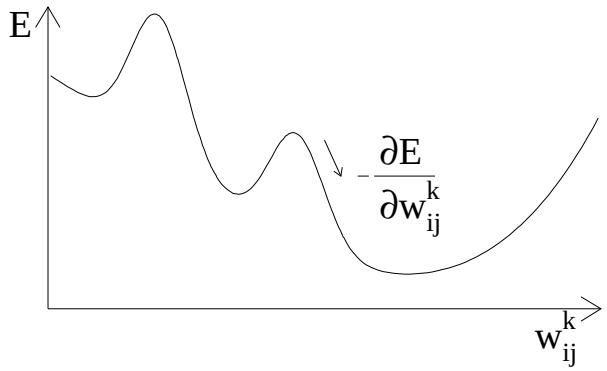


Abbildung 9.21: Beispiel für den Verlauf einer Fehlerfunktion im Gewichtsraum.

im allgemeinen sehr schwierig ist, das globale Minimum zu finden. Für unsere Anwendungen ist es aber in der Regel nicht wichtig, ob das Netz tatsächlich das globale Minimum gefunden hat, wenn es nur ein relativ gutes gefunden hat.

Die Fehlerfunktion soll also entlang des negativen Gradienten im Gewichtsraum schrittweise verkleinert werden. Dazu korrigieren wir jedes Gewicht w_{ij}^k entsprechend:

$$\Delta w_{ij}^k = -\eta \frac{\partial E}{\partial w_{ij}^k} \tag{9.29}$$

Wenn der Lernparameter η genügend klein ist (damit es keine Oszillationen um das Minimum gibt), kann die Korrektur nach jedem angebotenen Muster p erfolgen:

$$\Delta w_{ij}^k = -\eta \frac{\partial E^{(p)}}{\partial w_{ij}^k}$$

Dann stellt jedes Muster bereits einen Iterationsschritt dar; in der Regel ist dieses Verfahren schneller, als wenn man vor jeder Gewichtskorrektur erst über alle N Muster mittelt. Aus Stabilitätsgründen kann es allerdings manchmal vorteilhaft sein über eine kleine Zahl m von Mustern zu mitteln (m≈10).

Eine effiziente Methode, die Gewichtskorrekturen für die verschiedenen Lagen zu berechnen, ist der Backpropagation-Algorithmus, den wir allerdings hier aus Zeitgründen nicht näher besprechen.

Training:

Im folgenden sollen einige Begriffe, die beim Training von FF-Netzen auftreten, erläutert werden:

Trainingsdatensatz: Der Trainingsdatensatz enthält N Muster, die jeweils den Eingabevektor $\vec{x}^{(p)}$ und die erwartete Antwort $\vec{y}^{(p)}$ enthalten:

$$(\vec{x}^{(p)}, \hat{\vec{y}}^{(p)}),\ p = 1, \ldots, N \tag{9.30}$$

Lernzyklen: Im allgemeinen muß das Lernen relativ langsam erfolgen ($\eta < 1$), damit das Minimum sicher gefunden werden kann. Um zum Minimum zu kommen, muß der Trainingsdatensatz in der Regel wiederholt dargeboten werden (Lernzyklen).

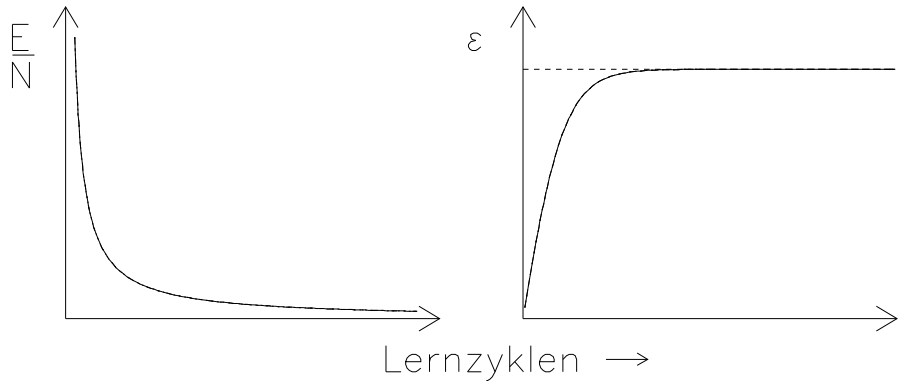


Abbildung 9.22: Kontrolle der Konvergenz: typische Verläufe der Fehlerfunktion (links) und der Effizienz (rechts).

Konvergenzkontrolle: Die **Konvergenz** des Verfahrens wird nach jedem Zyklus (oder nach q Zyklen) getestet durch Auswertung der Fehlerfunktion E (oder meistens E/N) oder der **Effizienz** der Selektion für jede Klasse i:

$$\epsilon_i = \frac{N_i^{net}}{N_i^{in}} \tag{9.31}$$

Dabei ist N_i^{net} die Anzahl der Muster, die vom Netz richtig in die i-te Klasse eingeordnet werden, und N_i^{in} die Anzahl der dem Netz angebotenen Muster der Klasse i. Die Effizienz sollte in einen Sättigungswert übergehen, der je nach Überlapp der Klassen zwischen 50% und 100% liegen sollte (100% kann nur für disjunkte Klassen erwartet werden). Abbildung 9.22 zeigt das erwartete Verhalten der Fehlerfunktion und der Effizienz.

Generalisierung: Die Bewährungsprobe für ein Netz ist schließlich der Nachweis, dass es das Gelernte auf einen ihm unbekannten Testdatensatz anwenden kann. Geprüft wird auch hier die Fehlerfunktion und die Effizienzen für die verschiedenen Klassen. Im allgemeinen sind die Effizienzen etwas niedriger und die Fehlerfunktion etwas größer als für die Trainingsdaten. Bei zu großer Diskrepanz ist zu prüfen, ob das Netz durch 'Overtraining' zu stark an die Trainingsdaten angepaßt ist. Das ist dann auch ein Hinweis, dass das Netz wahrscheinlich zuviele Freiheitsgrade hat.

Praktische Regeln zum Netzwerktraining:

Wahl von 'intelligenten' Variablen: Um gute Resultate mit Neuronalen Netzen zu erzielen, ist es in der Regel wichtig, die benutzten Variablen geschickt auszuwählen und eventuell vorzuverarbeiten.

Kontrolle von Lerngeschwindigkeit und Konvergenzverhalten: Es gibt viele verschiedene Methoden, um das Lernen, das häufig sehr zeitaufwendig sein kann, effektiver zu machen. Dazu gehört die dynamische Anpassung des Lernparameters an die Variation der Fehlerfunktion mit den Gewichten. Statistische Schwankungen

im Trainigsdatensatz können durch Hinzufügen eines "Trägheitsterms", der proportional zur Gewichtsänderung im vorhergehenden Schritt ist, gedämpft werden:

$$\Delta w_{ij}^k(t+1) \;=\; -\eta \frac{\partial E}{\partial w_{ij}^k}(t) + \alpha\, \Delta w_{ij}^k(t). \tag{9.32}$$

Dabei ist der Trägheitsparameter α auf das Problem abzustimmen.

Beschränkung der Komplexität eines Netzes:

Wieviele Lagen sind notwendig? Mit 2 Lagen können linear separierbare Probleme behandelt werden (siehe Lösungen der AND-, OR-Probleme mit dem Perzeptron).

Mindestens 3 Lagen werden gebraucht, wenn das Problem nicht linear separierbar ist (zum Beispiel, wenn eine Klasse in zwei disjunkten Bereichen, getrennt durch eine andere Klasse, liegen; siehe XOR-Problem). Ohne Beweis sei angegeben: Mit einem 3-Lagen-Netz kann

- jede kontinuierliche Funktion $y \;=\; f(\vec{x})$ approximiert werden,
- jede Boolsche Funktion $y \;=\; f(x_1, \ldots, x_n)$, mit $y,\; x_i \;= 1$ oder 0, dargestellt werden.

Wieviele Knoten pro Lage? Ein geschlossenes Volumen in n Dimensionen kann im allgemeinen durch n+1 Hyperebenen (oder weniger, wenn es zu einer oder mehreren Seiten offen ist,) eingeschlossen werden. Mehr als n+1 Hyperebenen pro geschlossenem, zu selektierendem Volumen liefert mehr Freiheit, den Konturen zu folgen (für das Quadrat ist offensichtlich n+2=4 eine bessere Wahl der Anzahl der Hyperebenen). Wir halten also fest:

- In der Regel sind mindestens n+1 Knoten in der ersten versteckten Lage notwendig.
- Die Zahl der Knoten in der zweiten versteckten Lage hängt von der Komplexität des Problems ab, insbesondere von der Anzahl der nicht-zusammenhängenden Volumina. Es ist wahrscheinlich nicht mehr als ein Knoten pro Volumen notwendig.
- Es sollten so wenig Knoten wie möglich definiert werden, um die Generalisierungsfähigkeit des Systems sicherzustellen.

