Empirische Risiko-Minimierung für dynamische Datenstrukturen

Dissertation

zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik
der Universität Dortmund

vorgelegt von
Thomas Fender

Dortmund 2003

1. Gutachter: Prof. Dr. Ursula Gather

2. Gutachter: Prof. Dr. Claus Weihs

Tag der mündlichen Prüfung: 16. Dezember 2003

Inhaltsverzeichnis

1 Einleitung

Die wissenschaftliche Beschreibung von Regeln und Zusammenhängen in physikalischen, biologischen, ökonomischen und soziologischen Systemen geschieht heute zunehmend mit Hilfe mathematischer Modelle. Für die Erstellung derartiger Modelle werden Beobachtungen der Systemvariablen benötigt (vgl. Cherkassky und Mulier, 1998). Mit diesen sollen einerseits wesentliche Muster erkannt werden, die gewissermaßen zum Lernen eines geeigneten Modells führen, andererseits können Messwerte und Beobachtungen in einem zweiten Schritt benutzt werden, um die gefundenen Modelle zu verifizieren.

Durch die bemerkenswerten Weiterentwicklungen hauptsächlich der Computerwissenschaften bei der Datenerfassung, -speicherung und -organisation werden immer häufiger große Mengen von Daten auch aus komplexen, hochdimensionalen Systemen erhoben, die genutzt werden können und sollten, um Zusammenhänge zwischen Systemvariablen zu extrahieren. Statistische Methodik muss darauf ausgerichtet werden, derartige interessanten Muster und Zusammenhänge aus Daten zu extrahieren und zu formulieren. Das Ziel der Datenanalyse sollte daher sein, direkt aus den Daten zu lernen (Hastie, Tibshirani und Friedman, 2001). In den Hintergrund tritt dann die Analyse von Daten unter der Annahme einer festen Modellklasse bzw. Verteilungsannahme.

Die wichtigste Herausforderung bei der Entwicklung einer Theorie für das *Lernen aus Daten* ist die Notwendigkeit ein möglichst generelles Konzept zu entwerfen, das ohne Voraussetzungen an die Daten für eine Vielzahl von Datensituationen gültig ist. Ein erstes allgemeingültiges Konzept des Lernens wurde von Valiant (1984) postuliert. Danach ist ein Verfahren oder eine Methode lernfähig, wenn es Muster bzw. Zusammenhänge in den Daten identifiziert und charakterisiert, so dass sie generalisierbar sind. Das bedeutet, Lernergebnisse, die durch ein Verfahren aus einer Stichprobe eines Datensatzes extrahiert wurden, müssen auch auf einer anderen Stichprobe aus denselben Daten gültig sein. Zusätzlich wird von einem Lernverfahren gefordert, dass der Lernprozess in ausreichend schneller Zeit durchführbar ist.

Lernen aus Daten lässt sich kategorisieren in das *unüberwachte Lernen*, mit dem Verknüpfungen und Muster zwischen Variablen erkannt werden sollen, und in das *überwachte* bzw. *prediktive Lernen*, bei dem mit einer gewissen Anzahl von Werten bzw. Variablen ein ausgehender Wert vorhergesagt oder geschätzt werden soll. Prediktives Lernen wird im Bereich Statistik genauso genutzt und entwickelt wie im Bereich künstliches Lernen. Dazu gehören Verfahren zur Funktionenapproximation, zur (nichtparametrischen) Regression und zur Mustererkennung sowie Verfahren des maschinellen Lernens (Friedman, 1994). Überwachtes Lernen ist also eine in der Statistik ebenso bekannte wie auch angewendete Methodik.
Das Lernkonzept von Valiant (1984) wurde von Vapnik (1995) in den Kontext der mathematischen Statistik übertragen. Motivation für diese *statistische Lerntheorie* ist der Wunsch, den Verlust zu minimieren, der eintritt, wenn im Rahmen eines Lernprozesses ein funktionaler Zusammenhang in den Daten spezifiziert wurde. Aus der statistischen Entscheidungstheorie stammt die Idee, auf Basis empirischer Daten das Risiko, d. h. den erwarteten Verlust, zu minimieren. Im Gegensatz zum klassischen verteilungsbasierten Ansatz, bei dem mit vorgegebenen Modell- bzw. Verteilungsannahmen und statistischer Inferenz das Risiko geschätzt und dann minimiert wird, ist das *Prinzip der empirischen Risiko-Minimierung (ERM-Prinzip)* ein wahrscheinlichkeitstheoretisches Konzept, bei dem das empirische Risiko direkt minimiert werden kann. Der Nachweis, dass dieses Prinzip Konsistenz, d. h. Konvergenz gegen die beste, das Risiko minimierende Lösung, garantiert, erfordert die Gültigkeit von Gesetzen großer Zahlen, für die zusätzlich die Konvergenzrate kontrollierbar ist.

Die für die Konvergenzaussagen benötigte Theorie von Glivenko und Cantelli benötigt die Voraussetzung, dass die Daten unabhängig voneinander erhoben wurden. Aus diesem Grund kann die bisherige Theorie der empirischen Risiko-Minimierung beispielsweise auf die große Klasse zeitabhängiger Daten nicht angewendet werden. Allerdings ist eine solche Anwendung wünschenswert, da der Anteil der dynamisch erhobenen Daten stetig zunimmt, so dass die Nutzung stochastischer Prozesse für zahlreiche statistische Analysen unverzichtbar ist.
Dies ist die Motivation, nach solchen Abhängigkeitsstrukturen in den Daten zu suchen, für die geeignete wahrscheinlichkeitstheoretische Gesetzmäßigkeiten gültig sind,

die die Anwendung der Gesetze der großen Zahlen weiterhin erlauben. Für stochastische Prozesse gibt es eine Vielzahl von Abhängigkeitskonzepten, für die solche Gesetze Gültigkeit besitzen (Doob, 1953). Allerdings muss dabei unterschieden werden zwischen ereignisorientierten und prozessorientierten Ansätzen zur Darstellung der Abhängigkeiten in stochastischen Prozessen (Davidson, 1994). Die Prozessorientiertheit erlaubt die Modellierung der Abhängigkeit über bedingte Wahrscheinlichkeiten. Dies hat den Vorteil, dass die Vergangenheit nur noch als ein Ereignis Auswirkungen auf den Erhebungsprozess hat, wie dies beispielsweise bei Zeitreihen der Fall ist. In dieser Arbeit wird das klassische Konzept der Martingale und das von McLeish (1975) entwickelte und hauptsächlich in der Ökonometrie angewandte Mixingal-Konzept genutzt, um das Prinzip der empirischen Risiko-Minimierung auf dynamische Datenstrukturen zu verallgemeinern.
Damit auch bei obigen Abhängigkeitsstrukturen Konsistenz für das Prinzip gewährleistet ist, werden die Abhängigkeiten über die empirischen Verluste modelliert. Dieses Vorgehen stellt die breite Anwendbarkeit des ERM-Prinzips sicher. Die direkte Modellierung der Daten als abhängige Beobachtungen, ist dagegen in der Regel nicht möglich, da die benutzten Abhängigkeitsstrukturen nur unter sehr wenigen Verlustfunktionalen invariant erhalten bleiben.

Die vorliegende Arbeit ist in drei Teile geteilt. Zuerst wird die Theorie der statistischen Lerntheorie mit Blick auf die empirische Risiko-Minimierung vorgestellt. Danach werden verschiedene Abhängigkeitsstrukturen in den Daten beschrieben. Der dritte Teil befasst sich dann mit der Verknüpfung der beiden ersten Teile zu einem Prinzip der empirischen Risiko-Minimierung bei dynamischen Datenstrukturen.
Im zweiten Kapitel zur statistischen Lerntheorie wird zuerst das Lernproblem und das Verfahren der empirischen Risiko-Minimierung in allgemeiner Weise vorgestellt. Danach wird ein Konsistenz-Konzept eingeführt, welches sicherstellt, dass die empirische Risiko-Minimierung als Prinzip für eine ganze Klasse von Lernproblemen gute Ergebnisse liefert. Bedingungen für die Konsistenz und für die Kontrolle der Konvergenzrate stellen dabei die Generalisierungsfähigkeit des Prinzips sicher.
Für die benötigten Konsistenz- und Konvergenz-Ergebnisse bei Abhängigkeiten in Datensätzen werden im dritten Kapitel Abhängigkeitsstrukturen in allgemeiner Form für

stochastische Prozesse eingeführt, bevor dann ausführlich die Ergebnisse für Martingale und Mixingale vorgestellt und bewiesen werden. Das Hauptaugenmerk liegt dabei, bedingt durch die Bedürfnisse bei der empirischen Risiko-Minimierung, auf den Gesetzen der großen Zahlen. Außerdem werden Bedingungen angegeben, unter denen diese Gesetze eine schnelle, d. h. exponentielle, Konvergenzrate besitzen.
Im vierten Kapitel können die Ergebnisse aus dem dritten Kapitel auf die statistische Lerntheorie angewendet werden. Es kann nachgewiesen werden, dass das Prinzip der empirischen Risiko-Minimierung auch bei dynamischen Datenstrukturen, speziell bei Martingal- oder Mixingal-Strukturen, gültig ist. Dabei werden hinreichende Bedingungen für die Konsistenz des ERM-Prinzips und für die Kontrolle der Konvergenzrate angegeben. Im letzten Kapitel werden die Auswirkungen der Ergebnisse des vierten Kapitels auf das statistische Lernen im Rahmen eines kurzen Ausblicks formuliert.

2 Statistische Lerntheorie

Aus dem Wunsch unbekannte Zusammenhänge in großen, hochdimensionalen Datensätzen mit komplexen Strukturen und hohem Informationsgehalt zu entdecken, resultiert die Notwendigkeit, mit wenigen Vorkenntnissen und Annahmen an die Struktur bzw. die Verteilung in den Daten geeignete Modelle zu finden, die die Datenstruktur wiedergeben und somit die wichtigen Muster aus den Daten extrahieren (vgl. Hastie, Tibshirani und Friedman, 2001). Die statistische Lerntheorie ist ein geeignetes Konzept, um funktionale Zusammenhänge in einer gegebenen Menge von Daten zu entdecken. Dies ist ein sehr generelles Problem und deckt zahlreiche statistische Problemstellungen ab, beispielsweise aus den Bereichen der Klassifikation oder der Regression. In diesem Kapitel wird eine allgemeine Definition des statistischen Lernproblems gegeben und mit dem Prinzip der empirischen Risiko-Minimierung ein generelles Konzept zur Lösung dieses Problems vorgestellt. Dieses Prinzip stellt dabei den konzeptionellen, theoretischen Teil der statistischen Lerntheorie dar. Die praktische Anwendung wird mit Hilfe geeigneter Algorithmen ermöglicht (Vapnik, 1995).

2.1 Das statistische Lernproblem

Das statistische Lernproblem ist bemerkenswert einfach zu stellen, ein generelles Konzept zur allgemeinen Lösung eines Lernproblems ist dagegen komplex. Einfache Lösungen sind in der Regel nur unter konkreten vereinfachenden Annahmen möglich. Die in sehr vielen statistischen Problemstellungen ähnliche Situation wird durch eine oft große Anzahl von gemeinsam erhobenen Variablen charakterisiert. Der eine Teil der Variablen sind dabei die *unabhängigen*, *erklärenden* bzw. *Input-Variablen*, der andere Teil sind die *abhängigen*, *Response-* bzw. *Output-Variablen* (vgl. Friedman, 1994). Das Ziel ist es, mit Hilfe eines erhobenen Datensatzes einen Zusammenhang zwischen den Variablen aufzudecken. Für Vapnik (1995) ist das Lernproblem die Aufgabe, eine Funktion zu finden, die einen bestehenden Zusammenhang in einer endlichen Menge von Beobachtungen wieder gibt. Mit Hilfe der *statistischen Lerntheorie* werden Lösungsansätze entwickelt, solche Zusammenhänge für gegebene Paare von In- und Output-Variablen zu finden, sie beschäftigt sich mit dem Lernen aus empirischen Daten anhand mathe-

matisch fundierter Algorithmen und Methoden (vgl. Schölkopf, 1998). Diese Art des Lernens wird in der Statistik allgemein als *Schätzung funktionaler Zusammenhänge* und in der Mathematik als *Funktionenapproximation* gesehen. Der Ansatz von Vapnik (1995) zur statistischen Lerntheorie kann besonders auf die Fälle angewandt werden, in denen eine funktionale Modellierung durch sachbezogenes Wissen schwierig oder gar unmöglich ist, denn ein gewisser funktionaler Zusammenhang wird zwar vorausgesetzt, muss aber nicht notwendig spezifiziert werden.

Die Formulierung des Lernproblems ist sehr generell. So unterschiedliche Methoden wie Diskriminanzanalyse oder Mustererkennung mit kategoriellen Response-Variablen, sowie Regressionsanalyse oder Dichteschätzung mit metrischem Output werden von dieser einfachen Formulierung abgedeckt. Erst eine nähere Eingrenzung führt zu einer Spezifizierung einer konkreten Problemstellung.
Das allgemeine Modell für die Suche nach einem Zusammenhang zwischen Variablen geht von einem Paar von Zufallsvariablen (X, Y) auf einem beliebigen Kreuzprodukt von Variablenmengen $\mathcal{X} \times \mathcal{Y}$ mit $X \in \mathcal{X}$ und $Y \in \mathcal{Y}$ aus, wobei sowohl der Variablen X als auch der Variablen Y eine Verteilung P_X sowie P_Y mit unbekannten Verteilungsfunktionen $F_X(x)$ bzw. $F_Y(y)$ und auch eine gemeinsame Verteilung $P_{X,Y}$ mit Verteilungsfunktion $F_{X,Y}(x, y)$ unterstellt wird. Der Zusammenhang wird dabei in sehr allgemeiner Weise nur durch die bedingte Verteilung $P_{Y|X}$ mit bedingter Verteilungsfunktion $F_{Y|X}$ modelliert. Es wird angenommen, dass sich die bedingte Verteilung aus der gemeinsamen Verteilung durch folgenden Zusammenhang ergibt:

$$P_{X,Y} = P_{Y|X} P_X.$$

Für die unbekannte bedingte Verteilungsfunktion folgt somit

$$F_{X,Y}(x, y) = F_{Y|X}(y|x) F_X(x).$$

Dagegen stellt eine Modellierung eines funktionalen Zusammenhangs zwischen Y und X eine Einschränkung an die zugelassenen Strukturen in den Variablen dar. Deshalb ist es beispielsweise sinnvoll, zusätzlich einen Zufallsfehler u additiv zu modellieren. Dadurch ist der funktionale Zusammenhang nicht deterministisch. Insgesamt ergibt sich ein statistisches Modell mit einer Funktion $f' : \mathcal{X} \to \mathcal{Y}$ aus einer Funktionenmenge

$\mathcal{F}'$ und einer Zufallsvariablen u in der Form

$$Y = f'(X) + u,$$

wobei der Zufallsfehler u unabhängig von X und identisch verteilt ist mit Erwartungswert $\mathrm{E}(u) = 0$. Dieses additive Fehler-Modell unterstellt, dass alle Abweichungen der Realität von dem deterministischen Modell $Y = f'(X)$ in dem Fehler u zusammengefasst werden, also alle nicht erfassten Auswirkungen auf die Variable Y ebenso wie der Messfehler. Die Einschränkung dieses Modells im Vergleich zu dem generellen Modell über die bedingte Verteilung ergibt sich daraus, dass in diesem Ansatz

$$f'(X) = \mathrm{E}(Y|X)$$

gilt und damit die Verteilung $P_{Y|X}$ durch X nur über den bedingten Erwartungswert $f'(X)$ spezifiziert wird. Trotzdem ist das additive Fehler-Modell häufig ein geeignetes Modell, um einen Zusammenhang zwischen Variablen darzustellen. Da die Realität nur bei einem rein deterministischen Zusammenhang aus den Daten exakt erfasst werden kann, wird in der Regel nur eine „gute“ Approximation des funktionalen Zusammenhangs aus den Daten „schätzbar“ sein.

Zur Schätzung oder Approximation des Zusammenhangs zwischen Input- und Output-Variablen werden die Realisationen des Variablenpaars $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ genutzt. Dazu wird ein Paar $(x, y) \in \mathcal{X} \times \mathcal{Y}$ gemäß der gemeinsamen Verteilung $P_{X,Y}$ mit Verteilungsfunktion $F_{X,Y}(x, y)$ gezogen bzw. generiert. Bei jeder Durchführung des Experiments entsteht so eine *Beobachtung* bzw. ein sogenanntes *Trainingsbeispiel* (x, y). Mit Hilfe von n zufällig gezogenen Beobachtungen

$$(x_1, y_1), \ldots, (x_n, y_n)$$

soll der Zusammenhang zwischen den erklärenden Variablen X und den Response-Variablen Y geschätzt werden. Die Menge der n Beobachtungen ist der *Trainingsdatensatz* oder kurz *Datensatz*. Da die Beobachtungen (x_i, y_i), $i = 1, \ldots, n$, durch die Art der Ziehung identisch verteilt sind, also Kopien des Variablenpaars (X, Y) sind und die Approximation des Zusammenhangs zwischen diesen Variablen durch einen funktionalen Zusammenhang ausreichend ist, wenn nur die Funktionenmenge $\mathcal{F}$ geeignet

und groß genug gewählt wird, kann im Weiteren ein funktionaler Zusammenhang

$$y = f(x) \text{ mit } f \in \mathcal{F}$$

für die Beobachtungspaare (x_i, y_i), $i = 1, \ldots, n$, angenommen werden. Von großer Bedeutung ist die Wahl der Menge $\mathcal{F}$, damit die „beste“ Lösung f aus dieser Menge den Zusammenhang auch ausreichend gut annähert. Dabei muss diese Menge nicht notwendig mit der Ausgangsmenge $\mathcal{F}'$ übereinstimmen, noch eine Teil- oder Obermenge sein. Es kann durchaus sinnvoll sein, eine komplizierte Funktion mit einer Lösung aus einer Menge einfacherer Funktionen zu approximieren. Beispielsweise liefert eine Funktion aus der Menge der Polynome häufig eine gute Lösung für weitaus kompliziertere Abbildungen.
Eine weitere Anpassung an das Lernziel erfolgt durch die Wahl eines geeigneten Maßes für die Güte des Lernergebnisses. Nur mit einem adäquaten Maß können die Lernfehler durch eine geeignete Verlustfunktion bestraft werden. Zum Beispiel ist es sinnvoll, bei einem Diskriminierungsproblem mit kategorieller Response-Variable falsche Zuordnungen mit immer dem gleichen Verlust zu bestrafen. Richtige Zuordnungen sollen dagegen keinen Verlust produzieren. Dies führt also zu einer Verlustfunktion und damit auch zu einem Maß, mit nur zwei Ausprägungen für „wahr“ bzw. korrekte Entscheidung und „falsch“ bzw. falsche Entscheidung. Bei einem Regressionsproblem sollten andererseits Fehler, also hier Abweichungen von dem wahren Wert des Outputs, möglicherweise proportional zur Größe der Abweichung bestraft werden.

Grundsätzlich heißt Statistisches Lernen, dass während eines Lernprozesses aus einer Menge vorgegebener Funktionen eine Funktion als Lösung im Sinne der Problemstellung gewählt wird. Die Regel, nach der diese Funktion gewählt wird, gehört zu den zentralen Schritten innerhalb der statistischen Lerntheorie (Vapnik, 1999). Das Modell zum statistischen Lernen durch Beobachtungen (Beispiele) wird wie folgt definiert.

2.1 Definition *Statistisches Lernen*
Seien $X \in \mathcal{X}$ und $Y \in \mathcal{Y}$ Zufallsvariablen, verteilt gemäß der gemeinsamen Verteilung $P_Z = P_{X,Y} = P_{Y|X} P_X$ aus einer Familie von Verteilungen $\mathcal{P}$, mit zugehörigen Beobachtungen (x_i, y_i), $i = 1, \ldots, n$. Sei weiter

$$\mathcal{F} = \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$$

eine Menge von Funktionen. Dann ist das *statistische Lernen* die Aufgabe, aus der gegebenen Menge $\mathcal{F}$ unter Ausnutzung der Beobachtungen diejenige Funktion

$$f^\star : \mathcal{X} \to \mathcal{Y}$$

zu wählen, die den über die Verteilungsannahme festgelegten Zusammenhang zwischen X und Y durch $y = f^\star(x)$ darstellt.

Dabei muss die Funktion $f^\star \in \mathcal{F}$, die den Zusammenhang zwischen den Zufallsvariablen X und Y exakt widerspiegelt, nicht existieren. In diesem Fall wird in der Menge $\mathcal{F}$ nach der Funktion gesucht, die den Zusammenhang am besten approximiert.
Die Definition des statistischen Lernproblems lässt sich vereinfachen, indem die Annahme an die Funktionenmenge durch eine Parametrisierung eingeschränkt wird. Sei dazu Λ eine feste, vorgegebene Parametermenge, so dass Funktionen aus der Menge $\mathcal{F}$ durch $f(\cdot, \alpha)$, $\alpha \in \Lambda$, dargestellt werden. Die Charakterisierung der Funktion f durch einen Parameter α ist eindeutig, die Menge $\mathcal{F}$ wird durch die Menge Λ der möglichen Parameter α ebenfalls eindeutig dargestellt:

$$\mathcal{F} = \{f(\cdot, \alpha), \alpha \in \Lambda\}.$$

Vereinfacht wird die Funktionenmenge $\mathcal{F}$ im Folgenden auch durch $f(\cdot, \alpha)$, $\alpha \in \Lambda$, beschrieben. Die mit der Parametrisierung verbundene Einschränkung der Definition des Lernproblems hängt dann von der Komplexität der Menge Λ ab. Einfache Formen sind Parametrisierungen durch reelle Skalare oder Vektoren, dann gilt $\Lambda \subseteq \mathbb{R}$ bzw. $\Lambda \subseteq \mathbb{R}^k$, aber ebenso kann Λ Bedingungen wie Stetigkeit oder Konvexität enthalten, oder Λ wird als die Menge aller Polynome, bzw. als Menge der Polynome bis zu einem gewissen Grad definiert. Es sind sogar abstrakte Bedingungen, die nicht geschlossen darstellbar sind, vorstellbar. Dadurch ist diese Form der Parametrisierung keine Einschränkung der Allgemeingültigkeit des statistischen Lernens, da sich alle Formen von

Funktionen durch einen geeigneten Parameter $\alpha \in \Lambda$ darstellen lassen.
Im weiteren Verlauf soll das statistische Lernproblem auf reelle Datensätze beschränkt werden. Das heißt, zu jedem reellen erklärenden Vektor $x \in \mathbb{R}^d$ wird unter der Annahme der bedingten Verteilung $F_{Y|X}(y|x)$ eine reelle Response $y \in \mathbb{R}$ beobachtet. Dies ist für die Einführung der statistischen Lerntheorie keine Beschränkung, alle weiteren Ergebnisse lassen sich auf beliebige Variablenmengen $\mathcal{X} \times \mathcal{Y}$ übertragen, solange auf diesen Mengen eine geeignete Metrik existiert.

2.2 Definition *Statistisches Lernen bei reellwertigen Daten*
Sei $X \in \mathbb{R}^d$ Zufallsvektor und $Y \in \mathbb{R}$ Zufallsvariable, verteilt gemäß der gemeinsamen Verteilungsfunktion $F_Z(z) = F_{X,Y}(x, y) = F_{Y|X}(y|x)F_X(x)$ aus der Familie der Verteilungsfunktionen $\mathbb{F}$, mit zugehörigen Beobachtungen (x_i, y_i), $i = 1, \ldots, n$. Sei $f(\cdot, \alpha)$, $\alpha \in \Lambda$, eine Menge von Funktionen. Dann ist das *statistische Lernen bei reellwertigen Daten* die Aufgabe, unter Ausnutzung der Daten aus der gegebenen Menge Λ denjenigen Parameter $\alpha^\star \in \Lambda$ zu wählen, so dass die Funktion

$$f(\cdot, \alpha^\star) : \mathbb{R}^d \to \mathbb{R}$$

den über die Verteilungsannahme festgelegten Zusammenhang zwischen X und Y durch $y = f(x, \alpha^\star)$ am besten approximiert.

Mit der Einführung des reellen statistischen Lernproblems ist es sehr viel einfacher, Gütekriterien für die Exaktheit der Approximation durch die Funktion $f(\cdot, \alpha^\star)$ festzulegen, da es so ausreicht, Kriterien auf den reellen Zahlen zu wählen. Um im geeigneten Sinne den Abstand zwischen den Responsevariablen y_i, $i = 1, \ldots, n$, und der gewählten Approximation $f(x_i, \alpha^\star)$ zu messen, wird ein Abstandsmaß auf den reellen Zahlen benötigt, das der Aufgabenstellung bei der Funktionenapproximation entspricht und den Verlust widerspiegelt, der durch die Fehler bei der Approximation entsteht. Allgemein wird der *Verlust* definiert als eine Funktion $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ mit $L(y, f(x, \alpha))$, die der Responsevariablen y und der zugehörigen Approximation $f(x, \alpha)$ eines Tupels (x, y) einen Abstand zuordnet. Durch die parametrische Darstellung des statistischen Lernproblems wird die Funktion $f(x, \alpha)$ eindeutig durch α bestimmt, so dass der Verlust als Funktion $Q : \mathbb{R}^{d+1} \times \Lambda \to \mathbb{R}$ des Tupels $z = (x, y) \in \mathbb{R}^{d+1}$ und des Parameters

$\alpha \in \Lambda$ dargestellt werden kann:

$$Q(z, \alpha) = L(y, f(x, \alpha)).$$

Der erwartete Verlust, das *Risiko*, ergibt sich aus

$$R(\alpha) = \int L(y, f(x, \alpha)) dF_{X,Y}(x, y) = \int Q(z, \alpha) dF_Z(z).$$

Das Ziel ist, den Parameter $\widetilde{\alpha}$ bzw. die Funktion $f(\cdot, \widetilde{\alpha})$ zu finden, die das Risiko-Funktional $R(\alpha)$ minimiert. Dies ist die beste funktionale Approximation auf der Menge von Funktionen $f(x, \alpha)$, $\alpha \in \Lambda$, für den Zusammenhang zwischen den Zufallsvariablen X und Y unter dem vorliegenden Verlust-Konzept. Die Idee dabei ist, dass bei Wahl einer geeigneten Verlustfunktion die Risiko-Minimierung eine Funktion $f(x, \alpha)$ liefert, die der besten Approximation $f(x, \alpha^\star)$ aus der Menge der Funktionen $\mathcal{F}$ nahe kommt. In der Situation, in der die Verteilung $P_Z = P_{X,Y}$ bzw. $F_Z(z) = F_{X,Y}(x, y)$ unbekannt ist, muss nutzbare Information, die auf die Verteilung und damit auf den funktionalen Zusammenhang zwischen X und Y schließen lässt, aus den Beobachtungen $z_1, \ldots, z_n$ mit $z_i = (x_i, y_i) \in \mathbb{R}^{d+1}$ gewonnen werden. Daraus ergibt sich das allgemeingültige statistische Lernproblem.

2.3 Definition *Statistisches Lernproblem*
Seien $z_1, \ldots, z_n$ Beobachtungen mit $z_i = (x_i, y_i) \in \mathbb{R}^{d+1}$ und verteilt gemäß der gemeinsamen, unbekannten Verteilung $F_Z(z) = F_{X,Y}(x, y) = F_{Y|X}(y|x) F_X(x)$. Sei $f(\cdot, \alpha)$, $\alpha \in \Lambda$, eine Menge von Funktionen und sei

$$Q(z, \alpha) = L(y, f(x, \alpha))$$

die Verlustfunktion. Dann ist das *statistische Lernproblem* beschrieben durch die Aufgabe, aus der gegebenen Parametermenge Λ, unter Ausnutzung der Beobachtungen $z_1, \ldots, z_n$ denjenigen Parameter $\widetilde{\alpha} \in \Lambda$ zu finden, der das Risiko

$$R(\alpha) = \int Q(z, \alpha) dF_Z(z)$$

minimiert:

$$\widetilde{\alpha} = \arg \inf_{\alpha \in \Lambda} R(\alpha) = \arg \inf_{\alpha \in \Lambda} \int Q(z, \alpha) dF_Z(z).$$

Diese Formulierung des statistischen Lernproblems ist sehr generell, da keinerlei Eigenschaften an die Struktur der Menge der Funktionen $f(\cdot, \alpha)$, $\alpha \in \Lambda$, vorgegeben werden und für die Beobachtungen $z_1, \ldots, z_n$ wird nur die Annahme der identischen Verteilung gemäß einer Verteilungsfunktion F_Z vorausgesetzt. Das Risikofunktional ist ebenfalls sehr generell formuliert, so dass beliebige Verlust-Konzepte darstellbar sind. Damit sind zahlreiche statistische Problemstellungen abgedeckt, unter anderen zwei der wichtigsten Aufgaben, die Klassifikation bzw. die Mustererkennung und die Regression (Evgeniou et al., 2002), die als Beispiele dienen können.

2.4 Beispiel *Klassifikation*
Sei für das Paar $(x, y) \in \mathbb{R}^{d+1}$ die Outputvariable y beschränkt auf die Menge $\{0, 1\}$ und sei der Output über die bedingte Verteilung $F_{Y|X}(y|x)$ abhängig von dem d-dimensionalen Inputvektor x, verteilt gemäß $F_X(x)$. Sei $f(\cdot, \alpha)$, $\alpha \in \Lambda$, die Menge der Indikatorfunktionen mit Ausprägungen aus der Menge $\{0, 1\}$ und sei die Verlustfunktion definiert als

$$L(y, f(x, \alpha)) = \begin{cases} 0, & \text{falls} \quad y = f(x, \alpha) \\ 1, & \text{falls} \quad y \neq f(x, \alpha). \end{cases}$$

Das Klassifikationsproblem ist die Minimierung des Risikofunktionals $R(\alpha)$ bezüglich der Menge der Indikatorfunktionen, wobei die gemeinsame Verteilung

$$F_Z(z) = F_{X,Y}(x, y)$$

unbekannt ist, aber eine Menge von Beobachtungspaaren $z_1, \ldots, z_n$ mit

$$z_i = (x_i, y_i) \in \mathbb{R}^d \times \{0, 1\}, \quad i = 1, \ldots, n,$$

vorliegen. Auf Grund der besonderen Form der Verlustfunktion gibt das Risiko $R(\alpha)$ die Wahrscheinlichkeit des Klassifikationsfehlers, also wenn der Output y und der Wert der Indikatorfunktion verschieden sind, an. Damit kann das Klassifikationsproblem auch als die Minimierung der Wahrscheinlichkeit des Klassifikationsfehlers betrachtet werden.

Das Problem der Klassifikation oder auch der Mustererkennung ist ein einfaches statistisches Lernproblem, da die Verlustfunktion eine Indikatorfunktion ist, also nur zwischen zwei Ausprägungen unterschieden werden muss. Damit ergibt sich die einfache

Berechnung des Risikos als Wahrscheinlichkeit. Gute Ergebnisse bei Anwendung der statistischen Lerntheorie im Klassifikationskontext gibt es beispielsweise bei der Bilderkennung (vgl. Li et al., 2002). Wesentlich komplexer wegen der Modellierung eines stetigen Verlusts ist das Regressionsproblem.

2.5 Beispiel *Regression*

Seien zwei Mengen von Elementen $\mathcal{X} \subseteq I\!R^d$ und $\mathcal{Y} \subseteq I\!R$ durch einen stochastischen Zusammenhang in dem Sinne verbunden, dass jedem Vektor $x \in \mathcal{X}$ durch eine bedingte Verteilung $F_{Y|X}(y|x)$ ein Skalar $y \in \mathcal{Y}$ zugeordnet wird. Diese bedingte Verteilungsannahme drückt dann den stochastischen Zusammenhang zwischen x und y für jedes Paar $(x, y) \in \mathcal{X} \times \mathcal{Y}$ aus. Werden Vektoren x zufällig als Beobachtungen aus der Menge $\mathcal{X}$ gemäß einer Verteilung $F_X(x)$ gezogen, so ergeben sich die Werte für y mittels $F_{Y|X}(y|x)$ als zufällige Experimente und es existiert eine gemeinsame Verteilung $F_{X,Y}(x, y) = F_{Y|X}(y|x)F_X(x)$ nach der die Paare

$$(x_1, y_1), \ldots, (x_n, y_n)$$

als Beobachtungen gezogen werden. Soll der stochastische Zusammenhang mit Hilfe dieser Beobachtungen geschätzt werden, bedeutet dies, dass die unbekannte, bedingte Verteilung $F_{Y|X}(y|x)$ geschätzt werden muss. Allerdings ist in der Regel die Bestimmung des funktionalen Zusammenhangs zwischen den Elementen der Mengen $\mathcal{X} \in I\!R^d$ und $\mathcal{Y} \in I\!R$ ausreichend, dies führt zu dem Problem der Schätzung des bedingten Erwartungswerts

$$r(x) = \mathrm{E}(Y|X = x) = \int_{\mathcal{Y}} y dF_{Y|X}(y|x).$$

Diese zu schätzende Funktion $r(x)$ heißt *Regression*, die Approximation dieser Funktion aus einer Menge von Funktionen $f(\cdot, \alpha)$, $\alpha \in \Lambda$, ist die *Regressionsschätzung*, d. h. unter L_2-Norm $\|\cdot\|_2$ muss

$$\int (f(x, \alpha) - r(x))^2 dF(x, y)$$

minimiert werden.

Es lässt sich einfach zeigen, dass bei Wahl einer quadratischen Verlustfunktion

$$L(y, f(x, \alpha)) = (y - f(x, \alpha))^2$$

das Problem der Regressionsschätzung auf ein statistisches Lernproblem zurückführbar ist. Existieren die zweiten Momente von y, von der Regression $r(x)$ und von den Funktionen der Menge $f(\cdot, \alpha)$, $\alpha \in \Lambda$, d. h.

$$\int y^2 dF_{X,Y}(x,y) < \infty, \quad \int (r(x))^2 dF_{X,Y}(x,y) < \infty$$
$$\text{und} \quad \int (f(x,\alpha))^2 dF_{X,Y}(x,y) < \infty,$$

kann das Risiko aufgespalten werden in

$$\begin{aligned} R(\alpha) &= \int (y - f(x,\alpha))^2 dF(x,y) \\ &= \int (y - r(x))^2 dF(x,y) + \int (f(x,\alpha) - r(x))^2 dF(x). \end{aligned}$$

Da der erste Summand in der letzten Gleichung nicht vom Parameter α abhängt, ist die Funktion $f(x, \alpha^\star)$, die das Risikofunktional minimiert entweder die Regression $r(x)$, falls $r(x) \in \{f(x,\alpha), \alpha \in \Lambda\}$ gilt, oder die Funktion $f(x, \alpha^\star)$ ist diejenige aus der Menge, die bezüglich der L_2-Norm den kleinsten Abstand zu $r(x)$ hat, falls die Regression nicht zu $f(x,\alpha)$, $\alpha \in \Lambda$, gehört.

Die Verlustfunktion $Q(z,\alpha) = L(y, f(x,\alpha))$ kann im Regressionskontext beim statistischen Lernproblem beliebige nichtnegative Werte annehmen, im Gegensatz zum Klassifikationsproblem. Ein Spezialfall der Regression ist die Signalextraktion, auch hier findet die statistische Lerntheorie ihre Anwendung (vgl. Cherkassky und Shao, 2001). Ebenso wie bei der klassischen Regression kann auch die L^1-*Regression* als statistisches Lernproblem dargestellt werden. In diesem Fall wird

$$L(y, f(x,\alpha)) = |y - f(x,\alpha)|$$

als Verlustfunktion gewählt.

Grundsätzlich ist das generelle Problem im Kontext des statistischen Lernens die Minimierung des Risikofunktionals

$$R(\alpha) = \int Q(z,\alpha) dF_Z(z)$$

bei gegebener Verlustfunktion $Q(z,\alpha)$ und unbekannter Verteilung P_Z bzw. Verteilungsfunktion $F_Z(z)$ unter Ausnutzung empirischer Daten $z_1, \ldots, z_n$ mit $z_i = (x_i, y_i)$.

Die direkte Minimierung ist nicht möglich, aber aus den Beobachtungen lässt sich die empirische Verteilungsfunktion $F_{emp}(z)$ berechnen. Diese Schätzung kann in das Risikofunktional $R(\alpha)$ an Stelle der unbekannten Verteilung eingesetzt werden, so dass die Berechnung eines empirischen Risikos möglich wird:

$$R_{emp}(\alpha) = \int Q(z, \alpha) dF_{emp}(z), \quad \alpha \in \Lambda.$$

Allerdings ist es nicht notwendig die empirische Verteilung explizit zu berechnen, denn das empirische Risiko kann durch

$$R_{emp}(\alpha) \; = \; \int Q(z, \alpha) dF_{emp}(z) \; = \; \frac{1}{n} \sum_{i=1}^{n} Q(z_i, \alpha), \quad \alpha \in \Lambda,$$

direkt errechnet werden. Dieses induktive Vorgehen zur Lösung statistischer Lernprobleme führt zum Prinzip der empirischen Risiko-Minimierung.

2.6 Definition *empirische Risiko-Minimierung (ERM)*
In einem statistischen Lernproblem mit Risiko

$$R(\alpha) = \int Q(z, \alpha) dF_Z(z)$$

seien $z_1, \ldots, z_n$ Beobachtungen mit $z_i = (x_i, y_i) \in I\!R^{d+1}$ und verteilt gemäß der gemeinsamen, unbekannten Verteilung $F_Z(z)$. Dann heißt

$$R_{emp}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} Q(z_i, \alpha), \quad \alpha \in \Lambda$$

empirisches Risikofunktional bzw. *empirisches Risiko*. Das Prinzip, $R_{emp}(\alpha)$ an Stelle von $R(\alpha)$ zu minimieren, heißt *empirische Risiko-Minimierung*. Mit $\widetilde{\alpha}_n \in \Lambda$ bezeichne den Parameter, der das empirische Risiko minimiert:

$$\widetilde{\alpha}_n = \arg \inf_{\alpha \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} Q(z_i, \alpha).$$

Das empirische Risiko ist auf Basis der Daten $z_1, \ldots, z_n$ explizit berechenbar und minimierbar. Der Parameter $\widetilde{\alpha}_n \in \Lambda$ kann somit als eine Approximation für $\widetilde{\alpha} \in \Lambda$, der das Risiko $R(\alpha)$ minimiert, aufgefasst werden. Um gewährleisten zu können, dass die

Funktion $f(\cdot, \widetilde{\alpha}_n)$ eine gute Darstellung des gesuchten funktionalen Zusammenhangs ist, müssen geeignete Bedingungen eingeführt werden, wann der Abstand zwischen den Verlusten $Q(x, \widetilde{\alpha}_n)$ und $Q(x, \widetilde{\alpha})$ klein ist. Dies führt zum Problem einer geeigneten Abstandsschätzung zwischen minimalem Risiko und dem Risiko bezüglich der aus der Minimierung des empirischen Risikos gewählten Funktion. Wenn also die Funktion $Q(z, \widetilde{\alpha}_n)$ durch empirische Risiko-Minimierung gewählt wurde, muss gefordert werden, dass mit einer Wahrscheinlichkeit $(1 - \eta)$, $0 < \eta \leq 1$, der Wert des Risikos $R(\widetilde{\alpha}_n)$ den Wert des kleinsten möglichen Risikowert $\inf_{\alpha \in \Lambda} R(\alpha)$ bei vorgegebener Menge von Funktionen höchstens um einen Wert $\mathcal{E}$ übersteigt. Dabei hängt $\mathcal{E}$ nur von der Stichprobengröße n, von der Wahrscheinlichkeit η und von einem Parameter H^Λ ab, der die generellen Entropie-Eigenschaften der Menge der Funktionen $f(\cdot, \alpha)$, $\alpha \in \Lambda$, charakterisiert, also deren Komplexität bzw. Größe. Mögliche Definitionen eines solchen Parameters werden in Abschnitt 2.3 eingeführt. Insgesamt muss für das Prinzip der empirischen Risiko-Minimierung garantiert werden, dass mit Wahrscheinlichkeit $(1 - \eta)$ die Ungleichung

$$R(\widetilde{\alpha}_n) - \inf_{\alpha \in \Lambda} R(\alpha) \leq \mathcal{E}(n, \eta, H^\Lambda)$$

gilt. Die $\mathcal{E}$-Nähe des geschätzten Risikos zum minimalen Risiko garantiert die Nähe der Funktion $f(\cdot, \widetilde{\alpha}_n)$ zu der ursprünglich gesuchten Funktion $f(\cdot, \widetilde{\alpha})$ in dem jeweiligen statistischen Lernproblem. Dabei gilt für jedes Problem bei der vorgegebenen Metrik und dem Verlust ein spezielles Konzept, um Nähe zu definieren. Für das Regressionsproblem ergibt sich daraus beispielsweise:

2.7 Beispiel *Regression*
Sei im statistischen Lernproblem zur Regression die Regressionsfunktion $r(x) = f(x, \widetilde{\alpha})$ aus der Menge der Funktionen $f(x, \alpha)$, $\alpha \in \Lambda$, und sei $f(x, \widetilde{\alpha}_n)$ die Schätzung mittels der empirischen Risiko-Minimierung, so dass für das Risiko $R(\widetilde{\alpha}_n)$ $\mathcal{E}$-Nähe gilt:

$$R(\widetilde{\alpha}_n) - \inf_{\alpha \in \Lambda} R(\alpha) \leq \mathcal{E}.$$

Dann ist die Funktion $f(x, \widetilde{\alpha}_n)$ $\sqrt{\mathcal{E}}$-nah zur Regressionsfunktion unter Nutzung der euklidischen Norm (vgl. Vapnik, 1998):

$$\|f(x, \widetilde{\alpha}_n), r(x)\|_2 = \sqrt{\int (f(x, \widetilde{\alpha}_n) - r(x))^2 dF(x)} \leq \sqrt{\mathcal{E}}.$$

Die Minimierung des Abstandes der Schätzung $f(x, \widetilde{\alpha}_n)$ zur optimalen Funktion $f(x, \widetilde{\alpha})$ aus der Menge der Funktionen $f(\cdot, \alpha)$, $\alpha \in \Lambda$, kann also auf das Problem zurückgeführt werden, den Abstand der zugehörigen Risiken zu minimieren, also das kleinste $\mathcal{E}$ in der Ungleichung

$$R(\widetilde{\alpha}_n) - \inf_{\alpha \in \Lambda} R(\alpha) \leq \mathcal{E}(n, \eta, H^{\Lambda})$$

bei fester vorgegebener Stichprobengröße zu finden. Dazu muss zuerst ein geeignetes *Konzept zur Konsistenz des Lernprozesses* entwickelt werden. Dieses Konzept muss gewährleisten, dass der statistische Lernprozess gute asymptotische Ergebnisse bezüglich der Approximation der Funktion $f(\cdot, \widetilde{\alpha})$ liefert. Das heißt, falls die Anzahl der Beobachtungen n gegen unendlich strebt, soll das Risiko $R(\widetilde{\alpha}_n)$ gegen das minimale Risiko $\inf_{\alpha \in \Lambda} R(\alpha)$, also $\mathcal{E}(n, \eta, H^{\Lambda})$ gegen Null konvergieren. Wie in Abschnitt 2.2 gezeigt wird, greifen die klassischen Konsistenzkonzepte hier zu kurz. Gleichzeitig müssen nicht nur notwendige sondern auch hinreichende Bedingungen für diese spezielle Konsistenz des Lernprozesses nachgewiesen werden. Damit ist sicher gestellt, dass die entwickelte Theorie konzeptuell nicht mehr verbessert werden kann.