Entfernen und Generieren von Verbindungen und Knoten: Um die Komplexität des Netzes so gering wie möglich zu halten, sind Techniken entwickelt worden, die erlauben, unwichtige Verbindungen und Knoten zu erkennen und zu entfernen oder auch notwendige Verbindungen und Knoten zu generieren.

Selbstgenerierung der Netz-Architektur: Bei diesem Vorgehen beginnt man zunächst mit einem sehr einfachen Netz und baut dann sukzessiv neue Verbindungen, Knoten und Lagen auf, deren Notwendigkeit dann wieder durch das Verhalten der Fehlerfunktion, der Konvergenz etc. geprüft werden kann.

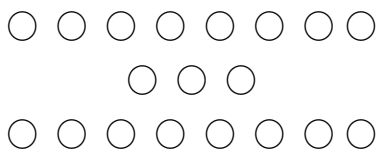
Tabelle 9.1: Vorzeichen der für das Encoder-Problem gefundenen Gewichte w_{ij} in der ersten Schicht.

i $j \rightarrow$	1	2	3	4	5	6	7	8
1	-	+	-	+	+	+	-	-
2	+	-	-	-	+	+	+	-
3	+	-	+	+	-	+	-	-

9.4.7 Typische Anwendungen für Feed-Forward-Netze

Beispiel für ein binäres Netz: 8-Bit-Encoder:

Wir trainieren ein (8-3-8)-Netz



mit 8 Mustervektoren $\vec{x}^p = (x_1^p, \ldots, x_8^p),\ p = 1, \ldots, 8$, und den erwarteten Netzantworten $\hat{\vec{y}}^p = (\hat{y}_1^p, \ldots, \hat{y}_8^p),\ p = 1, \ldots, 8$, denen folgende Binärwerte zugeordnet werden:

$$\begin{aligned} x_i^p &= \delta_{ip} \\ \hat{y}_i^p &= \delta_{ip} \end{aligned}$$

Wir erwarten also das gleiche Muster am Eingang und Ausgang. Wie schafft es das Netz diese Information durch das Nadelöhr von nur 3 Knoten in der versteckten Lage zu transportieren?

Das Netz wurde mit einem PC-Programm (NNSIMU) trainiert. Die Gewichte in der ersten Schicht ergaben sich alle zu etwa $|w_{ij}| \approx 5$. Das Interessante an den Gewichten ist eigentlich nur ihr Vorzeichen, siehe Tab. 9.1. Das Vorzeichen von w_{ij} gibt in diesem Fall direkt die Aktivität des i-ten versteckten Knotens an, wenn das j-te Muster anliegt. Aus der Tabelle erkennt man sofort, dass das Netz den Binärcode 'entdeckt' hat: die redundanten 8-Bit-Sequenzen sind in 3-Bit-Binärzahlen umgewandelt worden.

Funktionsapproximation:

Wie bereits in Abschnitt 9.4.6 ausgeführt, kann mit einem 3-lagigen Netz jede kontinuierliche Funktion,

$$\vec{x} = (x_1, \ldots, x_n) \rightarrow y = f(\vec{x}),$$

approximiert werden.

In Abb. 9.23 ist das Ergebnis eines Trainings der Funktion

$$y = \sin x, \quad 0 < x < \pi$$

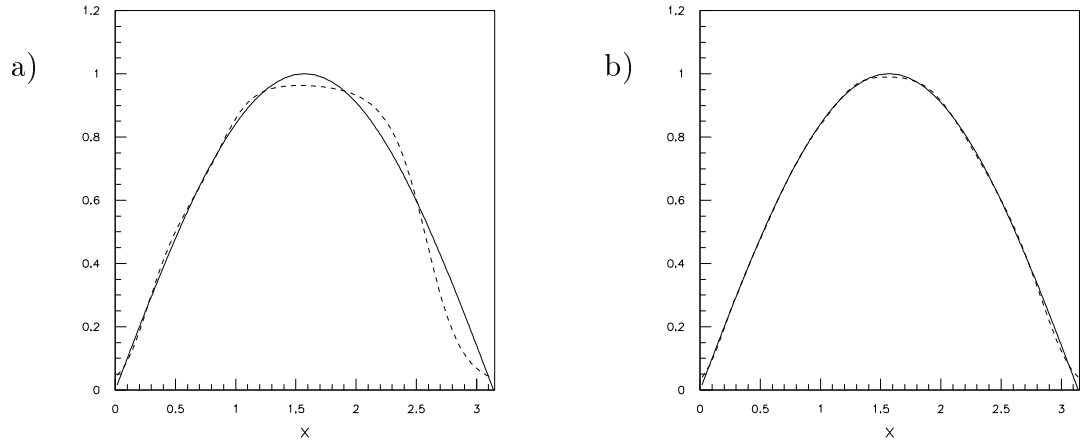


Abbildung 9.23: Approximation einer Sinus-Funktion durch ein (1-8-1)-Netz. Trainingszeiten: a) einige Sekunden, b) etwa 8 Stunden.

gezeigt. Trainiert wurde ein (1-8-1)-Netz mit 200 Musterpaaren (x, y), äquidistant verteilt auf der x-Achse. Nach einigen Lernzyklen, entsprechend einer Rechenzeit von einigen Sekunden, ergab sich die Approximation in Abb. 9.23a. Erst nach etwa 8 Stunden wurde die ausgezeichnete Reproduktion des Sinus durch das Netz in Abb. 9.23b erzielt (diese extrem lange Zeit für ein doch relativ einfaches Problem zeigt eigentlich nur, dass das benutzte Programm nicht sehr effektiv war).

In Abb. 9.24 sind einige Zwischenwerte des Netzes als Funktion von x dargestellt. Es läßt sich gut erkennen, wie daraus die Sinus-Funktion zusammengebaut wird. Außerdem wird durch einige fast verschwindende Aktivitäten nahegelegt, dass Knoten in der versteckten Lage (zum Beispiel der 6. und 8. Knoten) überflüssig sein könnten, die in einem nächsten Schritt entfernt werden könnten.

Klassifikationsprobleme:

Das Problem, Muster in verschiedene Klassen einzuordnen, tritt in unterschiedlichsten Zusammenhängen auf, zum Beispiel:

- Einteilung in disjunkte Klassen: als Beispiele mit kontinuierlichen Musterräumen hatten wir das Quadrat behandelt (siehe Abb. 9.18); Beispiele für diskrete Musterräume sind die Boolschen Funktionen (AND, OR, XOR, ...).

- Die Muster verschiedener Klassen können im allgemeinen auch in Verteilungen liegen, die sich überlappen. Ein einfaches Beispiel sind die überlappenden Gauß-Verteilungen in Abb. 9.20 (mehr dazu im nächsten Abschnitt).

  Gemeinsam ist diesen Fragestellungen, dass von einem bestimmten Muster nicht unbedingt gesagt werden kann, in welcher Klasse es liegt. Im allgemeinen kann nur eine Wahrscheinlichkeit angegeben werden, einer bestimmten Klasse anzugehören. Was die optimale Trennung ist und wie ein NN entscheidet, wird im nächsten Abschnitt besprochen.