Ein weiterer wichtiger Schritt ist die *Abschätzung der Konvergenzrate des Lernprozesses*, um eine ausreichende Konvergenzgeschwindigkeit zu erreichen. Nur die dazu entwickelten nicht-asymptotischen oberen Schranken für die asymptotische Konsistenzbedingung ermöglichen die Anwendung der statistischen Lerntheorie auf eine endliche Menge von Beobachtung und liefern damit die Generalisierungsfähigkeit der Algorithmen, die auf dem Konzept der empirischen Risiko-Minimierung basieren. Gleichzeitig müssen die Schranken verteilungsfrei sein, also nicht von der unbekannten Verteilung $F_Z(z)$ abhängen, damit die Konstruktivität gesichert ist.
Die Schranken können direkt aus den notwendigen und hinreichenden Bedingungen für die Konsistenz des Lernprozesses abgeleitet werden und können für das induktive Prinzip der *strukturellen Risiko-Minimierung (SRM)*, das die Kontrolle der Konvergenzrate des Lernprozesses erlaubt, genutzt werden. Dies funktioniert über den Trade-Off zwischen der Komplexität der Funktionen aus der Menge $f(x, \alpha)$, $\alpha \in \Lambda$, und dem Wert des empirischen Risikos, der mit den jeweiligen Funktionen erreicht werden kann. Zur Anwendung der statistischen Lerntheorie auf konkrete Datensituationen müssen *Al-*

gorithmen für den Lernprozess entwickelt werden, die das Konzept der strukturellen Risiko-Minimierung gewährleisten. Um das Risiko bei der Funktionenapproximation gemäß des SRM-Prinzips zu minimieren, müssen solche Algorithmen sowohl die Minimierung des empirischen Risikos bei einer gegebenen Menge von Funktionen kontrollieren, als auch die Wahl der Menge von Funktionen mit möglichst optimalen Eigenschaften.

Die ersten beiden der oben vorgestellten vier zentralen Probleme der statistischen Lerntheorie werden in dieser Arbeit für unabhängige Daten, wie in Vapnik (1995) erstmals einheitlich zusammengefasst, und für abhängige Daten gleichermaßen ausführlich behandelt, da Abhängigkeitsstrukturen in den Beobachtungen vornehmlich auf die empirische Risiko-Minimierung Auswirkungen haben. Das SRM-Prinzip kann dagegen mit den veränderten Voraussetzungen des ERM-Prinzips in gleicher Weise genutzt werden, und wird deshalb ebenso wie die daraus resultierenden Algorithmen, wie die Support Vector Machine (SVM), nur kurz erläutert.
Eine ausführliche Darstellung des SRM-Prinzips und der algorithmischen Umsetzung kann der Monographie zur statistischen Lerntheorie von Vapnik (1998) entnommen werden. Eine gute Zusammenfassung, um das Konzept des SRM-Prinzips und der Support Vector Machine nachzuvollziehen, gibt Vapnik (1999) und das vierte Kapitel im Buch von Cherkassky und Mulier (1998). Für eine generelle Zusammenfassung zur Formulierung des statistischen Lernproblems und zur Einführung des statistischen Lernen sei auf das ausführliche Buch von Vidyasagar (1997) und auf den Artikel von Mendelson (2003) verwiesen.

2.2 Konsistenz-Konzept für die statistische Lerntheorie

Die grundlegende Voraussetzung für gute Ergebnisse der statistischen Lerntheorie ist, dass das ERM-Prinzip unter bestimmten Bedingungen gute Konsistenzeigenschaften besitzt. Die zentrale Aufgabe in diesem Sinne ist die Suche nach diesen Bedingungen. Zuvor muss allerdings ein Konzept entwickelt werden, das die Konsistenz des Risikos gegen das minimale Risiko unter Berücksichtigung der generellen Struktur einer Menge

von Funktionen $f(x, \alpha)$, $\alpha \in \Lambda$, gewährleistet, denn im Gegensatz zu klassischen Konsistenzforderungen für Folgen reeller Zahlen muss hier Konsistenz für ein Prinzip auf einer Funktionenmenge gelten. Dabei wird in diesem Kapitel auf die klassische *Konvergenz nach Wahrscheinlichkeit* $X_t \xrightarrow{Pr} X$ zurückgegriffen. Eine ausführliche Erläuterung dieses Konvergenzbegriffs wird zusammen mit der fast sicheren Konvergenz im dritten Kapitel gegeben.
Sei $Q(z, \widetilde{\alpha}_n)$ die Funktion, die unter Ausnutzung der Beobachtungen $z_1, \ldots, z_n$ das empirische Risiko

$$R_{emp}(\alpha) = \sum_{i=1}^{n} Q(z_i, \alpha), \quad \alpha \in \Lambda,$$

minimiert. Dann ist nach dem klassischen Konsistenz-Konzept das Prinzip der empirischen Risiko-Minimierung *konsistent* für die Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, und für die unbekannte Verteilungsfunktion $F_Z(z)$, falls

$$R(\widetilde{\alpha}_n) \xrightarrow[n\to\infty]{Pr} \inf_{\alpha\in\Lambda} R(\alpha)$$

$$\text{und} \quad R_{emp}(\widetilde{\alpha}_n) \xrightarrow[n\to\infty]{Pr} \inf_{\alpha\in\Lambda} R(\alpha)$$

gilt. Das bedeutet, das ERM-Prinzip ist konsistent, wenn mit der Folge von Funktionen $Q(z, \widetilde{\alpha}_n)$, $n = 1, 2, \ldots$, die mit diesem Prinzip berechnet werden, sowohl das Risiko als auch das empirische Risiko gegen den für die gegebene Menge von Funktionen minimal möglichen Risikowert konvergieren. Dies drückt die Notwendigkeit aus, dass die Risikowerte $R(\widetilde{\alpha}_n)$ mit der Folge der durch das ERM-Prinzip berechneten Parametern $\{\widetilde{\alpha}_n\}_{i=1}^n$, $\widetilde{\alpha}_n \in \Lambda$, gegen den minimal möglichen Risikowert konvergieren, und dass gleichzeitig dieser minimale Wert durch die empirischen Risikowerte $R_{emp}(\widetilde{\alpha}_n)$ mit der gleichen Folge von Parametern erreicht wird und damit berechenbar ist.

Das Ziel bei der Beschreibung der Konsistenzbedingungen für das ERM-Prinzip ist eine Darstellung in Abhängigkeit von den generellen charakterisierenden Eigenschaften der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$. Mit der klassischen Definition ist eine solche Einbeziehung aber nicht möglich, da in diesem Rahmen triviale Fälle von Konsistenz auftreten können, die der Charakteristik der Menge der Funktionen nicht gerecht werden. Dabei bezeichnet *triviale Konsistenz* die Situation, in der für eine Menge von

Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, das ERM-Prinzip nicht konsistent ist, aber eine zusätzliche Funktion $\Phi(z)$ existiert, die die Funktionen minorisiert:

$$\inf_{\alpha \in \Lambda} Q(z, \alpha) > \Phi(z), \quad \text{für alle } z \in I\!R^{d+1}.$$

Daraus folgt direkt, dass für die erweiterte Menge von Funktionen

$$\{Q(z, \alpha), \alpha \in \Lambda\} \cup \{\Phi(z)\}$$

die Konsistenz der empirischen Risiko-Minimierung erzwungen wird. Dies liegt daran, dass für jede Verteilungsannahme P_Z und jede Anzahl von Beobachtungen n sowohl das empirische Risiko $R_{emp}(\alpha)$ als auch das Risiko $R(\alpha)$ durch die Funktion $\Phi(z)$ minimiert wird. Damit wird nur die minorisierende Eigenschaft der einen Funktion in der Menge der Funktionen berücksichtigt. Das Ziel muss aber die Entwicklung eines Konsistenz-Konzepts sein, das sich auf die generellen Eigenschaften der Funktionenmenge stützt, und nicht von einzelnen Funktionen in der Menge abhängt, die dann auch einzeln überprüft werden müssten.

2.8 Definition *nicht-triviale Konsistenz*
Das Prinzip der empirischen Risiko-Minimierung ist *nicht-trivial (strikt) konsistent* bezüglich der Menge von Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, und der Verteilungsfunktion $F_Z(z)$, falls für jede nichtleere Teilmenge

$$\Lambda(c) = \left\{ \alpha \;\middle|\; \int Q(z, \alpha) dF(z) > c,\ \alpha \in \Lambda, \right\}, \quad c \in I\!R,$$

von Λ die Konvergenz

$$\inf_{\alpha \in \Lambda(c)} R_{emp}(\alpha) \xrightarrow[n \to \infty]{Pr} \inf_{\alpha \in \Lambda(c)} R(\alpha)$$

gilt.

Die Definition erlaubt eine geeignete Adaption des klassischen Konsistenzbegriffs an die Problematik der statistischen Lerntheorie in dem Sinne, dass nicht nur für die vollständige Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, sondern auch für jede Teilmenge $Q(z, \alpha)$, $\alpha \in \Lambda(c)$, Konsistenz gelten muss, wobei in diesen Teilmengen nur noch die

Funktionen berücksichtigt werden, deren Risikowert höher als ein Wert $c \in I\!R$ ist. Die Konsistenz wird somit auf jedem Risikoniveau überprüft, so dass die Konsistenz nicht von den Eigenschaften einzelner Elemente in der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, abhängen kann. Im Gegensatz zur Definition nach dem klassischem Konsistenzbegriff, muss in Definition 2.8 statt zwei Bedingungen nur eine erfüllt sein, dafür allerdings für eine nicht abzählbare Anzahl von Teilmengen der Funktionenmenge. Im nächsten Abschnitt wird gezeigt, dass die Bedingung in dieser Definition trotzdem nachprüfbar ist.
Weiterhin folgt die triviale Konsistenz direkt aus der nicht-trivialen Konsistenz. Die zweite Bedingung

$$R_{emp}(\widetilde{\alpha}_n) \xrightarrow[n\to\infty]{Pr} \inf_{\alpha\in\Lambda} R(\alpha)$$

für die klassische Konsistenz folgt wegen

$$R_{emp}(\widetilde{\alpha}_n) = \inf_{\alpha\in\Lambda} R_{emp}(\alpha)$$

direkt aus der Bedingung in Definition 2.8. Für die erste Bedingung folgt die Implikation mit folgendem Lemma.

2.9 Lemma *Vapnik (1998)*
Falls das ERM-Prinzip nicht-trivial konsistent ist, gilt folgende Konvergenz nach Wahrscheinlichkeit:

$$R(\widetilde{\alpha}_n) \xrightarrow[n\to\infty]{Pr} \inf_{\alpha\in\Lambda} R(\alpha).$$

Die Umkehrung des obigen Lemmas gilt im Allgemeinen nicht. Die Definition der klassischen Konsistenz ist mit dem Lemma in das weitergehende Konzept der nicht-trivialen Konsistenz eingebettet. Im Weiteren werden hinreichende und notwendige Bedingungen zum Nachweis der nicht-trivialen Konsistenz des ERM-Prinzips aufgezeigt. Die Analyse der Konsistenz der empirischen Risiko-Minimierung ist eng verbunden mit der Theorie zur Konvergenz empirischer Prozesse, so dass die Sätze dieser Theorie zum Nachweis der nicht-trivialen Konsistenz genutzt werden können.

2.10 Definition *Empirische Prozesse*
Sei $F_Z(z)$ eine auf $\mathbb{R}^{d+1}$ definierte Verteilungsfunktion und sei $Q(z,\alpha)$, $\alpha \in \Lambda$, eine Menge messbarer Funktionen bezüglich des Maßes $F_Z(z)$. Sei

$$z_1, \ldots, z_n, \ldots$$

eine Folge von Beobachtungsvektoren aus $\mathbb{R}^{d+1}$, identisch gezogen gemäß der Verteilungsfunktion $F_Z(z)$. Dann heißt

(i) die Folge von Zufallsvariablen

$$\zeta_n = \zeta(z_1, \ldots, z_n) = \sup_{\alpha\in\Lambda} \left| \int Q(z,\alpha) dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \right|, \; n = 1, 2, \ldots,$$

zweiseitiger empirischer Prozess und

(ii) die Folge von Zufallsvariablen

$$\zeta_n^+ = \zeta(z_1, \ldots, z_n) = \sup_{\alpha\in\Lambda} \left(\int Q(z,\alpha) dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \right)_+, \; n = 1, 2, \ldots,$$

einseitiger empirischer Prozess, wobei $(u)_+ = u$, falls $u > 0$, und $(u)_+ = 0$ sonst.

Beide Prozesse hängen dabei von dem Wahrscheinlichkeitsmaß $F_Z(z)$ und der Menge der Funktionen $Q(z,\alpha)$, $\alpha \in \Lambda$, ab und stellen den supremalen Abstand zwischen den Funktionen in Abhängigkeit von $\alpha \in \Lambda$ dar. Damit ermöglichen die Prozesse eine Worst-Case-Analyse des Abstands von Risiko und empirischem Risiko. Die Definition des einseitigen und zweiseitigen empirischen Prozesses ist sehr speziell und auf den hier notwendigen Kontext zur Suche nach Bedingungen für die nicht-triviale Konsistenz des ERM-Prinzips abgestimmt. Im Zusammenhang mit der Konsistenz des statistischen Lernprozesses sind für die zweiseitigen und einseitigen empirischen Prozesse vor allem Bedingungen für die Konvergenz von Interesse, also Bedingungen dafür, wann die Prozesse ζ_n und ζ_n^+ in Wahrscheinlichkeit gegen den Erwartungswert Null konvergieren:

$$P\left(\sup_{\alpha\in\Lambda} \left| \int Q(z,\alpha) dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \right| > \varepsilon \right) \underset{n\to\infty}{\longrightarrow} 0 \quad \forall \varepsilon > 0$$

$$\text{bzw.} \quad P\left(\sup_{\alpha\in\Lambda} \left(\int Q(z,\alpha) dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \right)_+ > \varepsilon \right) \underset{n\to\infty}{\longrightarrow} 0 \quad \forall \varepsilon > 0.$$

Bei finiter Funktionenmenge $Q(z, \alpha)$, $\alpha \in \Lambda$, sind die klassischen Konvergenzgesetze der mathematischen Statistik auf das Problem der Konvergenz der obigen Prozesse anwendbar, da das Supremum der jeweiligen Differenzen existiert (Vapnik, 1995). Im Gegensatz dazu muss bei infiniter Menge $Q(z, \alpha)$, $\alpha \in \Lambda$, mit $|\Lambda| = \infty$, sicher gestellt sein, dass die Eigenschaften der Menge zusammen mit dem Wahrscheinlichkeitsmaß $F_Z(z)$ die Konvergenz der beschriebenen empirischen Prozesse ζ_n bzw. ζ_n^+ erlauben. Das heißt, dass das Mittel $\frac{1}{n}\sum_{i=1}^{n} Q(z_i, \alpha)$ gleichmäßig zwei- bzw. einseitig gegen den Erwartungswert $\int Q(z, \alpha) dF_Z(z)$ im Funktionenraum $Q(z, \alpha)$, $\alpha \in \Lambda$, konvergiert. Dies ist eine Verallgemeinerung des schwachen Gesetzes der großen Zahlen auf Funktionenräume. Die allgemeingültige Definition eines stochastischen Prozesses, die die obige Definition der speziellen empirischen Prozesse mit einschließt, wird im dritten Kapitel mit den Definitionen 3.1 und 3.2 gegeben, die dort vorgestellten Eigenschaften, insbesondere die Gesetze der großen Zahlen, gelten dementsprechend auch für die hier definierten empirischen Prozesse, wenn die Menge der Funktionen finit ist, oder die Charakteristik einer infiniten Menge einfach genug ist, damit Bedingungen für die Konvergenz der Prozesse ζ_n oder ζ_n^+ gefunden werden können. Mit Hilfe dieser Prozesse und unter der Annahme geeigneter charakterisierender Eigenschaften, d. h. Entropie-Eigenschaften, der Menge $Q(z, \alpha)$, $\alpha \in \Lambda$, bei Berücksichtigung der Verteilungsannahme P_Z wird im nächsten Abschnitt die Konsistenz der statistischen Lerntheorie nachgewiesen und die Bedingungen für die Konsistenz diskutiert.

2.3 Konsistenz des statistischen Lernprozesses

Die nicht-triviale Konsistenz als geltendes Konsistenzkonzept zur Anwendung auf die empirische Risiko-Minimierung ist notwendig, um zu gewährleisten, dass die charakterisierenden Eigenschaften der gesamten Menge von Funktionen in die Betrachtung der Konsistenz mit einbezogen werden. Daraus ergibt sich allerdings die Notwendigkeit bezüglich einer nicht abzählbaren Anzahl von Teilmengen $\lambda(c)$, $c \in I\!R$, die Konsistenz des minimalen empirischen Risikos gegen das minimale Risiko überprüfen zu müssen. In diesem Abschnitt wird für unabhängige, identisch verteilte Beobachtungen $z_1, \ldots, z_n$ gezeigt, dass Konvergenz eines spezifischen empirischen Prozesses notwendig und hin-

reichend für die nicht-triviale Konsistenz des ERM-Prinzips ist, und dass es für diese Konvergenz ebenfalls notwendige und hinreichende Bedingungen gibt, die konstruktiv sind. Im vierten Kapitel wird die Konsistenz des statistischen Lernprozesses für spezielle Abhängigkeitsstrukturen in den Daten, also für identisch verteilte, aber abhängige Beobachtungen, unter Ausnutzung der Ergebnisse aus dem dritten Kapitel nachgewiesen.
Der Nachweis, dass es notwendige und hinreichende Bedingungen für die nicht-triviale Konsistenz des ERM-Prinzips gibt, erbrachte eine der wichtigsten Grundlagen für die Etablierung der statistischen Lerntheorie, da nur durch konstruktive Bedingungen die Theorie anwendbar wird. Vapnik und Chervonenkis (1989) haben das sogenannte Kerntheorem der statistischen Lerntheorie für die Situation mit unabhängigen, identisch verteilten Beobachtungen $z_1, \ldots, z_n$ nachgewiesen.

2.11 Satz *Kerntheorem der statistischen Lerntheorie*
Seien $a \in \mathbb{R}$ und $A \in \mathbb{R}$ Konstanten, so dass für alle Funktionen aus der Menge $Q(z, \alpha)$, $\alpha \in \Lambda$, und für eine gegebene Verteilungsfunktion $F_Z(z)$ die Ungleichungen

$$a \leq \int Q(z, \alpha) dF_Z(z) \leq A, \quad \alpha \in \Lambda,$$

gelten. Seien weiter $z_1, \ldots, z_n, \ldots$ unabhängige und gemäß $F_Z(z)$ verteilte Beobachtungen. Dann sind folgende Aussagen äquivalent:

(i) Für die gegebene Verteilungsfunktion $F_Z(z)$ ist die Methode der empirischen Risiko-Minimierung strikt konsistent auf der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$.

(ii) Für die gegebene Verteilungsfunktion gilt gleichmäßige einseitige Konvergenz des Mittels gegen den Erwartungswert auf der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$:

$$\lim_{n \to \infty} P\left(\sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \varepsilon \right) = 0, \quad \forall \varepsilon > 0.$$

Die Notwendigkeit der Bedingung (ii) für die Konsistenz des ERM-Prinzips in Satz 2.11 ist die wichtige Aussage des Kerntheorems. Denn dadurch ist bewiesen, dass jede Analyse der Konvergenzbedingungen des ERM-Prinzips eine Worst-Case-Analyse sein

muss. Bemerkenswert ist, dass nicht nur die hinreichende, sondern auch die notwendige Bedingung für die Konsistenz davon abhängt, ob noch für die schlechtest mögliche Wahl von $\alpha \in \Lambda$ der Abstand zwischen Risiko und empirischen Risiko für $n \to \infty$ in Wahrscheinlichkeit gegen Null konvergiert. Die Beschränkung des Risikos durch die Konstanten a bzw. A stellt dabei nur eine geringe Einschränkung dar. In der Regel können beide Werte ausreichend klein bzw. groß gewählt werden.

Basis für das Kerntheorem der statistischen Lerntheorie sind im wesentlichen wahrscheinlichkeitstheoretische Abschätzungen und kombinatorische Mengenbeziehungen, sowie im zentralen Maße Aussagen mittels des schwachen Gesetzes der großen Zahlen. Daraus ergibt sich als wesentliche Einschränkung aus den Voraussetzungen des Satzes die Unabhängigkeitsbedingung für die beobachteten Daten $z_1, \ldots, z_n$, denn dadurch wird die Problemstellung dahingehend vereinfacht, dass die Gesetze der großen Zahlen ohne weitere Bedingungen genutzt werden können. Eine Abschwächung der Voraussetzungen, so dass auch gewisse Abhängigkeitsstrukturen in den Daten zugelassen werden können, ist allerdings sinnvoll, da so das Prinzip der empirischen Risiko-Minimierung auch auf beispielsweise zeitabhängige Beobachtungen ausgeweitet werden kann. Die Erweiterung des Kerntheorems in Kapitel 4 führt zu Aussagen, die Abhängigkeitsstrukturen erlauben und Satz 2.11 als Spezialfall integrieren, so dass der dort geführte Nachweis des Kerntheorems auch für den obigen Satz gilt.

In der klassischen mathematischen Statistik wird das Problem der gleichmäßigen einseitigen Konvergenz nicht betrachtet, es wird speziell für die Konsistenz des statistischen Lernprozesses wegen der Aussagen des Kerntheorems 2.11 wichtig. In diesem Kontext liegt eine asymmetrische Situation vor, denn die Konsistenz des ERM-Prinzips bezieht sich auf die Minimierung des Risikos unter Ausnutzung des minimalen empirischen Risikos. Da bei der Anpassung einer Funktion $f(x, \alpha)$ an den Output y über das empirische Risiko nur eine Auswahl der Paare $z = (x, y) \in \mathbb{R}^{d+1}$ berücksichtigt werden muss, ist das minimale empirische Risiko bei fester Menge von Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, immer kleiner als das minimale Risiko, das über die gesamte Verteilung P_Z gebildet wird.

Satz 2.11 gilt für eine feste Verteilung P_Z mit zugehöriger Verteilungsfunktion $F_Z(z)$. Allerdings soll die statistische Lerntheorie für jede Verteilung aus einer Familie von Verteilungen $\mathcal{P}$ gelten. Damit die Bedingungen für die Konsistenz der Methode der empirischen Risiko-Minimierung für jedes Wahrscheinlichkeitsmaß aus $\mathcal{P}$ gilt, muss das Kerntheorem der statistischen Lerntheorie so verallgemeinert werden, dass die Bedingungen für die Konsistenz auf der gesamten zu $\mathcal{P}$ gehörigen Familie der Verteilungsfunktionen $\mathbb{F}$ gelten.

2.12 Korollar *Vapnik (1998)*
Seien $a \in I\!R$ und $A \in I\!R$ Konstanten, so dass für alle Funktionen aus der Menge $Q(z, \alpha)$, $\alpha \in \Lambda$, und für alle Verteilungsfunktionen $F_Z(z)$ in der Familie der Verteilungen $\mathbb{F}$ die Ungleichungen

$$a \leq \int Q(z, \alpha) dF_Z(z) \leq A, \quad \alpha \in \Lambda,\ F_Z(z) \in \mathbb{F},$$

gelten. Gelte weiter für alle Verteilungsfunktionen $F_Z(z) \in \mathbb{F}$, dass $z_1, \ldots, z_n, \ldots$ jeweils gemäß $F_Z(z)$ verteilte und unabhängige Beobachtungen sind. Dann sind folgende Aussagen äquivalent:

(i) Für jede Verteilungsfunktion in der Menge $\mathbb{F}$ ist die Methode der empirischen Risiko-Minimierung strikt konsistent auf der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$.

(ii) Für jede Verteilungsfunktion in der Menge $\mathbb{F}$ gilt gleichmäßige einseitige Konvergenz des Mittels gegen den Erwartungswert auf der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$:

$$\lim_{n \to \infty} P\left(\sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \varepsilon \right) = 0, \quad \forall \varepsilon > 0.$$

Mit dem Kerntheorem 2.11 und dem daraus folgenden Korollar 2.12 wird das Problem, strikte Konsistenz der Methode der empirischen Risiko-Minimierung nachzuweisen, verschoben auf den Existenznachweis der gleichmäßigen einseitigen Konvergenz des Mittels gegen den Erwartungswert, also der Konvergenz des einseitigen empirischen Prozesses.

Daraus ergibt sich die Suche nach notwendigen und hinreichenden Bedingungen für diese Konvergenz.

Bei der Lösung dieses Problems können die spezifischen Eigenschaften der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, unter Ausnutzung empirischer Daten genutzt werden, denn schon bei der Definition der nicht-trivialen Konsistenz wird der Einfluss der generellen Entropie-Eigenschaften der Funktionenmenge deutlich. Für den einfachsten Fall, die Annahme einer endlichen Menge von Funktionen, d. h. die Annahme von $|\Lambda| = K < \infty$, kann die Bedingung für die gleichmäßige einseitige Konvergenz

$$P\left\{\sup_{1\leq k\leq K}\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right\} \underset{n\to\infty}{\longrightarrow} 0, \quad \text{für alle } \varepsilon > 0,$$

durch eine obere Schranke für die obige Wahrscheinlichkeit ermittelt werden, die von der Größe der Funktionenmenge abhängt, so dass unter Nutzung dieser Schranke die notwendige und hinreichende Bedingung deutlich wird. Für eine endliche Menge reeller, beschränkter Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, mit $|Q(z, \alpha)| \leq B$ für alle $\alpha \in \Lambda$ und $B < \infty$ gelten die folgenden Ungleichungen:

$$\begin{aligned} P\left\{\sup_{1\leq k\leq K}\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right\} & \\ \leq \sum_{k=1}^{K} P\left\{\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right\} & \\ \leq K\, P\left\{\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right\}. & \end{aligned}$$

Unter Ausnutzung der Hoeffding-Ungleichung (Hoeffding, 1963) mit der Konstanten B für die Funktionswerte gilt die Ungleichung

$$P\left\{\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right\} < \exp\left\{-\frac{\varepsilon^2 n}{2B^2}\right\}.$$

Insgesamt folgt daraus die obere Schranke

$$\begin{aligned} P\left\{\sup_{1\leq k\leq K}\left(\int Q(z,\alpha_k)dF_z(z)-\frac{1}{n}\sum_{i=1}^{n}Q(z_i,\alpha_k)\right)>\varepsilon\right\} & \\ < \quad K\ \exp\left\{-\frac{\varepsilon^2 n}{2B^2}\right\} & \\ = \quad \exp\left\{\left(\frac{\ln K}{n}-\frac{\varepsilon^2}{2B^2}\right)n\right\}, & \end{aligned}$$

die die Wahrscheinlichkeit angibt, dass die Differenz zwischen empirischem Risiko und Risiko kleiner als ein ε ist. Ein ähnliches Ergebnis existiert auch für die einfachere endliche Menge von Indikatorfunktionen $Q(z,\alpha)$, $\alpha \in \Lambda$, mit $|\Lambda| < \infty$ und $Q(z,\alpha) \in \{0,1\}$. Dann gilt unter Verwendung der Chernoff-Ungleichung (Chernoff, 1952), die ein Spezialfall der Hoeffding-Ungleichung (Hoeffding, 1963) ist, ebenfalls eine Abschätzung für die gleichmäßige Konvergenz der Mittel gegen den Erwartungswert:

$$\begin{aligned} P\left\{\sup_{1\leq k\leq K}\left(\int Q(z,\alpha_k)dF_z(z)-\frac{1}{n}\sum_{i=1}^{n}Q(z_i,\alpha_k)\right)>\varepsilon\right\} & \\ < \quad K\ \exp\left\{-2\varepsilon^2 n\right\} & \\ = \quad \exp\left\{\left(\frac{\ln K}{n}-2\varepsilon^2\right)n\right\}. & \end{aligned}$$

Aus den jeweiligen oberen Schranken lässt sich direkt ablesen, dass für alle $\varepsilon > 0$ die gleichmäßige einseitige Konvergenz erreicht werden kann, wenn

$$\frac{\ln K}{n} \underset{n\to\infty}{\longrightarrow} 0$$

gilt. Diese hinreichende Bedingung ist in diesem speziellen Fall einer finiten Anzahl von Funktionen immer erfüllt, so dass automatisch auch die einseitige gleichmäßige Konvergenz gelten muß. Diese Bedingung ist eine Aussage über die Größe oder Mächtigkeit der Menge der Funktionen im Verhältnis zur Anzahl der Beobachtungen.

Bedingungen dieser Art werden auch im Weiteren Indikatoren für die gleichmäßige Konvergenz sein, auch wenn die Anzahl der Funktionen in der Menge $Q(z,\alpha)$, $\alpha \in \Lambda$, infinit ist. Um geeignete Bedingungen entwickeln zu können, müssen dazu Maße für

die Mächtigkeit oder die Mannigfaltigkeit der Funktionenmenge definiert werden, die in Abhängigkeit von der Anzahl der Beobachtungen die Eigenschaften der Menge widerspiegelt. Dabei ist die wichtigste Idee, dass trotz der unendlichen Anzahl der Elemente nur eine endliche Anzahl von Klassen in dieser Funktionenmenge bezüglich der gegebenen Beobachtungen $z_1, \ldots, z_n$ *unterscheidbar* sind. Um die Unterscheidbarkeit angemessen definieren zu können, muss ein Entropie-Konzept gelten, dass die Ähnlichkeiten in der Menge der Funktionen unter Berücksichtigung der gegebenen Beobachtungen quantifiziert. Dazu wird die VC-Entropie von Vapnik und Chervonenkis für Indikatorfunktionen (Vapnik und Chervonenkis, 1979) und für beschränkte reelle Funktionen (Vapnik und Chervonenkis, 1995) eingeführt.

Sei $Q(z, \alpha)$, $\alpha \in \Lambda$, eine Menge von Funktionen mit $|Q(z, \alpha)| \leq B$ und seien $z_1, \ldots, z_n$ identisch verteilte Beobachtungen, gezogen gemäß der Verteilungsfunktion $F_Z(z)$. Sei weiter

$$q(\alpha) = (Q(z_1, \alpha), \ldots, Q(z_n, \alpha)), \quad \alpha \in \Lambda,$$

die Menge der Vektoren, die sich aus den Verlusten für die verschiedenen α und den Beobachtungen ergeben. Ein Konzept für die Ähnlichkeit in Funktionenmengen in Abhängigkeit von den Beobachtungen kann aus der Ähnlichkeit der Vektoren $q(\alpha)$, $\alpha \in \Lambda$, abgeleitet werden.

2.13 Definition *minimales ε-Netz*
Sei $Q(z, \alpha)$, $\alpha \in \Lambda$, eine Menge beschränkter, reeller Funktionen mit $|Q(z, \alpha)| \leq B$ und seien $z_1, \ldots, z_n$ Beobachtungen. Dann hat die Menge von Vektoren $q(\alpha) \in I\!R^n$, $\alpha \in \Lambda$, unter Supremumsnorm ein *minimales ε-Netz*

$$q(\alpha_1), \ldots, q(\alpha_K),$$

falls eine minimale Anzahl von Vektoren $q(\alpha_1), \ldots, q(\alpha_K)$ existiert mit einem

$$K = K^{\Lambda}(\varepsilon; z_1, \ldots, z_n),$$

so dass es für jeden Vektor $q(\alpha)$, $\alpha \in \Lambda$, ein $q(\alpha_r)$ aus der Menge der K Vektoren $q(\alpha_1), \ldots, q(\alpha_K)$ gibt, das ε-dicht zu $q(\alpha)$ ist, d. h.

$$\|q(\alpha), q(\alpha_r)\|_\infty = \sup_{1 \leq i \leq n} |Q(z_i, \alpha) - Q(z_i, \alpha_r)| \leq \varepsilon.$$

In der Definition 2.13 ist $K^{\Lambda}(\varepsilon; z_1, \ldots, z_n)$ eine Zufallsvariable, die durch die Zufallsvektoren $z_1, \ldots, z_n$ mit gemeinsamer unbekannter Verteilungsfunktion $F_{Z_1,\ldots,Z_n}(z_1, \ldots, z_n)$ konstruiert wird. Die Bildung des Erwartungswerts über die logarithmierte Zufallsvariable ergibt die VC-Entropie für beschränkte, reelle Funktionen (Vapnik, 1995).

2.14 Definition *VC-Entropie für reelle Funktionen*
Sei $Q(z, \alpha)$, $\alpha \in \Lambda$, eine Menge beschränkter, reeller Funktionen mit $|Q(z, \alpha)| \leq B$ und seien $z_1, \ldots, z_n$ Beobachtungen mit gemeinsamer Verteilungsfunktion

$$F_{Z_1,\ldots,Z_n}(z_1, \ldots, z_n).$$

Haben die n-dimensionalen Vektoren

$$q(\alpha) = \left(Q(z_1, \alpha), \ldots, Q(z_n, \alpha)\right), \quad \alpha \in \Lambda,$$

ein minimales ε-Netz bezüglich der Supremumsnorm $\|\cdot\|_{\infty}$ mit Anzahl $K^{\Lambda}(\varepsilon; z_1, \ldots, z_n)$ von Vektoren, dann heißt

$$H^{\Lambda}(\varepsilon, n) = \mathrm{E}\left(\ln K^{\Lambda}(\varepsilon; z_1, \ldots, z_n)\right)$$

VC-Entropie der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, wobei der Erwartungswert bezüglich $F_{Z_1,\ldots,Z_n}(z_1, \ldots, z_n)$ gebildet wird.

Die VC-Entropie für reelle Funktionen ist die Verallgemeinerung des Entropiebegriffs für Indikatorfunktionen durch die Nutzung von ε-Netzen. Für die Menge von Indikatorfunktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, ist

$$q(\alpha) = \left(Q(z_1, \alpha), \ldots, Q(z_n, \alpha)\right), \quad \alpha \in \Lambda,$$

eine Menge binärer Vektoren, die im n-dimensionalen Raum einen Einheitswürfel mit $K < n$ Ecken bilden. Die Anzahl unterschiedlicher Ecken bezüglich der Beobachtungen $z_1, \ldots, z_n$ gibt somit einen Wert $K = K^{\Lambda}(z_1, \ldots, z_n)$ für die Entropie der Menge der Indikatorfunktionen an, der nicht von einem ε abhängt. Das minimale ε-Netz für Indikatorfunktionen ist für alle $\varepsilon < 1$ eine Teilmenge der Ecken des Einheitswürfels. Die *VC-Entropie für eine Menge von Indikatorfunktionen* $Q(z, \alpha)$, $\alpha \in \Lambda$, vereinfacht sich damit

zu

$$H^{\Lambda}(n) = \mathrm{E}\left(\ln K^{\Lambda}(z_1, \ldots, z_n)\right)$$

(vgl. Vapnik und Chervonenkis, 1979, und Vapnik, 1982). Die VC-Entropie hängt durch die Bildung des Erwartungswerts von der unbekannten Verteilungsfunktion $F_{Z_1,\ldots,Z_n}(z_1, \ldots, z_n)$ ab und ist dadurch nicht direkt berechenbar. Trotzdem gibt es unter Ausnutzung der VC-Entropie notwendige und hinreichende Bedingungen für die zweiseitige bzw. einseitige gleichmäßige Konvergenz bei unabhängig erhobenen Beobachtungen und damit auch für das Prinzip der empirischen Risiko-Minimierung.

2.15 Satz *Vapnik und Chervonenkis (1981)*
Sei $Q(z, \alpha)$, $\alpha \in \Lambda$, eine Menge beschränkter, reeller Funktionen mit $|Q(z, \alpha)| \leq B$ und seien $z_1, \ldots, z_n$ unabhängig und identisch verteilte Beobachtungen gemäß der Verteilung $F_Z(z)$. Damit zweiseitige gleichmäßige Konvergenz

$$P\left(\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0 \quad \forall\varepsilon > 0$$

über der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, gilt, ist es notwendig und hinreichend, dass für alle $\varepsilon > 0$ die folgenden Bedingungen erfüllt sind:

$$\frac{H^{\Lambda}(\varepsilon, n)}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

Die Bedingungen für die zweiseitige gleichmäßige Konvergenz für eine infinite Anzahl von Funktionen hat also dieselbe Form wie für eine endliche Anzahl, wobei hier mit der VC-Entropie eine Angabe über die Komplexität der Funktionenmenge gemacht wird und nicht die Anzahl der Funktionen K genutzt wird.
Durch die Notwendigkeit der Einführung des ε-Netzes für reelle Funktionen muss für jedes $\varepsilon > 0$ die Bedingung gelten. Für den Spezialfall der Indikatorfunktionen vereinfacht sich die Aussage des Satzes zu folgendem Korollar.

2.16 Korollar *Vapnik und Chervonenkis (1968)*
Sei $Q(z,\alpha)$, $\alpha \in \Lambda$, eine Menge von Indikatorfunktionen mit $Q(z,\alpha) \in \{0,1\}$ und seien $z_1,\ldots,z_n$ unabhängig und identisch verteilte Beobachtungen gemäß der Verteilung $F_Z(z)$. Damit zweiseitige gleichmäßige Konvergenz

$$P\left(\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0, \quad \forall\varepsilon > 0,$$

über der Menge der Funktionen $Q(z,\alpha)$, $\alpha \in \Lambda$, gilt, ist es notwendig und hinreichend, dass die Bedingung

$$\frac{H^{\Lambda}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0$$

erfüllt ist.