- Mustererkennung: Eine der großen Herausforderungen für die Neuroinformatik ist die Verarbeitung und das Erkennen von visuellen, auditiven oder anderen kognitiven Mustern. Von den bisherigen Beispielen unterscheidet sich diese

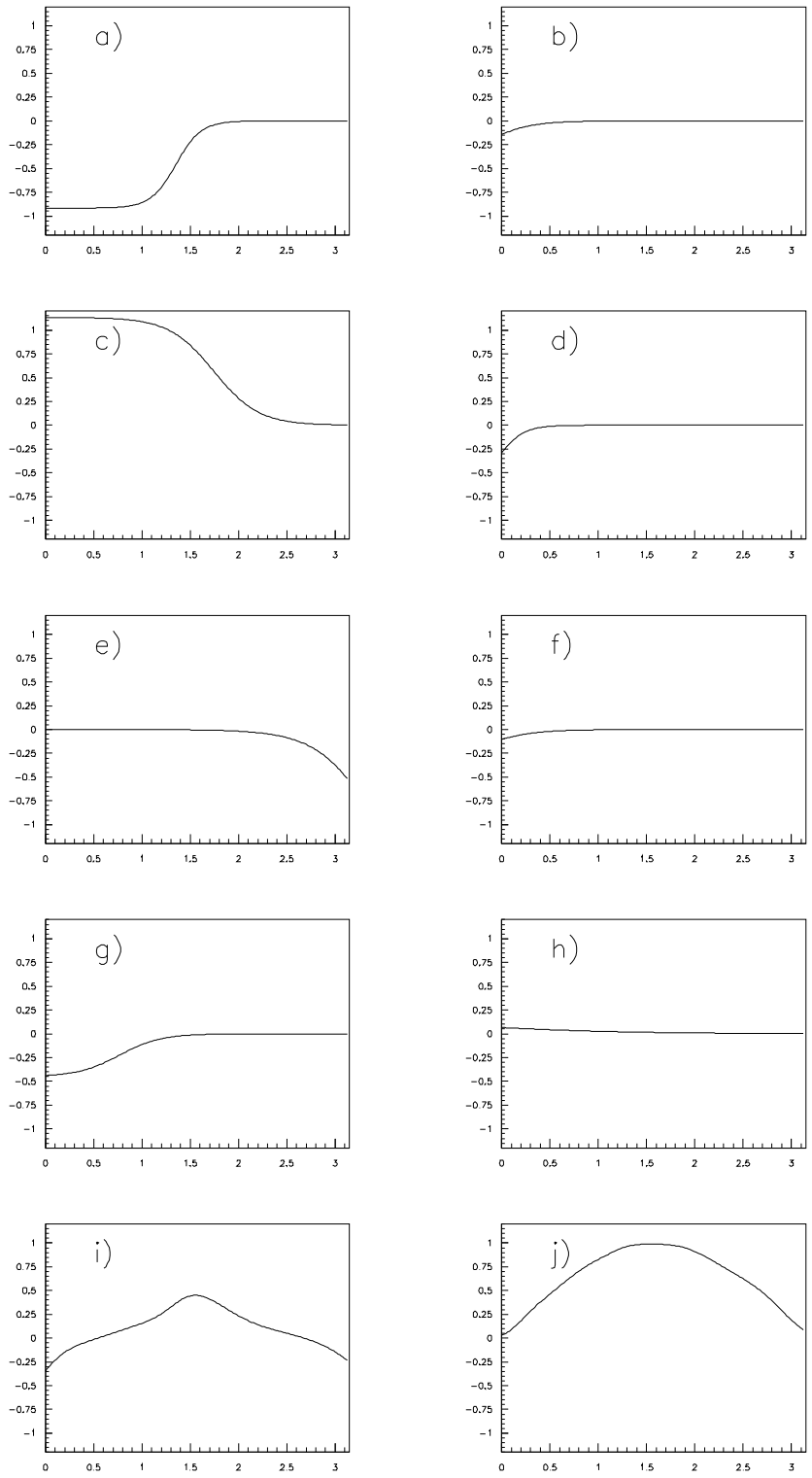


Abbildung 9.24: Für das in Abb. 9.23b benutzte Netz sind als Funktion von x dargestellt: a) bis h) die 8 gewichteten Ausgänge der versteckten Knoten $v_i = w_i' g(z_i)$; i) die Aktivität des Ausgangsknotens $z' = \sum_{i=1,\ldots,8} v_i$, j) das Ausgangssignal $y = g(z')$.

Problemstellung im wesentlichen durch ihre sehr viel größere Komplexität. Ein Bild beispielsweise muß in sehr viele Pixel unterteilt werden, die als Eingabe für das Netz dienen; die Netze werden damit sehr umfangreich. Ein besonderes Problem ist auch die Dynamik, durch die neben der räumlichen auch die zeitliche Dimension ins Spiel kommt. Besonders wichtige Eigenschaften der Netze sind Fehlertoleranz und Rauschunterdrückung.

9.4.8 BP-Lernen und der Bayes-Diskriminator

Die Bayes-Diskriminante:

Es seien Musterklassen $C_i,\ (i = 1, \ldots, m)$, gegeben. Der Bayes-Diskriminator ordnet einen Mustervektor $\vec{x}$ in diejenige Klasse C_i ein, für die die folgende Bayes-Diskriminanten-Funktion maximal ist:

$$P(C_i|\vec{x}) = \frac{p(\vec{x}|C_i)\, P(C_i)}{\sum_{j=1}^{m} p(\vec{x}|C_j)\, P(C_j)} \tag{9.33}$$

Dabei ist

$P(C_i\|\vec{x})$	(a posteriori) Wahrscheinlichkeit, dass $\vec{x}$ in Klasse C_i ist,
$P(C_i)$	(a priori) Wahrscheinlichkeit für Klasse C_i,
$p(\vec{x}\|C_i)$	Wahrscheinlichkeitsverteilung für $\vec{x}$, wenn es in Klasse C_i liegt.

Die Wahrscheinlichkeiten sind normiert:

$$\sum_i P(C_i) \;=\; 1; \qquad \int_{\Omega^n} p(\vec{x}|C_i)\, d^n x$$

Es ist wichtig zu beachten, dass Ω^n das 'beobachtete' Volumen ist, d.h. im allgemeinen ist die tatsächliche Verteilung noch mit einer Akzeptanzfunktion η zu korrigieren:

$$p(\vec{x}|C_i) \;\rightarrow\; p(\vec{x}|C_i)\, \eta(\vec{x}|C_i)$$

Beispiel: Bei der Teilchenidentifikation durch Flugzeitmessung (TOF) wird der Impuls p und die Geschwindigkeit β gemessen. Daraus läßt sich das Quadrat der Masse ('TOF-Masse') bestimmen:

$$m^2_{TOF} \;=\; p^2(\frac{1}{\beta^2} - 1)$$

Die verschiedenen Klassen entsprechen den Teilchensorten Pion, Kaon und Proton $(C_i,\ i = \pi, K, p)$, die mit der Häufigkeit $P(C_i)$ auftreten. Unter der Annahme, dass m^2_{TOF} für eine Teilchensorte i Gauß-verteilt ist um die tatsächliche Masse m_i^2 des Teilchens, ergibt sich für die Verteilung von m^2_{TOF} unter der Hypothese i:

$$p(m^2_{TOF}|C_i) \;=\; \frac{1}{\sqrt{2\pi}\sigma_i} \exp \frac{-(m^2_{TOF} - m_i^2)^2}{2\sigma_i^2}$$

Ein typisches Meßergebnis ist in Abb. 9.25 gezeigt. Die Entscheidung wird dann für das Teilchen gefällt, für das die Diskriminanten-Funktion in (9.33) maximal ist.

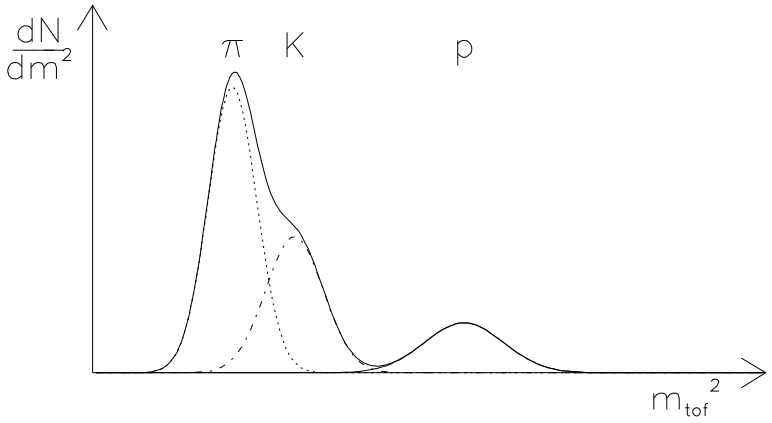


Abbildung 9.25: Typische Verteilung der Massenquadrate, berechnet aus einer Flugzeitmessung für Pionen , Kaonen und Protonen.