Der Nachweis für Satz 2.15 kann ebenso wie für Korollar 2.16 aus den zitierten Artikeln oder aus dem Buch zur statistischen Lerntheorie von Vapnik (1998) entnommen werden. Dort gibt es eine ähnliche Aussage auch für Mengen unbeschränkter reeller Funktionen, die hier aber nicht weiter betrachtet werden sollen. Damit die Ergebnisse aus Satz 2.15 und dem Korollar 2.16 für Aussagen über notwendige und hinreichende Bedingungen zur nicht-trivialen Konsistenz des Prinzips der empirischen Risiko-Minimierung genutzt werden können, muss die Anwendung der obigen Bedingungen auch auf einseitige gleichmäßige Konvergenz ausgedehnt werden.
Gleichmäßige zweiseitige Konvergenz kann dargestellt werden durch

$$P\left(\left\{\sup_{\alpha\in\Lambda}\left(R(\alpha) - R_{emp}(\alpha)\right) > \varepsilon\right\} \vee \left\{\sup_{\alpha\in\Lambda}\left(R_{emp}(\alpha) - R(\alpha)\right) > \varepsilon\right\}\right) \underset{n\to\infty}{\longrightarrow} 0 \quad \forall\varepsilon > 0,$$

so dass sofort deutlich wird, dass diese Konvergenz eine hinreichende Bedingung für die Konvergenz des Prinzips der empirischen Risiko-Minimierung ist. Allerdings wird in der statistischen Lerntheorie die unsymmetrische Situation betrachtet, dass die Generalisierungsfähigkeit im Lernprozess mit dem ERM-Prinzips nur durch die Minimierung des Risikos und nicht durch die Maximierung erreicht wird. Das bedeutet für eine Funktion $Q(z,\alpha)$, die die zweiseitige gleichmäßige Konvergenz nicht erfüllt, dass trotzdem eine Funktion $Q^*(z,\alpha')$ existieren kann, die entsprechend konvergiert und geeignet nah zur ursprünglichen Funktion ist. Im folgenden Satz von Vapnik und Chervonenkis (1989)

wird gezeigt, dass unter Ausnutzung dieser Idee notwendige und hinreichende Bedingungen für einseitige gleichmäßige Konvergenz gefunden werden können.

2.17 Satz *Vapnik und Chervonenkis (1989)*
Sei $Q(z,\alpha)$, $\alpha \in \Lambda$, ein Menge beschränkter, reeller Funktionen mit $|Q(z,\alpha)| \leq B$ und seien $z_1, \ldots, z_n$ unabhängig und identisch verteilte Beobachtungen gemäß der Verteilung $F_Z(z)$. Dann ist für die einseitige gleichmäßige Konvergenz

$$P\left(\sup_{\alpha\in\Lambda}\left(\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right) > \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0 \quad \forall\varepsilon > 0$$

notwendig und hinreichend, dass für jedes positive ε, ϵ und δ eine Menge von Funktionen existiert, so dass gilt:

(i) Für jede Funktion $Q(z,\alpha)$ existiert eine Funktion $Q^*(z,\alpha')$, die die Bedingungen

$$Q(z,\alpha) \geq Q^*(z,\alpha'),$$

$$\int \big(Q(z,\alpha) - Q^*(z,\alpha')\big)dF_Z(z) < \epsilon$$

erfüllt.

(ii) Die VC-Entropie der Menge von Funktionen $Q^*(z,\alpha')$, $\alpha' \in \Lambda'$, erfüllt die Ungleichung

$$\lim_{n\to\infty} \frac{H^{\Lambda'}(\varepsilon,n)}{n} < \delta.$$

Zum Beweis, dass die Bedingung in diesem Satz von Vapnik und Chervonenkis (1989) hinreichend ist, kann dieselbe Technik wie für Satz 2.15 genutzt werden, d. h. es wird dieselbe Idee verfolgt, wie sie auch für eine finite Anzahl von Funktionen aufgezeigt wurde. Es wird allerdings die VC-Entropie der Menge $Q^*(z,\alpha')$, $\alpha' \in \Lambda'$, als Komplexitätsmaß für die ursprüngliche Menge $Q^*(z,\alpha)$, $\alpha \in \Lambda$, genutzt. Bemerkenswert ist dabei, dass die Bedingung in Satz 2.17 schwächer ist als die Bedingung für die gleichmäßige zweiseitige Konvergenz.
Der Satz 2.17 gilt ebenfalls für Indikatorfunktionen, wobei dann die notwendige und hinreichende Bedingung nicht mehr von ε abhängt. Die wichtigste Bemerkung ist aber,

dass der Satz nicht nur für eine feste Verteilung mit Verteilungsfunktion $F_Z(z)$ gilt, sondern für jede beliebige Verteilung P_Z aus der Klasse von Verteilungen $\mathcal{P}$. Dazu muss analog zu Satz 2.17 für jedes positive ε, ϵ und δ die Bedingung

$$\lim_{n\to\infty} \frac{H^{\Lambda'}(\varepsilon, n)}{n} < \delta$$

für jede Verteilungsfunktion $F_Z(z)$ aus der Menge der Verteilungsfunktionen $\mathbb{F}$ gültig sein. Mit dieser zentralen Bemerkung wird zusammen mit Korollar 2.12 deutlich, dass die Bedingungen in Satz 2.17 notwendig und hinreichend für die nicht-triviale Konsistenz des Prinzips der empirischen Risiko-Minimierung sind.

2.4 Konvergenzrate des statistischen Lernprozesses

Nach der Bestimmung der Bedingungen für die nicht-triviale Konsistenz des Prinzips der empirischen Risiko-Minimierung ist es zur Vervollständigung des konzeptionellen Teils der statistischen Lerntheorie von Bedeutung, Aussagen darüber zu treffen, unter welchen Bedingungen die Konvergenzrate der geschätzten Risikowerte $R(\widetilde{\alpha}_n)$, $n = 1, 2, \ldots$, gegen das minimale Risiko $\inf_{\alpha\in\Lambda} R(\alpha)$ exponentiell ist. Dazu müssen für ein gegebenes Wahrscheinlichkeitsmaß obere Schranken gefunden werden, so dass diese für ausreichend große n eine Abschätzung der Konvergenzrate gewährleisten. Gleichzeitig wird unter Ausnutzung einer solchen oberen Schranke der Wert des Risikos $R(\widetilde{\alpha}_n)$ ebenfalls abschätzbar.

Für eine Menge von Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, werden in diesem Abschnitt zuerst verteilungsabhängige Schranken entwickelt, aus denen sich aber direkt verteilungsunabhängige Schranken entwickeln lassen, wobei dafür generell angenommen wird, dass die Beobachtungen $z_1, \ldots, z_n$ identisch gemäß der Verteilungsfunktion $F_Z(z)$ erhoben werden. Die Grundlage für die Bestimmung und die Analyse der beiden Typen von Schranken sind jeweils unterschiedliche Entropie-Konzepte.

2.18 Definition *Entropie-Konzepte*
Sei $Q(z, \alpha)$, $\alpha \in \Lambda$, eine Menge beschränkter, reeller Funktionen mit $|Q(z, \alpha)| \leq B$ und seien $z_1, \ldots, z_n$ Beobachtungen mit gemeinsamer Verteilungsfunktion

$$F_{Z_1,\ldots,Z_n}(z_1, \ldots, z_n).$$

Sei $K^\Lambda(\varepsilon; z_1, \ldots, z_n)$ die Anzahl der Vektoren im minimalen ε-Netz der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$. Dann heißt

$$H^\Lambda_{ann}(\varepsilon, n) = \ln \mathrm{E}(K^\Lambda(\varepsilon; z_1, \ldots, z_n))$$

die *annealed VC-Entropie* und

$$G^\Lambda(\varepsilon, n) = \ln \sup_{z_1,\ldots,z_n} K^\Lambda(\varepsilon; z_1, \ldots, z_n)$$

die *Growth-Funktion* der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, wobei der Erwartungswert bei der annealed VC-Entropie bezüglich $F_{Z_1,\ldots,Z_n}(z_1, \ldots, z_n)$ gebildet wird.

Für Indikatorfunktionen kann wie bei der VC-Entropie auf das Konzept des minimalen ε-Netzes verzichtet werden, so dass die *annealed VC-Entropie für Indikatorfunktionen* durch

$$H^\Lambda_{ann}(n) = \ln \mathrm{E}(K^\Lambda(z_1, \ldots, z_n))$$

definiert ist und die *Growth-Funktion für Indikatorfunktionen* durch

$$G^\Lambda(n) = \ln \sup_{z_1,\ldots,z_n} K^\Lambda(z_1, \ldots, z_n).$$

Die beiden Konzepte sind zusammen mit der VC-Entropie für beschränkte, reelle Funktionen so definiert, dass für alle n und beliebiges, festes $\varepsilon > 0$ die Ungleichungen

$$H^\Lambda(\varepsilon, n) \leq H^\Lambda_{ann}(\varepsilon, n) \leq G^\Lambda(\varepsilon, n)$$

gelten. Für Indikatorfunktionen gelten entsprechende Ungleichungen:

$$H^\Lambda(n) \leq H^\Lambda_{ann}(n) \leq G^\Lambda(n).$$

Die jeweils erste Ungleichung folgt aus der Jensen-Ungleichung, die zweite Ungleichung folgt direkt aus den Definitionen der Entropien. Die annealed VC-Entropie ist ebenso wie die VC-Entropie nicht verteilungsfrei, beide hängen von der Verteilung P_Z ab,

weshalb diese beiden Entropie-Konzepte nicht unabhängig von der zu lösenden Aufgabe sind. Das Entropie-Konzept mit der Growth-Funktion ist dagegen verteilungsfrei, aber wie die anderen Entropien nicht konstruktiv, da es keine allgemeingültige Möglichkeit gibt, die Anzahl der Vektoren des minimalen ε-Netzes für eine beliebige Menge von Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, zu berechnen (vgl. Vapnik, 1993). Da aber die Growth-Funktion in direktem Zusammenhang mit der von Vapnik und Chervonenkis entwickelten VC-Dimension (Vapnik und Chervonenkis, 1968, 1971) steht, können die drei Entropien durch die konstruktive VC-Dimension nach oben abgeschätzt werden. Zusätzlich existieren empirische Verfahren zur Schätzung der VC-Dimension für eine gegebene Menge von Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, so dass die oben beschriebenen Schranken vollständig berechenbar werden (vgl. Vapnik, Levin und Le Cun, 1994).

Auf die konkrete Einführung der VC-Dimension kann hier allerdings verzichtet werden, da die Bedingungen zur Abschätzung der Konvergenzrate durch die nichtkonstruktiven, oberen Schranken direkt auf Schranken, die die VC-Dimension nutzen, übertragen werden können (vgl. Kapitel 4 und 5 in Vapnik, 1998). Allgemein ist die VC-Dimension in der gesamten Lerntheorie (Valiant, 1984) von besonderer Bedeutung, um verteilungsfreie Konvergenz nachweisen zu können. Dazu ist die wichtigste Voraussetzung immer die Finitheit der VC-Dimension (vgl. Blumer et al., 1989). In diesem Zusammenhang werden auch in Alon et al. (1997) geeignete Schranken für den Fall von reellwertigen Verlustfunktionen in der statistischen Lerntheorie angegeben.
Alle diese Ansätze zur Konstruktion von Schranken zur Kontrolle der Konvergenzrate haben dabei gemein, dass die Unabhängigkeit der Beobachtungen vorausgesetzt wird. Für das restliche Kapitel wird angenommen, dass unabhängig erhobene Beobachtungen betrachtet werden, bevor im dritten und vierten Kapitel für die statistische Lerntheorie Abhängigkeitsstrukturen in den Daten eingeführt werden.

Am einfachsten Fall, also für Mengen $Q(z, \alpha)$, $\alpha \in \Lambda$, mit endlicher Anzahl von Funktionen, soll die Idee für die Entwicklung der Schranken aufgezeigt werden, die eine schnelle Konvergenz der statistischen Lerntheorie garantieren. Im Abschnitt 2.3 wurden bereits obere Schranken für finite Funktionenmengen angegeben, die auch geeignet sind, die Schnelligkeit der Konvergenzrate abzuschätzen. Für eine Menge mit einer fi-

niten Anzahl K beschränkter, reeller Funktionen $Q(z,\alpha)$, $\alpha \in \Lambda$, mit $|Q(z,\alpha)| \leq B$ und für unabhängig und identisch gemäß $F_Z(z)$ verteilte Beobachtungen $z_1, \ldots, z_n$ gilt die Ungleichung

$$\begin{aligned} P\left\{ \sup_{1\leq k\leq K} \left(\int Q(z,\alpha_k) dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k) \right) > \varepsilon \right\} & \\ < \; K \; \exp\left\{ -\frac{\varepsilon^2 n}{2B^2} \right\} & \\ = \; \exp\left\{ \left(\frac{\ln K}{n} - \frac{\varepsilon^2}{2B^2} \right) n \right\}. & \end{aligned}$$

Aus dieser Ungleichung ergibt sich nicht nur die notwendige und hinreichende Bedingung

$$\frac{\ln K}{n} \underset{n\to\infty}{\longrightarrow} 0$$

für gleichmäßige einseitige Konvergenz. Die gleiche Bedingung ist auch hinreichend dafür, dass die exponentielle Schranke nicht trivial ist, das heißt, für jedes $\varepsilon > 0$ konvergiert die Schranke

$$\exp\left\{ \left(\frac{\ln K}{n} - \frac{\varepsilon^2}{2B^2} \right) n \right\}$$

gegen Null, falls

$$\frac{\ln K}{n} \underset{n\to\infty}{\longrightarrow} 0$$

gilt. Mit der Einführung von η mit $0 < \eta \leq 1$ als Wert für die obere Schranke kann

$$K \; \exp\left\{ -\frac{\varepsilon^2 n}{2B^2} \right\} = \eta$$

gesetzt werden, so dass sich durch Auflösung nach ε

$$\varepsilon = 2\,B\,\sqrt{\frac{\ln K - \ln \eta}{2n}}$$

ergibt. Damit kann die obige Ungleichung in die folgende äquivalente Form gebracht werden:
Mit Wahrscheinlichkeit $(1-\eta)$ ist die Ungleichung

$$R(\alpha_k) - R_{emp}(\alpha_k) \;\; \leq \;\; 2\,B\,\sqrt{\frac{\ln K - \ln \eta}{2n}}$$

gleichzeitig für alle K beschränkten, reellen Funktionen in der Menge $Q(z, \alpha_k)$, $k = 1, \ldots, K$, gültig.
Da die obige Aussage für alle K Funktionen gilt, kann dadurch der Wert des Risikos $R(\widetilde{\alpha}_n)$ für die durch Minimierung des empirischen Risikos gewählte Funktion $Q(z, \widetilde{\alpha}_n)$ abgeschätzt werden:

$$R(\widetilde{\alpha}_n) \;\leq\; R_{emp}(\widetilde{\alpha}_n) + 2\,B\,\sqrt{\frac{\ln K - \ln \eta}{2n}}.$$

Durch diese Abschätzung wird deutlich, dass sich der Risikowert $R(\widetilde{\alpha}_n)$ für große n in der Nähe des minimalen empirischen Risikos bewegt, wenn der zweite Summand klein genug ist, wofür

$$\frac{\ln K}{n} \underset{n\to\infty}{\longrightarrow} 0$$

die hinreichende Bedingung ist. Weiterhin gilt für beliebiges $\alpha_k \in \{\alpha_1, \ldots, \alpha_K\}$ die Hoeffding-Ungleichung (Hoeffding, 1963)

$$P\left\{\left(\frac{1}{n}\sum_{i=1}^{n} Q(z_i, \alpha_k) - \int Q(z, \alpha_k) dF_z(z)\right) > \varepsilon\right\} < \exp\left\{-\frac{\varepsilon^2 n}{2B^2}\right\},$$

woraus folgt, dass mit Wahrscheinlichkeit $(1 - \eta)$ die Abschätzung

$$R(\widetilde{\alpha}) = \inf_{\alpha\in\Lambda} R(\alpha) \;\geq\; R_{emp}(\widetilde{\alpha}) - 2\,B\,\sqrt{\frac{-\ln \eta}{2n}} \;\geq\; R_{emp}(\widetilde{\alpha}_n) - 2\,B\,\sqrt{\frac{-\ln \eta}{2n}}.$$

für das minimale Risiko existiert. Zusammen mit der Abschätzung für den Risikowert gilt deshalb für die Differenz des geschätzten Risikos zum minimalen Risiko mit Wahrscheinlichkeit $(1 - 2\eta)$ die Abschätzung

$$R(\widetilde{\alpha}_n) - \inf_{\alpha\in\Lambda} R(\alpha) \;\leq\; 2B\,\left(\sqrt{\frac{\ln K - \ln \eta}{2n}} + \sqrt{\frac{-\ln \eta}{2n}}\right).$$

Dabei ist

$$\frac{\ln K}{n} \underset{n\to\infty}{\longrightarrow} 0$$

wiederum eine hinreichende Bedingung dafür, dass diese Differenz der Risikos klein wird für großes n. Die beiden Abschätzungen für den Risikowert $R(\widetilde{\alpha}_n)$ und die Differenz $R(\widetilde{\alpha}_n) - \inf_{\alpha\in\Lambda} R(\alpha)$ erlauben zusammen mit den zugehörigen Voraussetzungen

eine vollständige Analyse der Generalisierungsfähigkeit der Methode der empirischen Risiko-Minimierung für eine finite Anzahl beschränkter, reeller Funktionen und unabhängig gezogener Beobachtungen mit identischer Verteilung. Für den Spezialfall einer endlichen Menge von Indikatorfunktionen gelten alle oben gemachten Aussagen gleichermaßen, dann mit der Konstanten $B = \sqrt{1/2}$.

Mit der gleichen Vorgehensweise können ähnliche Ergebnisse auch für Mengen mit infiniter Anzahl von Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, unter Verwendung der annealed VC-Entropie oder verteilungsfrei mit der Growth-Funktion hergeleitet werden. Im Weiteren werden Schranken für die Menge von Indikatorfunktionen angegeben. Auf die Ausweitung auf beschränkte, reelle Funktionen kann hier verzichtet werden, denn für solche Mengen ergibt sich die Herleitung der benötigten Schranken daraus, dass für die reellen Funktionen $Q(z, \alpha)$ ein weiteres Kapazitätskonzept entwickelt wird, die Schranken ansonsten aber genauso gebildet werden. Die Idee basiert darauf, dass die Charakteristik der Menge der reellen Funktionen durch die Eigenschaften und damit durch die annealed VC-Entropie oder die Growth-Funktion einer zugehörigen, geeigneten Menge von Indikatorfunktionen dargestellt werden kann. Aus diesem Grund gibt es bis auf ein differierendes Entropie-Konzept keinen Unterschied zwischen Indikatorfunktionen und beschränkten, reellen Funktionen bezüglich der Entwicklung exponentieller Schranken zur Analyse der Konvergenzrate. Für eine ausführliche Beschreibung des Entropiebegriffs für beschränkte, reelle Funktionen sei hier auf das Kapitel 5.2 in Vapnik (1998) verwiesen.

Zur Herleitung der Schranken für eine infinite Menge von Indikatorfunktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, mit $Q(z, \alpha) \in \{0, 1\}$ und $|\Lambda| = \infty$ bei unabhängig erhobenen Beobachtungen $z_1, \ldots, z_n$ ist genauso die Rate der Konvergenz der empirischen Risikowerte $R_{emp}(\alpha)$ gegen das Risiko $R(\alpha)$ zu untersuchen (vgl. Vapnik und Chervonenkis, 1991). Die exponentielle Abschätzung der gleichmäßigen zweiseitigen Konvergenz ist grundlegend für die Bestimmung der Konvergenzrate des ERM-Prinzips, da darauf alle weiteren Ergebnisse aufbauen. So lassen sich diese Ergebnisse beispielsweise auch auf infinite Mengen aus beschränkten, reellen Funktionen übertragen (vgl. Vapnik, 1995, 1998). Unter Ausnutzung des Konzeptes der annealed VC-Entropie $H^{\Lambda}_{ann}(n)$ für Indikatorfunktionen bei

einer Anzahl von Beobachtungen n gilt der folgende Satz.

2.19 Satz *Vapnik (1998)*
Sei $Q(z,\alpha)$, $\alpha \in \Lambda$, eine Menge von Indikatorfunktionen mit $Q(z,\alpha) \in \{0,1\}$ und seien $z_1,\ldots,z_n$ unabhängig und identisch verteilte Beobachtungen gemäß der Verteilung $F_Z(z)$. Dann gilt folgende Ungleichung:

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right\}$$
$$< \; 4\,\exp\left\{\left(\frac{H^{\Lambda}_{ann}(2n)}{n} - \left(\varepsilon - \frac{1}{n}\right)^2\right)n\right\}.$$

Hinreichend für die Nicht-Trivialität dieser exponentiellen Schranke ist die Bedingung

$$\frac{H^{\Lambda}_{ann}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

Der Beweis des Satzes wird in Vapnik (1998) ausführlich geführt. Eine ähnliche Aussage wurde bereits 1968 von Vapnik und Chervonenkis (1968, 1971) nachgewiesen, allerdings ist die Schranke nicht so eng gefasst wie in Satz 2.19. Mit einer Reihe technischer Modifikationen können für die Abschätzung in Satz 2.19 noch engere Schranken gefunden werden (vgl. beispielsweise Parrondo und Van den Broeck, 1993). Diese Schranken sind allerdings konzeptionell nicht grundlegend anders, so dass hier auf eine nähere Beschreibung verzichtet wird. Eine Übertragung der Aussagen auf den Fall, dass Abhängigkeitsstrukturen in den Beobachtungen vorliegen, wird bei Nutzung des gleichen Entropie-Konzeptes in Kapitel 4 nachgewiesen. Das dort erzielte Ergebnis ist allerdings eher mit der älteren, etwas weiteren Schranke von Vapnik und Chervonenkis (1968, 1971) vergleichbar. Die Idee zum Nachweis von Satz 2.19 geht zwingend von der Unabhängigkeit der Beobachtungen aus.

Genau wie bei finiten Funktionenmengen können mit Hilfe des Satzes 2.19 Aussagen über die Generalisierungsfähigkeit des Prinzips der empirischen Risiko-Minimierung bei einer infiniten Anzahl von Indikatorfunktionen gemacht werden. Diese Aussagen sind

aber nicht verteilungsfrei, da der Erwartungswert, der für die annealed VC-Entropie verwendet wird, von der Verteilung P_Z abhängt.

2.20 Satz *Vapnik (1998)*
Sei $Q(z,\alpha)$, $\alpha \in \Lambda$, eine Menge von Indikatorfunktionen mit $Q(z,\alpha) \in \{0,1\}$ und seien $z_1,\ldots,z_n$ unabhängig und identisch verteilte Beobachtungen gemäß der Verteilung $F_Z(z)$. Dann gilt für die Menge $Q(z,\alpha)$, $\alpha \in \Lambda$, mit Wahrscheinlichkeit $(1-\eta)$, $0 < \eta \leq 1$, die Ungleichung:

$$R(\widetilde{\alpha}_n) \quad \leq \quad R_{emp}(\widetilde{\alpha}_n) + \sqrt{\frac{H^{\Lambda}_{ann}(2n) - \ln(\eta/4)}{n}} + \frac{1}{n}.$$

Der Wert des Risikos $R(\widetilde{\alpha}_n)$ für die durch Minimierung des empirischen Risikos gewählte Funktion $Q(z,\widetilde{\alpha}_n)$ kann durch das minimale empirische Risiko und im Wesentlichen durch einen Ausdruck

$$\mathcal{E} = \mathcal{E}(n,\eta,H^{\Lambda}_{ann}) = \frac{H^{\Lambda}_{ann}(2n) - \ln(\eta/4)}{n}$$

abgeschätzt werden, der von der Entropie H^{Λ}_{ann}, der Wahrscheinlichkeit η und der Anzahl der Beobachtungen n abhängt. Genauso gibt es eine Abschätzung für die Differenz des geschätzten Risikos $R(\widetilde{\alpha}_n)$ zum minimalen Risiko $\inf_{\alpha\in\Lambda} R(\alpha)$ mit Wahrscheinlichkeit $(1-2\eta)$.

2.21 Satz *Vapnik (1998)*
Sei $Q(z,\alpha)$, $\alpha \in \Lambda$, eine Menge von Indikatorfunktionen mit $Q(z,\alpha) \in \{0,1\}$ und seien $z_1,\ldots,z_n$ unabhängig und identisch verteilte Beobachtungen gemäß der Verteilung $F_Z(z)$. Dann gilt für die Menge $Q(z,\alpha)$, $\alpha \in \Lambda$, mit Wahrscheinlichkeit $(1-2\eta)$, $0 < \eta \leq 1/2$, die Ungleichung:

$$R(\widetilde{\alpha}_n) - \inf_{\alpha\in\Lambda} R(\alpha) \quad \leq \quad \sqrt{\frac{H^{\Lambda}_{ann}(2n) - \ln(\eta/4)}{n}} - \sqrt{\frac{-\ln\eta}{2n}} + \frac{1}{n}.$$

Beide Sätze ergeben sich unter Ausnutzung von Satz 2.19 durch ähnliche Umformungen wie sie am Anfang dieses Abschnitts für finite Funktionenmengen vorgestellt wurden.

Für Satz 2.21 wird ebenfalls wieder die Chernoff-Ungleichung (Chernoff, 1952) genutzt, um die Aussage ableiten zu können, dass mit Wahrscheinlichkeit $(1-\eta)$ die Ungleichung

$$\inf_{\alpha\in\Lambda} R(\alpha) \;\geq\; R_{emp}(\widetilde{\alpha}_n) - \sqrt{\frac{-\ln\eta}{2n}}$$

gilt. Insgesamt ist die Bedingung

$$\frac{H^{\Lambda}_{ann}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0$$

in Abhängigkeit von der annealed VC-Entropie hinreichend für eine exponentielle Konvergenzrate des ERM-Prinzips und damit auch hinreichend für die nicht-triviale Konsistenz dieses Prinzips für eine Menge von Indikatorfunktionen bei Annahme unabhängig gezogener Beobachtungen. Die Schranken in den Sätzen 2.19 bis 2.21 hängen dabei alle von einem Wert $\mathcal{E}$ ab, der die Schnelligkeit der Konsistenz widerspiegelt. Wird in dem Ausdruck

$$\mathcal{E} = \mathcal{E}(n,\eta,H^{\Lambda}_{ann}) = \frac{H^{\Lambda}_{ann}(2n) - \ln(\eta/4)}{n}$$

die annealed VC-Entropie durch die Growth-Funktion ersetzt, ergeben sich wegen der Abschätzung

$$H^{\Lambda}_{ann}(n) \;\leq\; G^{\Lambda}(n)$$

etwas weitere Schranken

$$R(\widetilde{\alpha}_n) \;\leq\; R_{emp}(\widetilde{\alpha}_n) + \sqrt{\mathcal{E}(n,\eta,G^{\Lambda})} + \frac{1}{n}$$

und

$$R(\widetilde{\alpha}_n) - \inf_{\alpha\in\Lambda} R(\alpha) \;\leq\; \sqrt{\mathcal{E}(n,\eta,G^{\Lambda})} - \sqrt{\frac{-\ln\eta}{2n}} + \frac{1}{n},$$

wobei dann die Schranken für jede Verteilung $P_Z \in \mathcal{P}$ gültig sind, da der Ausdruck

$$\mathcal{E}(n,\eta,G^{\Lambda}) = \frac{G^{\Lambda}(2n) - \ln(\eta/4)}{n}$$

durch die Verwendung der Growth-Funktion verteilungsfrei ist. Da die Schranken mit $\mathcal{E}(n,\eta,G^{\Lambda})$ obere Schranken für die Schranken mit $\mathcal{E}(n,\eta,H^{\Lambda}_{ann})$ sind, ist

$$\frac{G^{\Lambda}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0$$

eine hinreichende Bedingung für die Nicht-Trivialität der Schranken und damit auch eine hinreichende Bedingung für eine exponentielle Konvergenzrate und gleichzeitig für die nicht-triviale Konsistenz des Prinzips der empirischen Risiko-Minimierung. Im Gegensatz zu der Bedingung unter Verwendung der annealed VC-Entropie lässt sich bei Nutzung der Growth-Funktion auch die Notwendigkeit der Bedingung

$$\frac{G^{\Lambda}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0$$

für die nicht-triviale Konsistenz des ERM-Prinzips zeigen (vgl. Abschnitt 4.9 in Vapnik, 1998). Die Konstruktivität der Schranken kann durch die Abschätzung der Growth-Funktion mit Hilfe der VC-Dimension erreicht werden, da die VC-Dimension ein konstruktives Entropie-Konzept für eine Menge von Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, in Bezug auf die Beobachtungen $z_1, \ldots, z_n$ ist (vgl. Vapnik und Chervonenkis, 1971, 1979).

Durch den Nachweis, dass das Prinzip der empirischen Risiko-Minimierung konsistent in einem nicht-trivialen Sinne ist, und durch die Bestimmung von verteilungsfreien konstruktiven Schranken zur Bestimmung der Schnelligkeit der Konvergenzrate hat das ERM-Prinzip gute asymptotische Eigenschaften und ist damit als Konzept zum statistischen Lernen für ausreichend große Datenmengen geeignet. Bei vergleichsweise kleinen Beobachtungsmengen und einer komplexen Charakteristik der Menge der Funktionen können dagegen die oberen Schranken für das Risiko $R(\widetilde{\alpha}_n)$ relativ groß bleiben, obwohl das empirische Risiko klein ist, da dann der Wert $\mathcal{E}$ groß ist. Um eine solche Situation zu vermeiden, kann der Raum der zulässigen Funktionen gemäß seiner Charakteristik verkleinert werden, bis die Summe aus empirischem Risiko und dem Wert $\mathcal{E}$ minimal ist.
Das Konzept für die Suche eines solchen Trade-Offs, auf das in dieser Arbeit nicht eingegangen wird, baut auf dem ERM-Prinzip und den konstruktiven Schranken mit VC-Dimension auf und heißt *Prinzip der strukturellen Risiko-Minimierung (SRM-Prinzip)* (vgl. Vapnik, 1995, 1998, 1999). Durch die besondere Konstruktion des SRM-Prinzips sind die verteilungsfreien Schranken, die die VC-Dimension für dieses Prinzip nutzen, auch zur Kontrolle der Komplexität in statistischen Modellen nutzbar. Für den Fall der Regression mit quadratischer Verlustfunktion geben Cherkassky et al. (1999) einen

Einblick in die Vorgehensweise. Eine Anwendung für multivariate Spline-Regression geben Kohler, Krzyżak und Schäfer (2002).

Die in diesem Kapitel gemachten Aussagen über die Bedingungen für die Konsistenz und auch für die Konvergenzraten des ERM-Prinzips und damit auch für die Gültigkeit des SRM-Prinzips gelten ausschließlich für unabhängig und identisch verteilte Beobachtungen. Die mathematischen Grundlagen für die zentralen Aussagen zur Konsistenz und zur Konvergenzrate sind die klassischen Gesetze der großen Zahlen und die darauf aufbauenden exponentiellen Schranken zur Überwachung der Konvergenzrate. Um die Lernfähigkeit des ERM-Prinzip für Daten mit Abhängigkeitsstrukturen nachweisen zu können, werden im folgenden Kapitel zuerst geeignete Abhängigkeitsstrukturen vorgestellt, für die die Gesetze der großen Zahlen und auch exponentielle Schranken existieren. Im vierten Kapitel wird für diese Abhängigkeitsstrukturen ein Konzept zur Anpassung der Methoden zur empirischen Risiko-Minimierung entwickelt.

3 Abhängigkeitsstrukturen in Datensätzen

Im zweiten Kapitel ist die Annahme getroffen worden, dass die Daten unabhängig und identisch verteilt erhoben worden sind. Diese Annahme wird insbesondere für die Anwendung der Gesetze der großen Zahlen genutzt. Zusätzlich geht die Unabhängigkeit auch bei den Abschätzungen der Konvergenzrate durch die exponentiellen Schranken ein. Wenn eine Ausweitung des ERM-Prinzips auf Beobachtungen mit Abhängigkeitsstrukturen erfolgreich sein soll, müssen also Abschätzungen vom Typ der Hoeffding- und Chernoff-Ungleichung sowie geeignete Gesetze der großen Zahlen auch für abhängige Datenstrukturen anwendbar sein. In diesem Kapitel werden Konzepte für Abhängigkeiten allgemeingültig für beliebige stochastische Prozesse eingeführt, die die weitere Nutzung der Gesetze der großen Zahlen bei nur wenigen zusätzlichen Voraussetzungen an die Beobachtungen erlaubt. Weiterhin werden für sie geeignete exponentielle Schranken entwickelt, so dass im vierten Kapitel mit diesen Ergebnissen die zentralen Sätze des Prinzips der empirischen Risiko-Minimierung auf Daten mit Abhängigkeitsstrukturen ausgeweitet werden können.
Die Einführung der stochastischen Prozesse und der Abhängigkeitsstrukturen erfolgt in diesem Kapitel aus Gründen der einfacheren Darstellung für eine Folge von Zufallsvariablen $\{X_\tau, \tau \in \mathcal{T}\}$. Die Ergebnisse zur Martingal- und Mixingal-Theorie werden dementsprechend ebenfalls für Folgen in der Regel reellwertiger Zufallsvariablen $X_\tau \in I\!R$ angegeben. Die Anwendung der vorgestellten Abhängigkeitskonzepte und die Verwendung der Ergebnisse auf die Beobachtungen $z_1, \ldots, z_n$ und auf die zugehörigen Verluste $Q(z_1, \alpha), \ldots, Q(z_n, \alpha)$, im Kontext des ERM-Prinzip im vierten Kapitel bleibt dadurch trotzdem möglich.

3.1 Stochastische Prozesse

Beobachtungen treten in vielen statistischen Anwendungen als Daten auf, die in einem zeitlichen Kontext erhoben worden sind. Eine solche Datenstruktur lässt sich mit Hilfe stochastischer Prozesse, die den Kolmogorovschen Wahrscheinlichkeitsaxiomen genügen, beschreiben. Beobachtungen werden gemacht, während der Prozess fortschreitet, das heißt, es wird eine Familie von Zufallsvariablen $\{X_t, t \in \mathbb{T}\}$ mit der Variablen

X_t im Zeitpunkt t und der Indexmenge $\mathcal{T} = \mathbb{T}$, die den Zeitraum spezifiziert, beobachtet. Dieser Ansatz ist durch die Bedeutung zeitlicher empirischer Prozesse in der statistischen Analyse geprägt. Es besteht jedoch keine Notwendigkeit, $\mathcal{T}$ ausschließlich als Teilmenge der reellen Zahlen $\mathbb{R}$ zu betrachten bzw. die Menge $\mathcal{T}$ als Zeitraum aufzufassen. Die Theorie der stochastischen Prozesse kann unabhängig von einem zeitlichen Kontext entwickelt werden. Ein bekanntes Beispiel sind Raum-Zeit-Prozesse, die sich sowohl entlang einer Zeitachse und zusätzlich im zwei- oder dreidimensionalen Raum entwickeln können. Aber es sind auch stochastische Prozesse ohne Angabe einer Reihenfolge denkbar, also zum Beispiel unabhängig erhobene Beobachtungen. In dieser Arbeit können die Grundlagen zur Theorie der stochastischen Prozesse nur kurz vorgestellt werden, die erste umfassende und ganzheitliche Einführung stochastischer Prozesse stammt von Doob (1953), eine ausführliche Darstellung wird in Rao (1995) gegeben.
Eine allgemeinere Motivation führt zu der Interpretation, dass ein stochastischer Prozess eine Familie von Zufallsvariablen ist, die eine gewisse Verbindung, Verknüpfung oder Verwandtschaft untereinander auszeichnet.

3.1 Definition
Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein Wahrscheinlichkeitsraum, sei $\mathcal{T}$ eine beliebige Menge und sei $\mathbb{R}^{\mathcal{T}}$ das Kreuzprodukt von $\mathbb{R}$ über alle Elemente aus $\mathcal{T}$. Dann ist die messbare Abbildung $x : \Omega \to \mathbb{R}^{\mathcal{T}}$ mit

$$x(\omega) = \{X_\tau(\omega), \tau \in \mathcal{T}\}$$

ein *stochastischer Prozess*. $\mathcal{T}$ heißt *Indexmenge* und die Zufallsvariable $X_\tau(\omega)$ heißt *Koordinate* des Prozesses.

Ein stochastischer Prozess kann auch als eine Abbildung von $\Omega \times \mathcal{T}$ nach $\mathbb{R}$ aufgefasst werden, dann wird vorausgesetzt, dass X_τ für jedes τ eine messbare Zufallsvariable ist. Mit der Forderung nach *gemeinsamer* Messbarkeit in der Definition 3.1 wird dann allerdings eine gemeinsame Verteilung der $X_\tau(\omega)$ gefordert. Die Anforderungen an die Abbildung sind also in obiger Definition strenger.
Die Indexmenge $\mathcal{T}$ ist eine Menge mit beliebiger Struktur, die im Allgemeinen nicht

geordnet sein muss. Trotzdem gilt für viele relevante Fälle der statistischen Inferenz eine lineare Ordnung der Menge. Ein einfaches Beispiel für eine geordnete Indexmenge ist $\mathcal{T} = \{1, \ldots, n\}$, dann ist $x(\omega)$ ein n-dimensionaler Zufallsvektor

$$(X_1(\omega), \ldots, X_n(\omega)).$$

Wenn $\mathcal{T}$ wiederum geordnet und abzählbar, aber infinit ist, ist $x(\omega) = \{X_\tau(\omega), \tau \in \mathcal{T}\}$ eine Folge

$$\ldots, X_{\tau_{t-1}}(\omega), X_{\tau_t}(\omega), \ldots, X_{\tau_s}(\omega), X_{\tau_{s+1}}(\omega), \ldots$$

und somit ein Prozess mit diskretem Index. Ist dagegen $\mathcal{T}$ ein Intervall, wird $x(\omega) = \{X_\tau(\omega), \tau \in \mathcal{T}\}$ zu einem Prozess mit stetigem Index. In diesem Fall ist $x(\omega)$ eine Funktion in der reellen Variable τ und $\mathbb{R}^{\mathcal{T}}$ ist ein Raum von Zufallsfunktionen in τ. Diese Definition mit geordneter, aber überabzählbarer Indexmenge schließt die Möglichkeit nicht aus, dass $\mathcal{T}$ Informationen über den Abstand zwischen zwei Koordinaten X_τ und $X_{\tau^\star}$ bzw. den Indizes τ und $\tau^\star$ beinhaltet (vgl. Abschnitt 2.1 in Doob, 1953). Allerdings wird dieser Fall im Weiteren nicht näher betrachtet. Wenn nicht anders bezeichnet, wird von geordneten und abzählbaren Indexmengen ausgegangen.
Der Abstand zwischen einer Koordinate X_τ und einer in der Ordnung zurückliegenden Koordinate $X_{\tau^\star}$ wird mit *Lag* bezeichnet. Umgekehrt wird der Abstand zu einer folgenden Koordinate als *Lead* bezeichnet. Kann anhand der Ordnung die Länge des jeweiligen Abstandes über die Anzahl k der zwischen X_τ und $X_{\tau^\star}$ liegenden Koordinaten beziffert werden, so wird von einem *Lag der Ordnung k* bzw. einem *Lead der Ordnung k* gesprochen.

Ist $\mathcal{T}$ eine abzählbare Teilmenge der reellen Zahlen $\mathbb{R}$, dann ist $x(\omega)$ ein Prozess mit abzählbarer und linear geordneter Indexmenge $\mathcal{T}$. Wenn die X_τ Zufallsvariablen über die Zeit repräsentieren und im Falle äquidistanter Zeitabstände, ist es keine Einschränkung anzunehmen, dass die Indexmenge durch die natürlichen oder ganzen Zahlen $\mathbb{N}$ bzw. $\mathbb{Z}$ dargestellt wird. Für eine solche Indexmenge $\mathcal{T} = \mathbb{T}$, die einen zeitlichen Verlauf verkörpert, werden die Indizes $t \in \mathbb{T}$ auch als *Zeitverlauf* bezeichnet.

3.2 Definition
Ein stochastischer Prozess $x(\omega)$ mit abzählbarer Indexmenge $\mathcal{T} = \mathbb{T} \subseteq I\!R$, so dass

$$x(\omega) = \{X_t(\omega), t \in \mathbb{T}\}$$

gilt, heißt *stochastische Folge*. Falls $\mathbb{T} = I\!N$, gilt die Bezeichnung $\{X_t(\omega)\}_1^\infty$, für $\mathbb{T} = \mathbb{Z}$ gilt $\{X_t(\omega)\}_{-\infty}^{\infty}$.

Es ist für alle stochastischen Prozesse, die mit einem Startwert t_0 beginnen, keine Einschränkung der Allgemeinheit, $t_0 = 1$ festzulegen, denn für jeden anderen Prozess kann die Indizierung passend verschoben werden, so dass mit den abzählbaren Indexmengen $I\!N$ und $\mathbb{Z}$ alle stochastischen Prozesse dargestellt werden können (vgl. Abschnitt 1.1 in Rao, 1995).
Ein stochastischer Prozess kann auf zwei Arten betrachtet werden. Bisher wurde hier von einer Folge von Zufallsvariablen mit Index τ ausgegangen. Wenn dagegen der Prozess als eine Funktion in τ betrachtet wird, ist er ein Bündel von sogenannten *Trajektorien*, das heißt, für jedes feste $\omega \in \Omega$ ist $X.(\omega)$ eine Funktion von $\mathcal{T}$ in die reellen Zahlen $I\!R$.

3.3 Definition
Die Menge der Funktionen $\{X.(\omega) : \mathcal{T} \to I\!R, \omega \in \Omega\}$ heißen *Realisationen* oder *Trajektorien* des stochastischen Prozesses $\{X_\tau(\omega), \tau \in \mathcal{T}\}$. Für eine stochastische Folge $\{X_t(\omega), \omega \in \Omega\}$ heißen die Realisationen $\{X.(\omega) : T \to I\!R, \omega \in \Omega\}$ auch *Pfade*.