Approximation des Bayes-Diskriminators mit neuronalen Netzen:

Ein Netz sei auf die Trennung der beiden Klassen C_1 und C_2 trainiert worden, so dass die erwarteten Netzantworten jeweils sind:

$$\begin{aligned} \hat{y} &= 1 \quad \text{für } \vec{x} \text{ in } C_1 \\ \hat{y} &= 0 \quad \text{für } \vec{x} \text{ in } C_2 \end{aligned}$$

Dann berechnet sich der Erwartungswert der mittleren quadratischen Abweichungen der Netzantworten von den erwarteten Antworten:

$$E = \frac{1}{2} \int d\vec{x} \left[\alpha_1 p_1(\vec{x})(y(\vec{x}) - 1)^2 + \alpha_2 p_2(\vec{x})(y(\vec{x}))^2 \right] \tag{9.34}$$

Das Integral geht über den gesamten Musterraum; die α_i sind die Häufigkeiten, mit denen die Klassen C_i auftreten; die $p_i(\vec{x})$ sind die Wahrscheinlichkeitsverteilungen der Muster $\vec{x}$, wenn sie jeweils einer der beiden Klassen angehören. Mit den Definitionen aus dem vorigen Abschnitt gilt dann also:

$$\begin{aligned} \alpha_i &= P(C_i) \\ p_i(\vec{x}) &= p(\vec{x}|C_i) \end{aligned} \tag{9.35}$$

Bei überlappenden Verteilungen können in der Fehlerfunktion (9.34) die Fehleranteile beider Klassen ungleich Null sein. Dann wird das Minimum nicht mehr unbedingt für $y = 0$ oder 1 erreicht, sondern es gibt eine optimale Wahl des Netzes für y, die sich an jeder Stelle des Musterraumes aus folgender Bedingung herleiten läßt:

$$\frac{\partial E}{\partial y} = \alpha_1 p_1(\vec{x})(y(\vec{x}) - 1) + \alpha_2 p_2(\vec{x}) y(\vec{x}) = 0 \tag{9.36}$$

Die Auflösung nach y ergibt:

$$y(\vec{x}) = \frac{\alpha_1 p_1(\vec{x})}{\alpha_1 p_1(\vec{x}) + \alpha_2 p_2(\vec{x})} \tag{9.37}$$

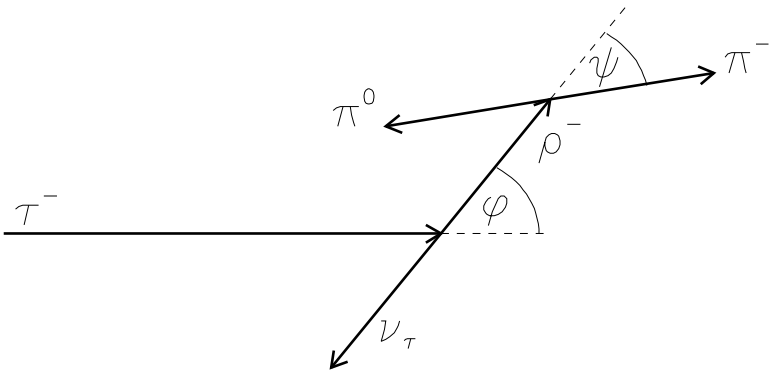


Abbildung 9.26: Darstellung der Zerfallswinkel in Reaktion (9.39).

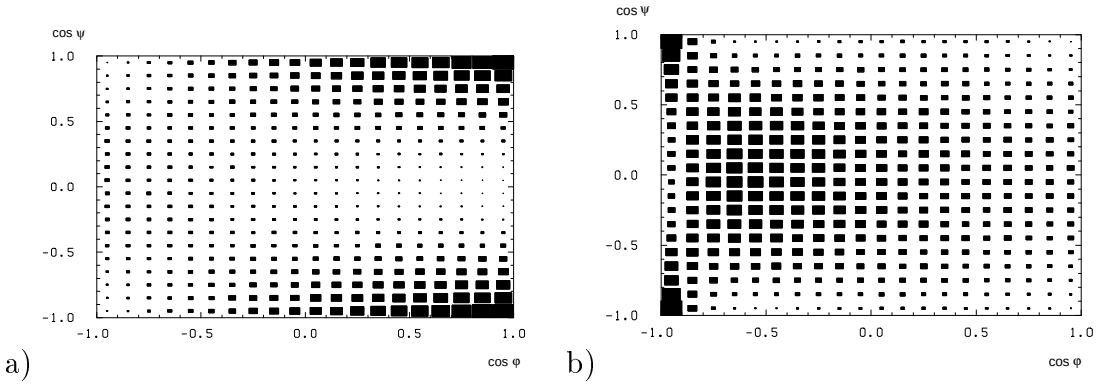


Abbildung 9.27: Winkelverteilung nach (9.40) für τ- Zerfälle im Helizitätszustand $+1$ (a) oder -1 (b).

Die Verallgemeinerung auf m Klassen lautet:

$$y_i(\vec{x}) \;=\; \frac{\alpha_i p_i(\vec{x})}{\sum_{j=1}^{m} \alpha_j p_j(\vec{x})} \tag{9.38}$$

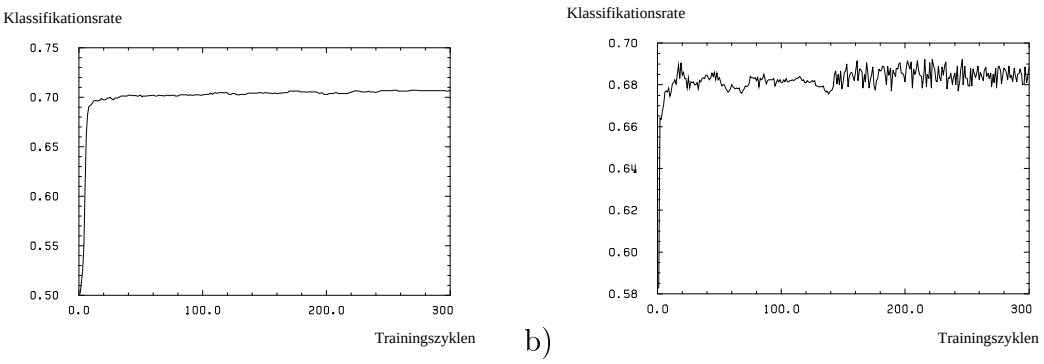
Das maximale y_i bestimmt, in welche Klasse das Muster einzuordnen ist. Bei zwei Klassen ist der Übergang offensichtlich gerade da, wo die beiden Wahrscheinlichkeiten gleich sind:

$$\alpha_1 p_1 \;=\; \alpha_2 p_2 \;\Longrightarrow\; y \;=\; 0.5$$

Im anschließenden Beispiel werden wir sehen, dass ein Netzwerk die optimale Lösung (9.38) approximieren kann.

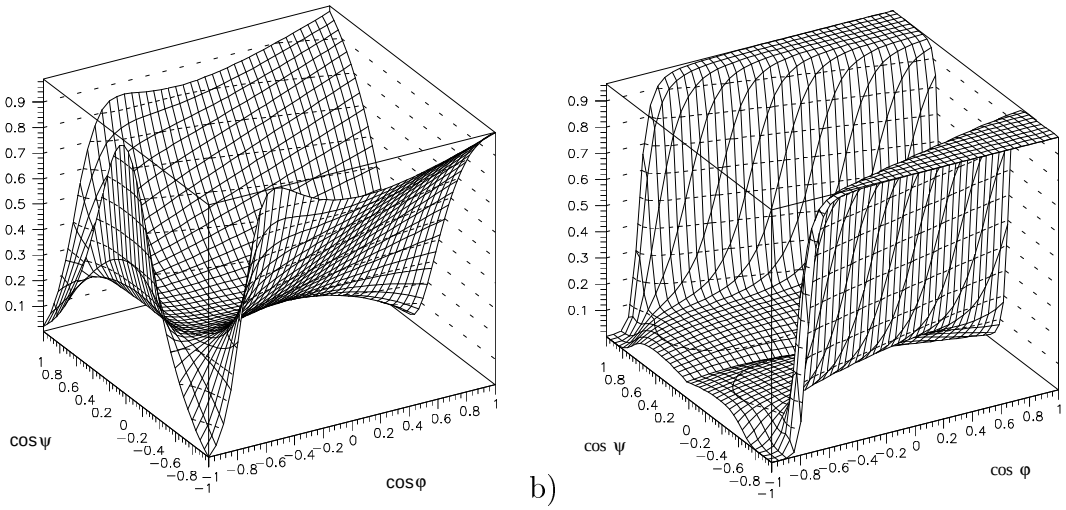
Beispiel für die Approximation des Bayes-Diskriminators durch ein Netz:

Als Beispiel für die Trennung von Klassen mit unterschiedlichen, aber überlappenden Verteilungen nehmen wir die Zerfallswinkelverteilungen von τ-Leptonen in den beiden möglichen Helizitätszuständen $h = +1$ und $h = -1$ (das τ-Lepton hat Spin $1/2$; die Helizität ist der auf ± 1 normierte Erwartungswert der Projektion des Spins