Ein stochastischer Prozess ist die Modellvorstellung, in die bestimmte Annahmen über den Bildungsmechanismus und die gemeinsame Verteilung der Koordinaten $X_\tau(\omega)$ eingehen. Eine Trajektorie ist eine Folge reeller Zahlen mit Index τ, also die Realisation eines stochastischen Prozesses für ein gegebenes festes $\omega \in \Omega$.
Im Fall $\mathcal{T} = \mathbb{T}$ ist die Realisation einer stochastischen Folge $\{X_t(\omega), t \in \mathbb{T}\}$ein zufällig gezogener Zeitpfad aus dem Bündel aller Pfade. Diese Folge von Beobachtungen heißt dann Zeitreihe.

3.4 Definition
Eine Folge von Beobachtungen $\{x_t, t \in \mathbb{T}_0\}$, $\mathbb{T}_0 \subseteq \mathbb{T}$, die durch die Realisation eines Ereignisses $\omega \in \Omega$ aus der Menge der Zeitpfade $\{X.(\omega) : \mathbb{T} \to I\!R\}$ gezogen wird, heißt *Zeitreihe*.

Eine Zeitreihe $\{x_t, t \in \mathbb{T}_0\}$ kann eine infinite Realisation einer stochastischen Folge sein, in diesem Fall muss $\mathbb{T}_0$ eine unendliche Teilmenge der Indexmenge $\mathbb{T}$ sein. Im Allgemeinen wird dann $\mathbb{T}_0 = \mathbb{T}$, also $\mathbb{T}_0 = I\!N$ oder $\mathbb{T}_0 = \mathbb{Z}$ gelten und damit $\{x_t, t \in \mathbb{T}_0\} = \{x_t\}_1^\infty$ bzw. $\{x_t, t \in \mathbb{T}_0\} = \{x_t\}_{-\infty}^\infty$. In vielen Fällen ist eine Zeitreihe allerdings endlich, da von einem Startpunkt $t = 1$ aus eine Folge von n Beobachtungen gezogen wird. Dann ist $\mathbb{T}_0$ eine echte Teilmenge der Indexmenge $\mathbb{T}$ mit $\mathbb{T}_0 = \{1, \ldots, n\}$. Im letzteren Fall ist die Zeitreihe somit in den realisierten Zeitpfad, und damit in die unendliche Folge möglicher Beobachtungen, eingebettet.

Oft wird eine Zeitreihe als eine zeitlich geordnete Folge von Beobachtungen einer einzigen Zufallsvariablen betrachtet. In diesem Fall ist die Zeitreihe eine Realisation einer stochastischen Folge mit identisch verteilten Koordinaten $X_t(\omega)$, $t \in \mathbb{T}$, das bedeutet, die einzelnen Koordinaten sind identische Kopien einer Zufallsvariablen. Das klassische Beispiel einer solchen Zeitreihe ist der *White Noise Prozess* mit identisch und sogar unabhängig erhobenen Koordinaten, die jeweils Erwartungswert Null haben. Dieser Prozess findet seine Anwendung nicht nur in der Modellierung stochastischer Prozesse, sondern auch im Regressionskontext (vgl. Davidson und MacKinnon, 1993). Andere wichtige Beispiele für Zeitreihen, die nicht notwendig identisch verteilte Koordinaten besitzen müssen, sind beispielsweise *Moving Average Prozesse (MA-Prozesse)* und *Autoregressive Prozesse (AR-Prozesse)* sowie die Kombinationen aus diesen Prozessen, die sogenannten *ARMA-Prozesse*. Eine ausführliche Einführung in die Zeitreihentheorie gibt beispielsweise Brockwell und Davis (1991).

Für die Analyse stochastischer Prozesse $\{X_t(\omega)\}_{-\infty}^\infty$ bzw. $\{X_t(\omega)\}_1^\infty$ spielt das Grenzverhalten für $t \to \infty$ eine große Rolle. Existiert ein Grenzwert so wirkt entlang des Prozesses eine Dämpfung, der Prozess beruhigt sich. Häufig ist der Grenzwert ein geeigneter Parameter des Prozesses für die statistische Inferenz. Da stochastische Prozesse

im Allgemeinen nicht deterministisch sind, kann auch nicht davon ausgegangen werden, dass für alle $\omega \in \Omega$ die jeweiligen Prozessrealisationen gegen denselben Grenzwert $X(\omega)$ konvergieren. Dementsprechend ist das Konzept der *fast sicheren Konvergenz* sinnvoller. Dazu reicht es auch, dass für alle $\omega \in C \subseteq \Omega$ mit $P(C) = 1$ die Konvergenz $X_t(\omega) \to X(\omega)$ für $t \to \infty$ gilt. Die Notationen

$$X_t \xrightarrow{f.s.} X, \quad X_t \to X \textit{ fast sicher} \text{ für } t \to \infty \quad \text{sowie} \quad \lim_{t\to\infty} X_t = X \textit{ fast sicher}$$

werden im Weiteren gleichberechtigt genutzt. Ein weiteres Konzept ist die *Konvergenz nach Wahrscheinlichkeit*,

$$X_t \xrightarrow{Pr} X \quad \text{für } t \to \infty,$$

für das für alle $\varepsilon > 0$ die Eigenschaft $\lim_{t\to\infty} P(\omega : |X_t(\omega) - X(\omega)| < \varepsilon) = 1$ gelten muss. Dieses Konzept ist ein schwächeres Konzept als die fast sichere Konvergenz. Es gilt für beliebige stochastische Prozesse, dass diese bei fast sicherer Konvergenz auch nach Wahrscheinlichkeit konvergieren. Die Umkehrung gilt im Allgemeinen nicht.
Die Anwendung der verschiedenen Konvergenz-Konzepte ist nicht nur zur Analyse des Grenzverhaltens des Prozesses selbst von Bedeutung. Die Konvergenz der Summe $\sum_{t=1}^{n} X_t$ oder des Beobachtungsmittels $\frac{1}{n}\sum_{t=1}^{n} X_t$ ist ebenso von herausragender Bedeutung für statistische Inferenz. Die fast sichere Konvergenz der Summe bzw. des Mittels unter gewissen Voraussetzungen ist dann ein *starkes Gesetz der großen Zahlen*. Für ein *schwaches Gesetz der großen Zahlen* muss unter meist schwächeren Bedingungen Konvergenz nach Wahrscheinlichkeit gelten. Die Gesetze der großen Zahlen gelten für stochastische Prozesse mit unabhängigen Koordinaten unter schwachen Annahmen, wie beispielsweise die quadratische Integrierbarkeit des Prozesses. Ähnliche Gesetze für Prozesse mit unterschiedlichen Abhängigkeitsstrukturen werden im Weiteren vorgestellt.

3.2 Abhängigkeit

Sollen Zusammenhänge zwischen den Zufallsvariablen X_t und X_{t-k} allgemein in der Form modelliert werden, dass die gemeinsame Verteilung dieser Zufallsvariablen vom (Zeit-) Punkt t und vom Lag k abhängt, so ist es nur bedingt sinnvoll, die gesamte

Folge von Zufallsvariablen $\{X_t(\omega)\}$ als einen Punkt in einem Wahrscheinlichkeitsraum zu betrachten. Für die Analyse der Abhängigkeitsstrukturen zwischen den einzelnen Elementen dieser Zufallsfolge gibt es eine hilfreiche Analogie zur Analyse verschiedener Folgen, also unterschiedlicher Beobachtungen $\omega \in \Omega$. Auf dem Wahrscheinlichkeitsraum Ω ist eine injektive Abbildung $T : \Omega \to \Omega$ eine Regel zum paarweisen Abgleich einer Realisation mit einer anderen in Ω. Andererseits wird jede Realisation ω auf eine unendliche Folge abgebildet, so dass T die Abbildung einer Folge auf eine andere induziert.

3.5 Definition
Auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ sei $T : \Omega \to \Omega$ eine injektive, messbare Abbildung. Dann heißt T *Maß-erhaltend*, falls $P(TE) = P(E)$ für alle $E \in \mathcal{A}$ gilt. Weiter heißt die Transformation durch die Abbildung T *Shift-Transformation* der Folge $\{X_t(\omega)\}_{-\infty}^{\infty}$, wenn

$$X_t(T\omega) = X_{t+1}(\omega)$$

gilt. Die Transformation mittels der inversen Abbildung $T^{-1} : \Omega \to \Omega$ heißt *Backshift-Transformation*:

$$X_t(T^{-1}\omega) = X_{t-1}(\omega).$$

Die Abbildung T heißt *Shift-Operator*, die Inverse T^{-1} *Backshift-Operator*.

Die Transformation T bildet also für alle t eine Realisation ω auf die Realisation ab, bei der der Wert X, der unter ω im Zeitpunkt t realisiert wird, nun im Zeitpunkt $(t+1)$ auftritt. Das heißt, jede Koordinate einer Folge wird mit dem vorherigen Periodenwert neu indiziert ($X_{t+1}(\omega) = X_t(\omega^\star)$ mit $\omega^\star = T\omega$). Diese Shift-Transformation ist nicht nur auf einschrittige Verschiebungen eingeschränkt. Allgemeiner kann die Beziehung zweier Punkte, die k-Schritte auseinander liegen, durch $X_t(T^k\omega) = X_{t+k}(\omega)$ dargestellt werden, so dass die Abbildung T^k die Charakteristik des Zusammenhangs zwischen den Koordinaten X_t und X_{t+k} darstellt.
Durch die Verknüpfung einer Zufallsvariablen $X_{t_0}(\omega) : \Omega \to \mathbb{R}$ und des Shifts T bzw. des Backshifts T^{-1} zu einem beliebigen, aber festen Zeitpunkt t_0 kann die gesamte unendliche Folge $\{X_t(\omega)\}_{-\infty}^{\infty}$ beschrieben werden. Durch die Anwendung von T^{-1} und

geeignete Iterationen wird die gesamte Folge entwickelt. Es können so viele Punkte innerhalb des Prozesses erzeugt werden wie benötigt. Das hat sehr vorteilhafte Auswirkungen auf die Analyse der Abhängigkeiten innerhalb des stochastischen Prozesses, da alle Informationen zu den Generierungsgesetzmäßigkeiten des stochastischen Prozesses so im Shift-Operator gebündelt werden (vgl. Abschnitt 13.1 in Davidson, 1994). Der Wahrscheinlichkeitsraum ist dabei äußerst komplex, denn zu jedem Punkt $\omega \in \Omega$ muss eine abzählbar unendliche Menge von Punkten $T^k\omega \in \Omega$ gehören, mit der die gesamte Folge reproduziert werden kann, unabhängig von dem Wert, der bezüglich der Zufallsvariablen X_1 realisiert wird. Die zeitlichen Zusammenhänge innerhalb einer stochastischen Folge können somit als Vergleich verschiedener Folgen aufgefasst werden, dem Original und der um k Perioden in die Zukunft verschobenen Folge.

Diese Darstellungsmöglichkeit ist deshalb so wichtig, damit statistische Inferenz an Zeitreihen überhaupt betrieben werden kann. Die zur Verfügung stehenden Daten entstammen meist einer einzelnen Realisation einer Zufallsfolge. Aber obwohl eine einzelne Realisation der Funktionswert für ein einziges ω ist, ist diese Realisation durch den Shift-Operator per Konstruktion doch mit abzählbar unendlich vielen ω verknüpft, abhängig von der Konstruktion der Leads und Lags der Folge. Durch die einzelne Realisation wird somit eine große Anzahl an ω abgebildet, diese Menge an Ereignissen muss aber speziell genug sein, um mit dieser einzelnen Beobachtung Inferenz bezüglich der Verteilung $P \in \mathcal{P}$ betreiben zu können. Zu diesem Zweck werden einige Restriktionen an das Verhalten der Zufallsfolge eingeführt. Die wichtigsten und am besten untersuchten sind die Begriffe *Unabhängigkeit* und *Stationarität*. Diese Konzepte haben den Vorteil, dass Parameterschätzungen und Aussagen zur statistischen Güte in stochastischen Modellen möglich sind, so gibt es in der Literatur zahlreiche Ergebnisse zur Wahrscheinlichkeits- und Konvergenztheorie. Der Nachteil ist das hohe Maß an Einschränkungen im Verhalten der Folge, wenn Unabhängigkeit bzw. Stationarität vorausgesetzt wird.

Von großer Bedeutung ist der Grad der Beziehung einer Koordinate zu ihren Nachbarn bei zufälligen Änderungen. Bei zeitlicher Ordnung des stochastischen Prozesses spiegelt diese Abhängigkeit das *Gedächtnis* wider. Vor allem interessiert der Zugewinn an

Information im gegenwärtigen Zustand im Vergleich zum Informationsgehalt der vorherigen Zustände. Bei einer Folge ohne Gedächtnis muss zu jedem Zeitpunkt der volle Informationsgehalt immer wieder neu und vollständig in den Koordinaten enthalten sein, das bedeutet gleichzeitig, die zeitliche Ordnung verliert für den stochastischen Prozess die Aussagekraft. Eine solche Folge heißt *seriell unabhängig*, formal bedeutet das, die Elemente der Folge $\{X_t(\omega)\}_{-\infty}^{\infty}$ sind paarweise unabhängig von den Elementen der Folge $\{X_t(T^k\omega)\}_{-\infty}^{\infty}$ für alle $k > 0$. Dies ist äquivalent zur *totalen Unabhängigkeit* jeder endlichen Folge von Koordinaten aus $\{X_t(\omega)\}_{-\infty}^{\infty}$, also

$$P\left(\{X_t(\omega)\}_k^l\right) = \prod_{t=k}^{l} P(X_t(\omega)) \text{ für alle } k, l \text{ mit } k < l.$$

Das ist eine deutlich stärkere Aussage im Vergleich beispielsweise zur paarweisen Unabhängigkeit aller (X_i, X_j) aus der Folge $\{X_t(\omega)\}_{-\infty}^{\infty}$. Serielle Unabhängigkeit ist eine Annahme an die gesamte Folge und damit eine einfache, aber auch sehr einschränkende Annahme an das Gedächtnis eines stochastischen Prozesses.
Aus dem Blickwinkel der Verteilungsannahme an den gesamten Prozess ist eine weitere mögliche Annahme, dass die gemeinsame Verteilung $P \in \mathcal{P}$ aller Koordinaten invariant bezüglich des Shift-Operators T ist.

3.6 Definition
Eine Zufallsfolge $\{X_t(\omega)\}_{-\infty}^{\infty}$ heißt *strikt stationär*, falls $\{X_t(\omega)\}_{-\infty}^{\infty}$ und $\{X_t(T^k\omega)\}_{-\infty}^{\infty}$ für alle $k \in \mathbb{T}$ dieselbe gemeinsame Verteilung haben.

Die obige Definition ist äquivalent zu der Aussage, dass die Shift-Transformation T Maß-erhaltend sein muss (vgl. Davidson, 1994, S. 193). Andererseits gibt es in diesem Sinne auch schwächere Bedingungen an die gemeinsame Verteilung des stochastischen Prozesses, falls die ersten beiden Momente der Koordinaten existieren. So ist es häufig sinnvoll zu verlangen, dass der Erwartungswert und die Varianz aller Koordinaten invariant gegenüber Zeitverschiebungen sind, und dass die Kovarianzen des stochastischen Prozesses nur vom Abstand der Zeitpunkte, also vom Lag, abhängen.

3.7 Definition
Sei $\mathrm{E}(|X_t|^2) < \infty$ und sei $\mu_t = \mathrm{E}(X_t)$ und $\gamma_{kt} = \mathrm{Cov}(X_t, X_{t+k})$ für alle $t \in \mathbb{Z}$ und $k \in \mathbb{N}_0$, so dass $\{\mu_t\}_{t=-\infty}^{\infty}$ und $\{\{\gamma_{kt}\}_{k=-\infty}^{\infty}\}_{t=-\infty}^{\infty}$ wohl definiert sind. Dann heißt eine Zufallsfolge $\{X_t(\omega)\}_{-\infty}^{\infty}$

(a) *mittelwertstationär*, falls $\mu_t = \mu$ für alle t und μ konstant.

(b) *kovarianzstationär*, falls $\gamma_{kt} = \gamma_k$ für alle t und $\{\gamma_k\}_{-\infty}^{\infty}$ eine Folge von Konstanten.

(c) *schwach stationär*, falls sie mittelwert- und kovarianzstationär ist.

Das Konzept der Stationarität ist verschieden von dem der Annahme identischer Verteilungen aller Koordinaten X_t, denn sowohl bei der strikten Stationarität als auch bei schwächeren Varianten werden auch immer Bedingungen an die gemeinsame Verteilung mit den Nachbarn im Prozess gemacht. Andererseits ist serielle Unabhängigkeit und die identische Verteilung hinreichend, aber nicht notwendig für die Stationarität einer Zufallsfolge. Das Konzept der Stationarität impliziert nicht Homogenität im Auftreten der Werte des Prozesses oder das Fehlen periodischer Muster. Die eigentliche Interpretation der Stationarität ist, dass das Verteilungsmuster des stochastischen Prozesses nicht vom Zeitindex t abhängt. Durch die Annahme der Stationarität wird es möglich, statistische Aussagen über den Prozess zu machen, obwohl es nur eine einzige Realisation gibt. Die Stationarität erlaubt es, die Verschiebung auf der Zeitachse als zusätzliche Realisation bei Verteilungsgleichheit (strikt stationär) bzw. bei gleichen Momenten (schwach stationär) zu betrachten. Stationarität ist eine sehr einschränkende Annahme, da in der Praxis häufig sowohl Trends als auch verschiedene saisonale Muster auftreten, also Nicht-Stationarität vorliegt. Aus diesem Grund ist es sinnvoll, zwischen lokaler Nicht-Stationarität und globaler Nicht-Stationarität zu unterscheiden. Im ersten Fall kann die Schwankung in den Momenten durch lokale Glättung bearbeitet werden, im globalen Fall sind dazu zum Beispiel die Bereinigung des Prozesses von Trend- oder saisonalen Einflüssen nötig.

Der informative Wert einer Realisation bei fester Länge eines stochastischen Prozesses hängt vom Grad der Abhängigkeit in dem Prozess ab. Auf der einen Seite ist eine Folge mit unabhängigen und identisch verteilten Koordinaten äquivalent zu einer einfachen

Zufallsstichprobe, die Information jeder Beobachtung geht vollständig in die statistische Analyse ein und alle klassischen statistischen Gesetze behalten ihre Gültigkeit. Auf der anderen Seite gibt es Prozesse, bei denen der Abhängigkeitsgrad so hoch ist, dass durch eine einzelne Realisation kaum Aussagen über die Parameter der wahren Verteilung des Prozesses gemacht werden können, selbst wenn die Länge (Stichprobenumfang) der Realisation gegen unendlich strebt. Ein wichtiges Beispiel für dieses Phänomen ist das Verhalten des Prozesses im zeitlichen Verlauf im Gegensatz zum Verhalten in einem Zeitpunkt t_0. Betrachtet man dazu die gemittelten Werte eines stochastischen Prozesses, so heißt das Mittel über mehrere Realisationen ω_i, $i = 1, \ldots, N$ in einem festen Zeitpunkt t_0

$$\overline{X}_{N,t_0} = \frac{1}{N} \sum_{i=1}^{N} X_{t_0}(\omega_i),$$

vertikales Mittel, das zeitliche Mittel mit einer Realisation der Länge n für ein $\omega_i \in \Omega$ ist

$$\overline{X}_n(\omega_i) = \frac{1}{n} \sum_{t=1}^{n} X_t(\omega_i).$$

Im Allgemeinen gilt für die Grenzwerte beider Mittel bezüglich N respektive n nicht, dass sie identisch sind. Das vertikale Mittel konvergiert für $N \to \infty$ bei unabhängiger Ziehung der ω_i gegen den marginalen Erwartungswert $\mathrm{E}(X_{t_0})$ im Zeitpunkt t_0. Für das zeitliche Mittel ist im Allgemeinen nicht zu erwarten, dass es gegen einen festen Wert konvergiert. Ist der Prozess nicht stationär ist dies intuitiv klar, aber selbst im Falle von Stationarität tritt die Konvergenz nicht notwendig ein, denn Zufallseffekte, die die Werte der verschiedenen Realisationen beeinflussen, müssen nicht notwendig vom Zeitindex t abhängen. Andererseits ist die Konsistenz des zeitlichen Mittels wünschenswert, um Aussagen über das Grenzverhalten des Mittels eines stochastischen Prozesses aus einer einzigen Realisation machen zu können.