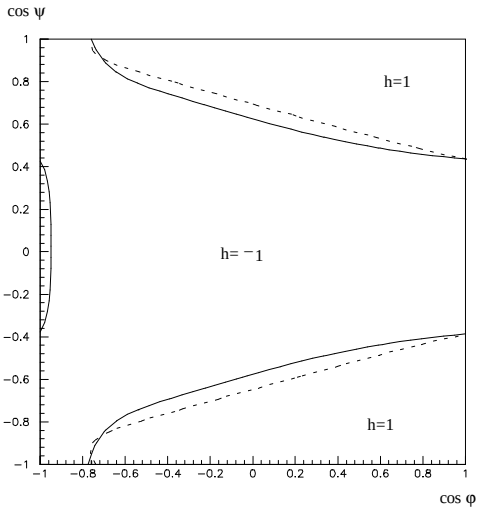


a) b)

Abbildung 9.28: Effizienzen für die Zuordnung des richtigen Helizitätszustandes. Das Netz wurde mit den Lernparametern a) $\eta = 0.001$, $\alpha = 0.9$ und b) $\eta = 0.1$, $\alpha = 0.9$ trainiert.



a) b)



c)

Abbildung 9.29: a) Bayes-Diskriminanten-Funktion aufgetragen über der ($\cos\phi$,-$\cos\psi$)-Ebene; b) dasselbe für den Ausgang y des Netzes. c) Klassifikationsgrenzen für die beiden Helizitäten (volle Linie: Bayes, gepunktete Linie: Netz).

eines Teilchens auf seine Flugrichtung). Wir nehmen an, die τ's seien in einem reinen Helizitätszustand ($h = \pm 1$) produziert worden.

Ein Zerfall, in dem sich die Spininformation im Endzustand gut messen läßt, ist der Zerfall des τ's in ein ρ-Meson mit Spin 1 und ein Neutrino mit Spin 1/2. Während das Neutrino nicht nachzuweisen ist, läßt sich die ρ-Spineinstellung über den ρ-Zerfall in zwei Pionen analysieren:

$$\tau \;\rightarrow\; \rho^- \nu_\tau \;\rightarrow\; \pi^- \pi^0 \nu_\tau \tag{9.39}$$

Die meßbaren Winkel sind der Winkel ϕ zwischen dem ρ und der Laborrichtung des τ (im Ruhesystem des τ) und der Winkel ψ zwischen dem π^- und dem ρ (im ρ Ruhesystem), siehe Abb. 9.26. Die beiden Winkelverteilungen sind Funktionen von $\cos\phi$ und $\cos\psi$:

$$\begin{aligned} P_{+1} &= \cos^2\psi \left[\cos\eta\cos\frac{\phi}{2} + \frac{m_\rho}{m_\tau}\sin\eta\sin\frac{\phi}{2}\right]^2 \\ &\quad + \frac{\sin^2\psi}{2}\left[\left(\sin\eta\cos\frac{\phi}{2} - \frac{m_\rho}{m_\tau}\cos\eta\sin\frac{\phi}{2}\right)^2 + \left(\frac{m_\rho}{m_\tau}\right)^2 \sin^2\frac{\phi}{2}\right] \\ P_{-1} &= \cos^2\psi \left[\cos\eta\sin\frac{\phi}{2} - \frac{m_\rho}{m_\tau}\sin\eta\cos\frac{\phi}{2}\right]^2 \\ &\quad + \frac{\sin^2\psi}{2}\left[\left(\sin\eta\sin\frac{\phi}{2} - \frac{m_\rho}{m_\tau}\cos\eta\cos\frac{\phi}{2}\right)^2 + \left(\frac{m_\rho}{m_\tau}\right)^2 \cos^2\frac{\phi}{2}\right] \end{aligned} \tag{9.40}$$

Dabei ist

$$\cos\eta \;=\; \frac{m_\tau^2 - m_\rho^2 + (m_\tau^2 + m_\rho^2)\cos\phi}{m_\tau^2 + m_\rho^2 + (m_\tau^2 - m_\rho^2)\cos\phi}$$

Abbildung 9.27 zeigt die sich ergebenden zwei-dimensionalen Verteilungen für die beiden Helizitäten.

Mit diesen Verteilungen wurde ein 3-lagiges FF-Netz darauf trainiert, die beiden Helizitäten zu unterscheiden. Die Netzkonfiguration war 2-8-1; der Trainingsdatensatz bestand aus 1000 Ereignissen, gleichviel von jeder Helizität. Abbildung 9.28 zeigt die Effizienz (Anzahl der richtig erkannten Ereignisse zur Gesamtzahl) in Abhängigkeit vom Lernzyklus für einen Testdatensatz. Mit dem Lernparameter $\eta = 0.001$ und dem Trägheitsparameter $\alpha = 0.9$ wird nach 300 Trainingszyklen eine Effizienz von nahezu 71% erreicht. Das kann verglichen werden mit der theoretisch berechenbaren Effizienz bei Benutzung des Bayes-Diskriminators, die sich zu 71.7% ergibt.

In Abb. 9.29 wird gezeigt, dass die Bayes-Diskriminanten-Funktion (Abb. 9.29a) von dem Ausgang y des Netzes (Abb. 9.29b) approximiert wird. Nach einem Schnitt bei $y = 0.5$ ergeben sich die Klassentrennungen, wie in Abb. 9.29c gezeigt. Ob noch eine bessere Approximation der Bayes-Trennung möglich ist, hängt neben einer ausreichenden Netzgröße auch von der Statistik des Trainingsdatensatzes ab. Es ist verständlich, dass zum Beispiel der kleine Zipfel bei $(-1, 0)$ von dem Netz nur dann richtig eingeordnet werden kann, wenn in diesem kleinen Bereich Ereignisse liegen.

9.5 Entscheidungsbäume

Wir betrachten wieder einen Datensatz von Ereignissen mit jeweils m Merkmalen, zusammengefasst in $\vec{x}$, die zwei verschiedenen Klassen angehören, zum Beispiel 'Signal' und 'Untergrund'. Im folgenden soll die Klassifizierung durch Entscheidungsbäume ('decision trees') eingeführt werden: Sequentielle Anwendung von Trennschnitten auf die Merkmale der Ereignisse verteilt die Daten auf verschiedene Äste, an deren Enden jeweils ein Blatt einer bestimmten Klasse zugeordnet ist. Zu derselben Klasse kann es mehrere Blätter geben, aber jedes Blatt ist nur auf einem Weg zu erreichen.

Im binären Entscheidungsbaum wird eine Serie von Fragen gestellt, welche alle mit Ja oder Nein beantwortet werden können. Diese Serie ergibt ein Resultat, welches durch eine Regel bestimmt ist. Die Regel ist einfach ablesbar, wenn man von der Wurzel her den Ästen des Baumes folgt, bis man zu einem bestimmten Blatt gelangt, welches das Resultat der Fragereihe darstellt. Bei kontinuierlich verteilten Merkmalen werden Trennschnitte bestimmt, die dann letztlich an jedem Punkt zu einer Ja-Nein-Entscheidung führt (größer/kleiner).

Generiert werden die Entscheidungsbäume üblicherweise bei der Wurzel beginnend schrittweise bis zu den Blättern gehend. Bei jedem Schritt wird genau das Merkmal gesucht, mit welchem man die Daten am besten klassifizieren kann. Um dieses zu ermitteln, muss die beste Aufteilung gefunden werden, das heißt die Aufteilung der Daten muss so gewählt werden, dass sie nach der Aufteilung möglichst rein sind. Ein Maß für die Reinheit ist zum Beispiel die Entropie. Aus der Entropie lässt sich dann berechnen, welches Merkmal für die Verzweigung den höchsten Informationsgewinn bietet. Ein weiteres Maß für die Bestimmung der optimalen Aufteilung ist der Gini-Index, der im Folgenden benutzt wird.

9.5.1 Aufwachsen eines Baumes

Ein Entscheidungsbaum wird mit einem Trainingsdatensatz in folgenden Schritten konstruiert:

1. Beginne mit dem Trainingsdatensatz an der Wurzel ('root node').

2. Suche aus allen Ereignissen das signifikanteste Merkmal zum Aufteilen an diesem Knoten.

3. Teile nach einem Kriterium wie zum Beispiel dem maximalen Gini-Index[1] auf:

$$G = 4\,P(1-P) \tag{9.41}$$

[1]Der Gini-Index ist ein Maß für Ungleichverteilung. Die hier benutzte Definition weicht etwas von der üblichen ab, bei der $G = (A_{gleich} - A)/A_{gleich}$ ist, wobei A_{gleich} die Fläche unter der Verteilungsfunktion für eine Gleichverteilung und A die Fläche unter der tatsächlichen Verteilungsfunktion ist. Aus Wikipedia: "Der Gini-Index oder auch Gini-Koeffizient ist ein statistisches Maß, das vom italienischen Statistiker Corrado Gini zur Darstellung von Ungleichverteilungen entwickelt wurde. Der Koeffizient kann beispielsweise als Kennzahl für die Ungleichverteilung von Einkommen oder Vermögen eingesetzt werden. Er wird besonders in der Wohlfahrtsökonomie verwendet."

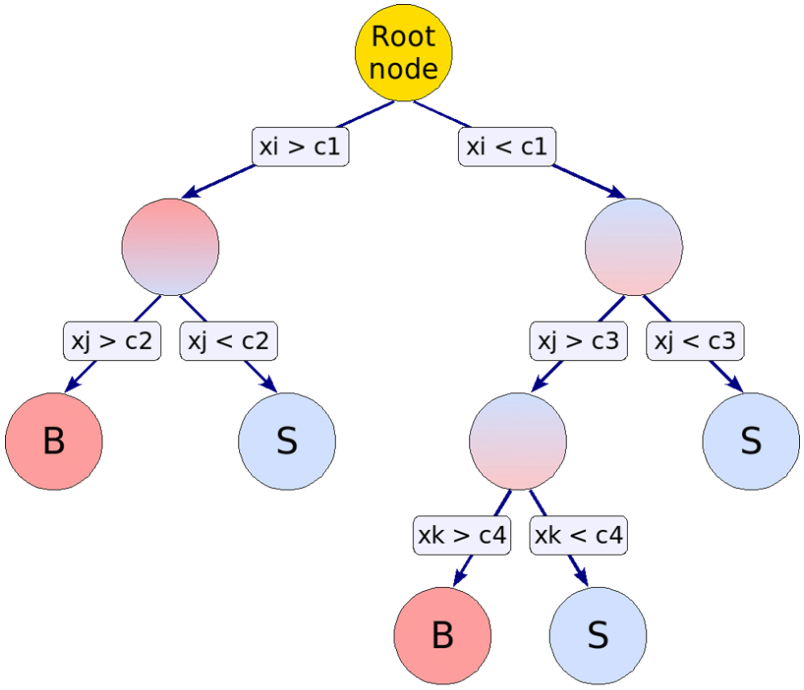


Abbildung 9.30: Prinzip eines Entscheidungsbaumes: an jeder Verzweigung (Knoten) wird das Merkmal ausgesucht, das die signifikanteste Trennung durch einen Selektionsschnitt erlaubt. Die Blätter (Endknoten) sind einer Klasse zugeordnet (hier S=Signal und B=Untergrund).

Dabei ist P die 'Reinheit' der Klasse 1, die bei einem Schnitt auf ein Merkmal an einem Knoten erreicht wird, und $1 - P$ die Reinheit der Klasse 2:

$$P = \frac{N_1}{N_1 + N_2} \tag{9.42}$$

Der Gini-Index wird für $P = 0.5$, entsprechend $G = 1$, maximal (mit der Normierung in (9.41) gilt $0 \leq G \leq 1$).

Der Gini-Index wird für die Bestimmung des Merkmals, das an einem Knoten die signifikanteste Trennung bietet und für die Bestimmung des Trennschnitts benutzt (maximiert).

4. Setze die Aufteilung fort, bis ein vorgegebenes Abbruchkriterium erfüllt ist, bis zum Beispiel eine minimale Anzahl Ereignisse in einem Knoten verbleibt oder bis eine maximale Reinheit erreicht ist.

5. Ein Blatt wird der Klasse zugeordnet, die die meisten Ereignisse in dem Blatt hat.

6. Evaluiere Effizienz und Reinheit mit einem unabhängigen und dem Baum bisher unbekannten Testdatensatz.

Für die Klassifizierung von Daten und die Lösung von Fragestellungen auf der Basis von Daten werden in den unterschiedlichsten Bereichen (Wirtschaft, Medizin,

Naturwissenschaften, ...) häufig Entscheidungsbäume benutzt. Die vorteilhaften Eigenschaften sind:

- Unabhängigkeit von gleichförmigen Variablentransformationen;
- Unanfälligkeit gegen Ausreißer in den Daten;
- Unterdrückung von 'schwachen' Variablen ohne Verlust der Leistungsfähigkeit.

Schwachstellen sind:

- Instabilität der Baumstruktur gegenüber kleinen Änderungen der Trainingsdaten;
- Anfälligkeit auf Übertraining (Abhilfe: 'pruning' = 'Ausasten');

Eine Klassifizierung mit einem Entscheidungsbaum hat also einige nicht ganz optimale Eigenschaften. Eine wesentliche Verbesserung stellen 'verstärkte Entscheidungsbäume' dar, wie im Folgenden besprochen wird.

9.5.2 Verstärkte Entscheidungsbäume

Ein weitaus besseres Klassifikationsvermögen wird dadurch erreicht, dass viele Bäume generiert werden und deren Ergebnisse gemittelt werden. Nach jeder Erzeugung eines Baumes gehen die falschen Zuordnungen mit einem höheren Gewicht in die nächste Erzeugung eines Baumes ein, wodurch sie mit höherer Wahrscheinlichkeit richtig eingeordnet werden. Die Klassenzugehörigkeit wird durch Mittelung der Entscheidung aller Bäume ermittelt ('verstärkte Entscheidungsbäume', 'boosted decision trees').

Das Training beginnt wie bei einem einzelnen Baum, wobei alle Ereignisse das Gewicht 1 haben. Bei der Erzeugung des nächsten Baumes wird jedem Ereignis ein Gewicht w_i zugeordnet, das von dem angewandten Algorithmus abhängt. Die Berechnung der Reinheit P in (9.42) ändert sich dann entsprechend zu

$$P = \frac{\sum_{i=1}^{N_1} w_i}{\sum_{i=1}^{N_1} w_i \; + \; \sum_{i=1}^{N_2} w_i} \tag{9.43}$$

Nach der Fertigstellung des Baumes werden die Gewichte wieder für den nächsten Baum berechnet. Das geht so weiter bis eine vorggebene Maximalzahl M von Bäumen generiert worden ist (typisch $M \approx 1000$).

Die Entscheidungsfunktion eines einzelnen Baumes sei:

$$y_k(\vec{x}) = \pm 1, \qquad k = 1, \ldots, M, \tag{9.44}$$

(zum Beispiel $y = +1$ für Klasse 1 und $y = -1$ für Klasse 2). Für die Gesamtentscheidung wird das gewichtete Mittel der einzelnen Entscheidungen gebildet:

$$y(\vec{x}) = \frac{\sum_{k=1}^{M} g_k \, y_k(\vec{x})}{\sum_{k=1}^{M} g_k} \tag{9.45}$$

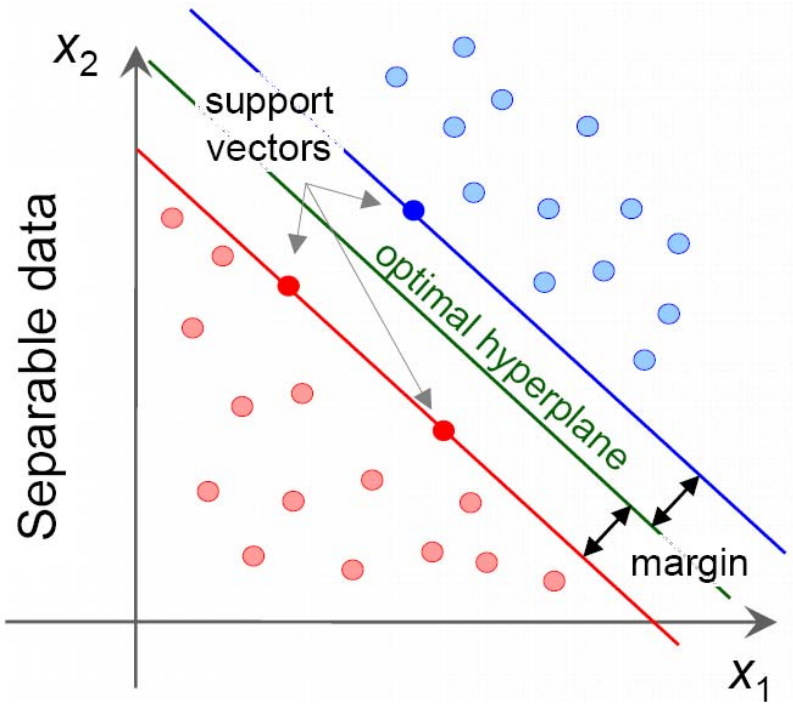


Abbildung 9.31: Zwei disjunkte Datenmengen, die durch eine Diskriminante mit der größten Trennspanne separiert werden.

Die Gewichte werden so gewählt, dass eine dem speziellen Algorithmus zugeordnete Verlustfunktion, die im Allgemeinen eine Funktion der richtigen und falschen Zuordnungen ist, minimiert wird. Als Beispiel ist der Algorithmus AdaBoost in [4] erklärt. In der TeV-Gamma-Astronomie (MAGIC, HESS) ist 'Random Forest'[2] beliebt.

Mit 'boosted decision trees' werden die Klassifizierungen wesentlich stabiler als mit einzelnen Bäumen. Durch die Mittelung der Einzelentscheidungen in (9.45) ergibt sich auch ein Maß für die Wahrscheinlichkeit der richtigen Einordnung. Die Eigenschaften scheinen durchaus mit Neuronalen Netzen vergleichbar oder vielleicht sogar überlegen zu sein.

9.6 Stützvektormaschinen

Das Konzept einer so genannten "Stützvektormaschine (SVM)" ('support vector machine') greift die Idee auf, dass eigentlich nur Merkmalvektoren in der Nähe der Trennung zwischen den Klassen wesentlich sind: aus einem Trainingsdatensatz werden die Vektoren, die im wesentlichen die Trennung definieren, als "Stützvektoren" ausgewählt.

Im Folgenden werden wir zunächst die lineare Variante der SVM besprechen und dann die vielleicht interessantere Variante für die Anwendung auf nicht linear separierbare Klassen.

[2] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

9.6.1 Lineare SVM-Klassifikation

Wir gehen zunächst von zwei disjunkten Klassen wie in Abb.9.31 aus. Die Klassen sollen durch eine lineare Diskriminante, also eine Hyperebene wie in Abschnitt 9.3 eingeführt, getrennt werden. Die Lage der diskriminierenden Hyperebene soll nun so optimiert werden, dass die nächsten Trainingsvektoren in beiden Klassen maximal von der Ebene entfernt sind, dass also der Trennungsstreifen möglichst breit wird. Ein solches Trainingsziel führt zu einer optimalen Generalisierungsfähigkeit.

Die Ränder des Trennungsstreifens sind zwei parallele Hyperebenen, die durch die Stützvektoren festgelegt werden sollen. Offensichtlich braucht man in m Dimensionen mindestens $m+1$ Stützvektoren. Zum Beispiel können m Vektoren eine Ebene festlegen und der verbleibende Vektor den Abstand der beiden Ebenen (siehe den zwei-dimensionalen Fall in Abb. 9.31). Die Aufgabe ist also, die maximale Trennung und die mindestens $m+1$ Stützvektoren zu bestimmen.

Die Diskriminante wird analog zu der Fisher-Diskriminante (Abschnitt 9.3.2) definiert:

$$\vec{w}^T \vec{x} + b = 0. \tag{9.46}$$

Wenn $\vec{w}$ ein Einheitsvektor ist, gibt b den Abstand vom Ursprung an, wenn im allgemeinen $\vec{w}$ kein Einheitsvektor ist, ist der Abstand von Ursprung durch $b/|\vec{w}|$ gegeben. Die beiden Randhyperebenen sollen in der Form

$$\vec{w}^T \vec{x} + b = \pm 1. \tag{9.47}$$

gegeben sein, was die Skala für $\vec{w}$ und b festgelegt. Dann ist der Abstand der Randebenen zur Diskriminante $d = 1/|\vec{w}|$. Für zwei Vektoren $\vec{x}^{(1)}$ und $\vec{x}^{(2)}$, die jeweils zu einer Randebene weisen, gilt:

$$\vec{w}^T(\vec{x}^{(1)} - \vec{x}^{(2)}) = 2. \tag{9.48}$$

Der tatsächliche Abstand zwischen den Hyperebenen ist

$$\frac{\vec{w}^T}{|\vec{w}|}(\vec{x}^{(1)} - \vec{x}^{(2)}) = \frac{2}{|\vec{w}|} = 2\,d. \tag{9.49}$$

Für alle Vektoren $\vec{x}$ gilt

$$|\vec{w}^T \vec{x} + b| \;\geq\; 1, \tag{9.50}$$

und zwar je nach Klassenzugehörigkeit

$$\vec{w}^T \vec{x} + b \;\geq\; +1 \quad \text{oder} \quad \vec{w}^T \vec{x} + b \;\leq\; -1. \tag{9.51}$$

Eine Testgröße für die Klassenzugehörigkeit wird deshalb durch folgende Funktion definiert:

$$y = y(\vec{x}) = \operatorname{sgn}\left(\vec{w}^T \vec{x} + b\right) = \pm 1 \tag{9.52}$$

Um einen möglichst großen Abstand der Randebenen zu bekommen, muss nach Gleichung (9.49) der Betrag des Normalenvektors minimiert werden,

$$|\vec{w}| = \text{Minimum}. \tag{9.53}$$

Dabei sollen gleichzeitig die Ereignisse beider Klassen außerhalb des Trennungsstreifens bleiben:

$$\text{Nebenbedingung}: \quad |\vec{w}^T \vec{x}_i + b| = y_i \left(\vec{w}^T \vec{x}_i + b\right) \geq 1, \quad i = 1, \ldots, N. \tag{9.54}$$

Die N Nebenbedingungen können mit der Methode der Lagrange-Multiplikatoren in eine 'Zielfunktion' einbezogen werden:

$$L(\vec{w}, b, \vec{\alpha} | \vec{x}_i,\, i = 1, \ldots, N) = \frac{1}{2}|\vec{w}|^2 - \sum_{i=1}^{N} \alpha_i \left(y_i \left(\vec{w}^T \vec{x}_i + b\right) - 1\right) \tag{9.55}$$

Diese Funktion soll bezüglich den Parametern $\vec{w}, b$ bei festem $\vec{\alpha}$ minimiert werden. Aus dem Verschwinden der Ableitungen,

$$\frac{\partial L}{\partial w_j} = 0 \quad (j = 1, \ldots, m); \qquad \frac{\partial L}{\partial b} = 0, \tag{9.56}$$

ergibt sich:

$$\vec{w} = \sum_{i=1}^{N} \alpha_i\, y_i\, \vec{x}_i \quad \text{und} \quad \sum_{i=1}^{N} \alpha_i\, y_i = 0. \tag{9.57}$$

Die Zwangbedingungen in (9.55) führen zu der Sattelpunkt-Bedingung (bezüglich der α_i), der Kuhn-Karush-Tucker-Bedingung:

$$\alpha_i \left\{y_i \left(\vec{w}^T \vec{x}_i + b\right) - 1\right\} = 0, \quad \forall\, i = 1, \ldots, N. \tag{9.58}$$

Das bedeutet, dass die α_i nur dann ungleich 0 sein können, wenn der Ausdruck in der geschweiften Klammer 0 ist, was aber nur für die Punkte auf dem Rand des Trennstreifens der Fall ist. Damit tragen nur die Merkmalsvektoren $\vec{x}_i$ mit $\alpha_i \neq 0$, die alle auf den Rändern liegen und Stützvektoren (support vectors) genannt werden, zu der Definition von $\vec{w}$ in (9.57) bei:

$$\vec{w} = \sum_{i=1}^{N_{SV}} \alpha_i\, y_i\, \vec{x}_i \tag{9.59}$$

Dabei geht die Summe nur über die N_{SV} Stützvektoren.

Der Ausdruck für den Normalenvektor $\vec{w}$ in (9.57) enthält die bisher noch nicht bestimmten Lagrange-Multiplikatoren α_i. Die Ausdrücke in (9.57) werden in die Formel für L in (9.55) eingesetzt, was nach einiger Rechnung ergibt:

$$L(\vec{w}, b, \vec{\alpha}\,) \;\rightarrow\; L_D(\vec{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j\, y_i y_j\, \vec{x}_i^T \vec{x}_j \tag{9.60}$$

mit den Nebenbedingungen:

$$\sum_{i=1}^{N} \alpha_i\, y_i = 0 \quad \text{und} \quad \alpha_i \geq 0. \tag{9.61}$$

Damit ist das Problem jetzt auf das sogenannte 'duale Problem', die Maximierung von $L_D(\vec{\alpha})$ bezüglich $\vec{\alpha}$ unter den Nebenbedingungen (9.61) zurückgeführt, das mit

numerischen Methoden gelöst werden kann. Mit der Lösung für die α_i ist der Normalenvektor der Trennebene vollkommen bestimmt:

$$\vec{w} = \sum_{i=1}^{N_{SV}} \alpha_i \, y_i \, \vec{x}_i \tag{9.62}$$

Nur die Stützvektoren auf dem Rand des Trennbereiches tragen zur Festlegung von $\vec{w}$ bei. Mit einem der Stützvektoren, zum Beispiel auf der '+'-Seite, kann jetzt noch b berechnet werden:

$$\vec{w}^T \, \vec{x}_{SV+} + b = +1 \quad \Rightarrow b = 1 - \vec{w}^T \, \vec{x}_{SV+} \tag{9.63}$$

Damit kann für jeden zu klassifizierenden Vektor $\vec{x}$ die Entscheidungsfunktion y in (9.52) bestimmt werden:

$$y = y(\vec{x}) = \operatorname{sgn}\left(\vec{w}^T \, \vec{x} + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{N_{SV}} \alpha_i \, y_i \, \vec{x}_i^T \, \vec{x} + b\right) = \pm 1 \tag{9.64}$$

Bemerkenswert ist, dass nur Skalarprodukte des Testvektors mit den Stützvektoren zu berechnen und linear zu kombinieren sind. Die Tatsache, dass die Merkmalsvektoren nur in Skalarprodukten auftreten, macht man sich für eine Erweiterung des Merkmalsraumes in höhere Dimensionen mit einem verallgemeinerten Skalarprodukt zu Nutze, um auch nicht linear-separable Probleme zu lösen (siehe folgender Abschnitt).

Ohne hier in Details zu gehen, sei noch angemerkt, dass mit der linearen SVM auch moderat überlappende Klassen geteilt werden können, indem man die strikten Zwangsbedingungen (9.54) durch zusätzliche Terme mit so genannten 'Schlupfvariablen' aufweicht.

9.6.2 Nichtlineare Erweiterung mit Kernelfunktionen

Der oben beschriebene Algorithmus klassifiziert die Daten mit Hilfe einer linearen Funktion. Diese ist jedoch nur optimal, wenn auch das zu Grunde liegende Klassifikationsproblem linear separabel ist. In vielen Anwendungen ist dies aber nicht der Fall. Ein möglicher Ausweg ist, die Daten in einen Raum höherer Dimension abzubilden[3] (Abb.9.32):

$$\phi : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}, x \mapsto \phi(x) \qquad (d_1 < d_2). \tag{9.65}$$

Durch diese Abbildung wird die Anzahl möglicher linearer Trennungen erhöht (Theorem von Cover). Bei einer linearen Separierbarkeit gehen in die relevante Entscheidungsfunktion (9.64) die Datenpunkte $\vec{x}_i$ nur in Skalarprodukten ein. Daher ist es möglich, das Skalarprodukt $\vec{x}_i^T \vec{x}_j$ im Eingaberaum $\mathbb{R}^{d_1}$ durch ein Skalarprodukt $\langle \phi(x_i), \phi(x_j) \rangle$ im $\mathbb{R}^{d_2}$ zu ersetzen und stattdessen direkt zu berechnen. Die Kosten dieser Berechnung lassen sich sehr stark reduzieren, wenn eine positiv definite Kernel-Funktion als Skalarprodukt benutzt wird ('Kernel-Trick'):

$$k(\vec{x}_i, \vec{x}_j) = \langle \phi(\vec{x}_i), \phi(\vec{x}_j) \rangle \tag{9.66}$$

[3]Siehe auch http://de.wikipedia.org/wiki/Support_Vector_Machine

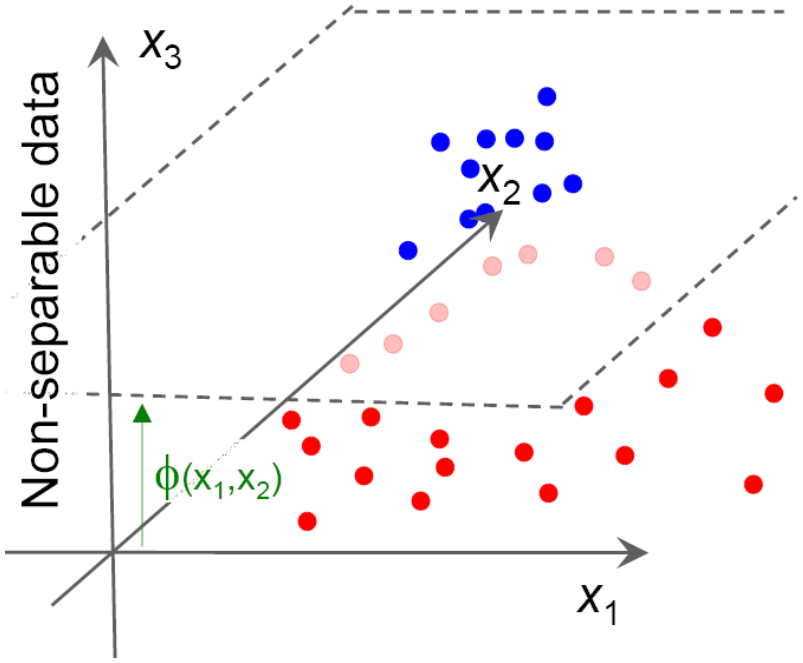


Abbildung 9.32: Beispiel eines in zwei Dimensionen nicht linear-separablen Datensatzes. Durch Transformationin eine höher dimensionalen Raum ist eine lineare Separation erreichbar.

Durch dieses Verfahren kann eine Hyperebene in einem höher-dimensionalen Raum implizit berechnet werden. Der resultierende Klassifikator hat die Form

$$y(\vec{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i \, y_i \, k(\vec{x}_i, \vec{x}) + b\right). \tag{9.67}$$

Obwohl durch die Abbildung ϕ implizit ein möglicherweise unendlich-dimensionaler Raum benutzt wird, generalisieren SVM immer noch sehr gut.

Die Kern-Funktionen müssen symmetrisch und positiv definit sein. Beispiele sind:

- Polynomial (homogen): $k(\vec{x}, \vec{x}') = (\vec{x} \cdot \vec{x'})^d$
- Polynomial (inhomogen): $k(\vec{x}, \vec{x}') = (\vec{x} \cdot \vec{x'} + 1)^d$
- Radiale Basisfunktion: $k(\vec{x}, \vec{x}') = \exp\left(-\frac{|\vec{x}-\vec{x'}|^2}{2\sigma^2}\right)$
- Sigmoid-Funktion: $k(\vec{x}, \vec{x}') = \tanh(\kappa \vec{x} \cdot \vec{x'} + c)$, für $\kappa > 0$ und $c < 0$.

Beispiel: Mit einem einfachen Beispiel soll die Beziehung der Kernel-Funktionen zu Skalarprodukten in höher-dimensionalen Räumen erläutert werden: Es seien zwei Vektoren $\vec{x}_1$ und $\vec{x}_2$ in einem zwei-dimensionalen Merkmalsraum gegeben:

$$\vec{x}_1 = (x_{11}, x_{12}), \qquad \vec{x}_2 = (x_{21}, x_{22}) \tag{9.68}$$

Als Kern-Funktion wählen wir die inhomogene Polynomial-Funktion mit $d = 2$ aus:

$$\begin{aligned} k(\vec{x}_1, \vec{x}_2) &= (\vec{x}_1 \cdot \vec{x}_2 + 1)^2 \\ &= (x_{11}x_{21} + x_{12}x_{22} + 1)^2 \\ &= 2x_{11}x_{21} + 2x_{12}x_{22} + (x_{11}x_{21})^2 + (x_{12}x_{22})^2 + 2x_{11}x_{21}x_{12}x_{22} + 1 \end{aligned} \tag{9.69}$$

Die Zuordnung

$$\phi(\vec{x}_1) = \phi((x_{11}, x_{12})) = (1, \sqrt{2}x_{11}, \sqrt{2}x_{12}, x_{11}^2, x_{12}^2, \sqrt{2}x_{11}) \tag{9.70}$$

ist eine nicht-lineare Abbildung des 2-dimensionalen Raumes auf einen 6-dimensionalen Raum, in dem das Skalarprodukt durch die Kernel-Funktion definiert ist:

$$\langle \phi(\vec{x}_i), \phi(\vec{x}_j) \rangle = k(\vec{x}_i, \vec{x}_j) \tag{9.71}$$

Tatsächlich braucht die Transformation in die höhere Dimension (die auch unendlich sein kann, zum Beispiel bei der Gauss-Funktion) nicht durchgeführt zu werden, da man nur die Skalarprodukte berechnen muss, die durch die Kernel-Funktion gegeben sind.