Einen Ausweg bietet hier das Konzept der Invarianz für stochastische Prozesse. In einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ heißt ein Ereignis $E \in \mathcal{A}$ *invariant unter der Transformation* T, falls die *symmetrische Differenz* $TE \triangle E = (TE \cup E) - (TE \cap E)$ der Mengen E und TE eine Nullmenge ist, also $P(TE \triangle E) = 0$ gilt. Die Menge der unter T invarianten Ereignisse ist eine σ-Algebra und sei bezeichnet durch $\mathcal{I}$. Eine Zufallsvariable $X(\omega)$ ist *invariant*, falls sie $\mathcal{I}$-$\mathcal{B}$-messbar ist, wobei $\mathcal{B}$ die zugehörige Borelmenge

über den reellen Zahlen bezeichnet, und hat damit die Eigenschaft $X(T\omega) = X(\omega)$. Daraus folgt, dass eine Folge invarianter Zufallsvariablen $\{X_t(\omega)\}_1^\infty$ trivial im Sinne von $X_t(\omega) = X_1(\omega)$ fast sicher für alle t ist. Dabei wird die Einschränkung auf einseitige stochastische Prozesse mit Indexmenge $\mathbb{T} = \mathbb{N}$ nur zur einfacheren Darstellung gewählt. Eine Generalisierung der Ergebnisse auf zweiseitige Prozesse $\{X_t(\omega)\}_{-\infty}^\infty$ ist ohne weitere Schwierigkeiten möglich und wird dort, wo es im Weiteren wichtig ist, explizit genannt.
Dies bedeutet in der Gesamtheit, dass sich ein stochastischer Prozess über die Zeit hinweg auf der Menge der unter der Shift-Transformation T invarianten Ereignisse $\mathcal{I}$ nicht verändert und führt somit zu einer ersten Aussage über das Grenzverhalten eines stochastischen Prozesses.

3.8 Satz *Doob (1953)*
Sei $\{X_t(\omega)\}_1^\infty$ ein stationärer Prozess, definiert durch eine messbare Abbildung X_1 und eine Shift-Transformation T, die Maß-erhaltend ist, so dass $X_t(\omega) = X_1(T^{t-1}\omega)$ für alle t. Sei $\mathcal{I}$ die σ-Algebra der invarianten Ereignisse $\omega \in \Omega$. Falls $\mathrm{E}(|X_1|) < \infty$, gilt

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} X_t(\omega) = \mathrm{E}(X_1|\mathcal{I}) \quad \text{fast sicher.}$$

Der Grenzwert für das zeitliche Mittel bei stationären Prozessen existiert demnach und entspricht dem Erwartungswert der durch die invarianten Ereignisse bedingten Verteilung der Zufallsvariablen X_{t_0} in einem beliebigen Zeitpunkt t_0.
Existieren bezüglich T invariante Ereignisse E, so gibt es Regionen im Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$, die von bestimmten Realisationen des Prozesses nicht erreicht werden können, denn es gilt wegen $P(TE \triangle E) = 0$ auch $P(TE \cap E^c) = 0$. Gilt andererseits $P(TE \triangle E) = 1$, heißt das, dass das Ereignis E^c mit Wahrscheinlichkeit Null in einem Prozess eintritt, in dem E eingetreten ist. Einige Realisationen eines Prozesses können unter Umständen also gewisse Werte nicht erreichen, dadurch sind das vertikale Mittel und der Grenzwert des zeitlichen Mittels in einem stationären Prozess im Allgemeinen nicht gleich (vgl. Gray, 1988). Der Wahrscheinlichkeitsraum wird nur dann von einer Realisation eines Prozesses fast sicher vollständig erreicht, wenn die unter T invarianten Ereignisse entweder fast sicher oder fast sicher nicht eintreten. Daraus

ergibt sich die Definition, dass eine Maß-erhaltende Transformation T *ergodisch* heißt, falls entweder $P(E) = 1$ oder $P(E) = 0$ für alle $E \in \mathcal{I}$ ist, wobei $\mathcal{I}$ die σ-Algebra der unter T invarianten Ereignisse E ist. Für stationäre Prozesse wird damit ebenfalls Ergodizität definiert.

3.9 Definition
Sei eine Transformation T Maß-erhaltend und ergodisch. Dann heißt ein stationärer Prozess $\{X_t(\omega)\}_1^\infty$ *ergodisch*, wenn $X_t(\omega) = X_1(T^{t-1}\omega)$ für alle t gilt.

Zusammen mit der gerade definierten Annahme und der Tatsache, dass sich bei einem stationären Prozess der Erwartungswert über die Zeit nicht verändert, kann wegen Satz 3.8 ein erstes Gesetz der großen Zahlen beschrieben werden.

3.10 Satz *Ergoden-Theorem*
Sei $\{X_t(\omega)\}_1^\infty$ ein stationärer, ergodischer stochastischer Prozess mit $\mathrm{E}(|X_1|) < \infty$, dann gilt:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} X_t(\omega) = \mathrm{E}(X_1) \quad \text{fast sicher.}$$

Dieses Gesetz nutzt Stationarität, was, wie bereits beschrieben, eine starke Annahme an einen stochastischen Prozess ist. In praktischen Anwendungen ist gerade diese Forderung nach stationärem Verhalten oft nicht haltbar. Auf der anderen Seite hat das Gesetz in dieser Form einen hohen theoretischen Nutzen, da die Ergodizität ein sehr schwaches Abhängigkeitskonzept ist. Eine Maß-erhaltende Transformation T wird mit fortschreitender Zeit die nicht invarianten Ereignisse in A mit denen in A^c vermischen. Durch die Maß-erhaltende Eigenschaft ist die Schnittmenge $TA \cap A^c$ nicht leer und durch wiederholte Anwendung von T wird eine Folge von Mengen $\{T^k A\}_0^\infty$ erzeugt, die verschiedene Mischungen der Elemente aus A und A^c enthält. Allgemein gilt, dass positive Abhängigkeit eines Ereignisses B bezüglich A, also $P(A \cap B) > P(A)P(B)$, die negative Abhängigkeit bezüglich A^c impliziert, d. h. $P(A^c \cap B) < P(A^c)P(B)$. Wenn die Vermischung fortgeschritten ist, also wenn die Transformation T nur oft genug auf ein Ereignis A und sein Komplement A^c angewendet wird, sollte die gemittelte

Abhängigkeit von B bezüglich der Mischungen aus A und A^c verschwinden (vgl. Abschnitt 13.4 in Davidson, 1994). Tatsächlich lässt sich dieser intuitive Ansatz als eine Charakterisierung der Ergodizität nachweisen.

3.11 Satz
Eine Maß-erhaltende Shift-Transformation T ist dann und nur dann ergodisch, wenn für jedes Paar von Ereignissen $A, B \in \mathcal{A}$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P(T^k A \cap B) = P(A)P(B)$$

gilt.

Mit diesem Satz folgt sofort eine Aussage über die Kovarianzen in einem stationären und ergodischen stochastischen Prozess $\{X_t(\omega)\}_1^\infty$ mit $\mathrm{E}(X^2) < \infty$. Dann gilt:

$$\frac{1}{n} \sum_{k=1}^{n} \mathrm{Cov}(X_1, X_k) \longrightarrow 0 \quad \text{für } n \to \infty.$$

Es gilt sogar die Umkehrung, falls der Prozess normalverteilt ist. Allerdings bedeutet diese im Grenzverhalten gemittelte Unabhängigkeit nicht die asymptotische Unabhängigkeit des Prozesses, mit $\mathrm{Cov}(X_1, X_k) \longrightarrow 0$ für $k \to \infty$.
Diese Eigenschaft besitzt dagegen ein sogenannter Mixing-Prozess. Eine Maß-erhaltende, ergodische Shift-Transformation T heißt *Mixing-Transformation*, falls für alle $A, B \in \mathcal{A}$

$$\lim_{n\to\infty} \sum_{k=1}^{n} P(T^k A \cap B) = P(A)P(B)$$

gilt. Ein stationärer, stochastischer Prozess $\{X_t(\omega)\}_1^\infty$ heißt *Mixing-Prozess*, falls $X_t(\omega) = X_1(T^{t-1}\omega)$ für alle t gilt, wobei T eine Mixing-Transformation ist. Für einen Mixing-Prozess gilt für zeitlich auseinander driftende Ereignisse im Grenzverhalten die Unabhängigkeit. Die wiederholte Anwendung der Mixing-Transformation auf ein Ereignis A führt zu einer so gründlichen Vermischung von A und A^c, dass für hinreichend großes k in der Komposition $T^k A$ keine Hinweise mehr über A zu finden sind (vgl. Abschnitt 13.3 in Davidson, 1994). So gilt zum Beispiel für einen reellen stochastischen Prozess, dass die Abhängigkeit der Ereignisse $A = \{\omega | X_t(\omega) \leq a\}$ und

$T^k A = \{\omega | X_{t+k}(\omega) \leq a\}$ mit zunehmenden Lag k voneinander abnimmt. Was intuitiv zu erwarten war, gilt auch formal für das Grenzverhalten der Kovarianz bei einem stationären Mixing-Prozess $\{X_t(\omega)\}_1^\infty$:

$$\mathrm{Cov}(X_1, X_k) \longrightarrow 0 \quad \text{für } k \to \infty.$$

Für eine genauere Einführung und für weitere Ergebnisse zur Ergodentheorie und zum Konzept des Mixing sei hier auf Gray (1988) bzw. auf Philipp und Stout (1975) verwiesen.

Ein ganz anderer Ansatz, die Abhängigkeitsstruktur eines stochastischen Prozesses zu untersuchen, ist es, direkt von deren Bildungsgesetzmäßigkeiten auszugehen. Die direkte Betrachtung der Abhängigkeiten zwischen den Koordinaten funktioniert über die Bildung zugehöriger *Sub-σ-Algebren* von $\mathcal{A}$, die durch den stochastischen Prozess generiert werden. Für den stochastischen Prozess $\{X_t, t \in \mathbb{Z}\}$ wird die Familie von Sub-σ-Algebren $\{\mathcal{A}_s^t, s \leq t\}$ mit $\mathcal{A}_s^t = \sigma(X_s, \ldots, X_t)$ erzeugt. Dabei ist $\mathcal{A}_s^t$ die von $(X_s, \ldots, X_t)$ induzierte σ-Algebra, also die kleinste σ-Algebra, so dass die Koordinaten des Prozesses im Bereich von s bis t messbar sind. Die $\mathcal{A}_s^t$ sind die inversen Bilder des $(t-s)$-dimensionalen Zylinders innerhalb der Borel-Menge $\mathcal{B}^\infty$ auf $\mathbb{R}^\infty$ bezüglich der Abbildung $(X_s, \ldots, X_t)$. Die Grenzen s und t können im Positiven wie im Negativen gegen unendlich laufen. Ein häufig benötigtes Beispiel ist die Familie der steigenden Sub-σ-Algebren $\{\mathcal{A}_{-\infty}^t, t \in \mathbb{Z}\}$, so dass diese Familie aus aufsteigenden Teilmengen

$$\ldots \subseteq \mathcal{A}_{-\infty}^{t-1} \subseteq \mathcal{A}_{-\infty}^{t} \subseteq \mathcal{A}_{-\infty}^{t+1} \subseteq \ldots \subseteq \mathcal{A}_{-\infty}^{\infty} \quad \subseteq \mathcal{A}$$

besteht. Jede σ-Algebra $\mathcal{A}_{-\infty}^t$ kann so interpretiert werden, dass in ihr die Information des gesamten stochastischen Prozesses bis zum Punkt t enthalten ist. Die σ-Algebra $\mathcal{A}_{-\infty}^{\infty} \subseteq \mathcal{A}$, für die der gesamte Prozess messbar ist, ist damit die kleinste durch den Prozess $\{X_t\}$ induzierte σ-Algebra, wobei bei Gleichheit der Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ insgesamt durch den stochastischen Prozess $\{X_t\}$ induziert ist. Für einen stochastischen Prozess $\{X_t, t \in \mathbb{N}\}$ gilt in gleicher Weise für die Familie $\{\mathcal{A}_1^t, t \in \mathbb{N}\}$, dass die Sub-σ-Algebren $\mathcal{A}_1^t = \sigma(X_1, \ldots, X_t)$ in ansteigender Form die Informationen des Prozesses bis zum Punkt t beinhalten.

Der Vorteil dieses Ansatzes ist, dass die Struktur des stochastischen Prozesses auf die Struktur der Familie der Sub-σ-Algebren übertragen wird, wodurch nur noch die für den Prozess wichtigen Ereigniskombinationen betrachtet werden. So müssen nur noch die von $\{X_t\}$ induzierten σ-Algebren auf ihre Abhängigkeitsstrukturen hin untersucht werden. Im Folgenden werden zwei Ansätze dieser Art, das Martingal-Konzept und das darauf aufbauende Konzept der Mixingale, eingeführt.

3.3 Martingale

Der Begriff Martingal hat eine lange Geschichte im spieltheoretischen Kontext und bedeutet ursprünglich das System, Verluste dadurch wieder auszugleichen, indem man den Einsatz nach jedem verlorenen Spiel verdoppelt. Das Konzept der Martingale lässt sich interpretieren als die Summe der Gewinne eines Spielers in einer Serie von fairen Wetten und wird bei Bachelier (1900) kurz erläutert. Wichtige theoretische Arbeiten stammen von Bernstein (1927, 1940) und von Lévy (1935, 1937), allerdings in einer beschränkten Form zur Verallgemeinerung von Grenzwertsätzen für Summen unabhängiger Zufallsvariablen. Der Name Martingal wurde erstmals von Ville (1939) eingeführt. Die daran anschließenden Arbeiten von Doob (1953) haben das Konzept so in die Theorie stochastischer Prozesse eingebettet, dass die Abhängigkeiten der Koordinaten sequentiell durch zugehörige Sub-σ-Algebren bedingt werden.

Die Beobachtungsrichtung bei Realisationen zeitlicher stochastischer Prozesse ist unidirektional. Die aktuelle Koordinate X_t wird von einer Umgebung bestimmt, in der die vorherigen Werte X_{t-k}, $k > 0$, gegeben und durch ihre Vorgänger bedingt deterministisch sind, wogegen die zukünftigen Werte zufällig bleiben. Die Vergangenheit ist bekannt, die Zukunft nicht. Sequentielle Bedingung auf die vergangenen Ereignisse ist deshalb gerade für Zeitreihen von hoher Bedeutung. Dabei wird das partielle Wissen durch die Spezifizierung der Sub-σ-Algebren der Ereignisse in $\mathcal{A}$ charakterisiert, für die bekannt ist, ob jedes der Ereignisse, das zu der σ-Algebra gehört, eingetreten ist oder nicht. Die Kumulierung der Information bei voran schreitender Zeit wird repräsentiert

durch die aufsteigende Folge von σ-Algebren $\{\mathcal{A}_t\}_{-\infty}^{\infty}$, so dass

$$\ldots \subseteq \mathcal{A}_{-1} \subseteq \mathcal{A}_0 \subseteq \mathcal{A}_1 \subseteq \mathcal{A}_2 \subseteq \ldots \quad \subseteq \mathcal{A}$$

gilt. Dabei sei die Bezeichnung $\mathcal{A}_t$ für die induzierten σ-Algebren sowohl für $\{X_k\}_{-\infty}^{t}$ mit $\mathcal{A}_t = \mathcal{A}_{-\infty}^{t}$ als auch für $\{X_k\}_1^t$ mit $\mathcal{A}_t = \mathcal{A}_1^t$ gültig, sofern im Weiteren durch den Zusammenhang keine Verwechslungen möglich sind.

Die Folge $\{\mathcal{A}_t\}_{-\infty}^{\infty}$ heißt *adaptiert* an den stochastischen Prozess $\{X_t\}_{-\infty}^{\infty}$, falls X_t $\mathcal{A}_t$-messbar ist für alle t. Die Paare $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ heißen *adaptierter Prozess* und *minimaler adaptierter Prozess*, falls $\mathcal{A}_t = \sigma(X_s, -\infty < s \leq t)$. Im Allgemeinen wird $\mathcal{A}_t$ als die vollständige vorliegende Information bis zum Zeitpunkt t interpretiert. Wenn X_t integrierbar ist, ist der bedingte Erwartungswert $\mathrm{E}(X_t|\mathcal{A}_{t-1})$ wohldefiniert und ist der beste Vorhersageschätzer bei Ein-Schritt-Prognosen bezüglich der euklidischen Norm. Die Formalisierung der intuitiven Idee aus der Spieltheorie, die erwarteten Gewinne aus einer Serie von fairen Wetten zu modellieren, führt zur Forderung, dass sich der erwartete Wert des Gewinns, bedingt auf das Wissen aus der Vergangenheit, also bedingt auf die adaptierten σ-Algebren, nicht anders verhält als die Zufallsvariable der Vorperiode. Dies führt zur folgenden Definition.

3.12 Definition *Martingal*
Sei $\{S_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein adaptierter Prozess auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$, wobei $\{\mathcal{A}_t\}_{-\infty}^{\infty}$ eine ansteigende Folge von σ-Algebren sei. Wenn für alle t die Bedingungen

(i) $\mathrm{E}(|S_t|) < \infty$,

(ii) $\mathrm{E}(S_t|\mathcal{A}_{t-1}) = S_{t-1}$ fast sicher

gelten, dann heißt diese Folge von Paaren *Martingal*.

Oftmals hat ein Martingal einen Anfangsindex, im Beispiel der fairen Wetten das erste Spiel, dann kann die Folge geschrieben werden als $\{S_t, \mathcal{A}_t\}_1^{\infty}$, wobei der Anfangswert S_1 eine beliebige, integrierbare Zufallsvariable ist. Neben den kumulierten Gewinnen

der Serie von Spielen ist auch der Zugewinn pro Spiel interessant. Dies sind nach obiger Modellierung als Martingal die Differenzen $X_t = S_t - S_{t-1}$ für alle t. Aus den Bedingungen, die an ein Martingal gestellt werden, ergeben sich in natürlicher Weise auch die Bedingungen für die Differenzen.

3.13 Definition *Martingal-Differenz*
Eine Folge von *Martingal-Differenzen* $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ist ein adaptierter Prozess auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$, der die Bedingungen

(i) $\mathrm{E}(|X_t|) < \infty$,

(ii) $\mathrm{E}(X_t|\mathcal{A}_{t-1}) = 0$ fast sicher

für alle t erfüllt.

Natürlich ergibt sich direkt aus der Konstruktion, dass $\{X_t\}$ mit $X_t = S_t - S_{t-1}$ eine Martingal-Differenz ist, wenn $\{S_t\}$ ein Martingal ist. Umgekehrt kann ein Martingal auch als die partielle Summe aus einer Folge von Martingal-Differenzen gebildet werden.

3.14 Beispiel *Random Walk*
Sei $\{X_t\}_1^{\infty}$ eine Folge von i.i.d. Zufallsvariablen mit $\mathrm{E}(X_t) = 0$ für alle t. Für

$$S_n = \sum_{t=1}^{n} X_t \quad \text{und} \quad \mathcal{A}_n = \sigma(X_n, X_{n-1}, \ldots, X_1)$$

ist $\{S_n, \mathcal{A}_n\}_1^{\infty}$ ein Martingal, denn $\mathrm{E}(|S_n|) \leq \sum_{t=1}^{n} \mathrm{E}(|X_t|) < \infty$ und wegen Unabhängigkeit und identischer Verteilung der X_t gilt

$$\mathrm{E}(S_n|\mathcal{A}_{n-1}) = \mathrm{E}(X_n|\mathcal{A}_{n-1}) + S_{n-1} = S_{n-1}.$$

Die Folge $\{S_n\}_1^{\infty}$ heißt *Random Walk*.

Eine allgemeingültige Definition eines Martingals $\{S_n, \mathcal{A}_n\}_{-\infty}^{\infty}$ durch $S_n = \sum_{t=-\infty}^{n} X_t$ kann aber zu Problemen führen, da für X_t mit positiver, gleichmäßig beschränkter Varianz über alle t, der Erwartungswert $\mathrm{E}(|S_n|)$ zwar beschränkt für alle n ist, aber

nicht gleichmäßig über alle n. Außerdem existieren Martingale, die nicht als Summe von Martingal-Differenzen erklärt sind, wie das folgende Beispiel zeigt.

3.15 Beispiel
Sei Z integrierbare und $\mathcal{A}$-$\mathcal{B}$-messbare Zufallsvariable mit $\mathrm{E}(Z) = 0$. Sei weiter $\{\mathcal{A}_n\}_1^\infty$ eine aufsteigende Folge von σ-Algebren mit $\lim\limits_{n\to\infty} \mathcal{A}_n = \mathcal{A}$ und sei $S_n = \mathrm{E}(Z|\mathcal{A}_n)$. Dann gilt

$$\mathrm{E}(S_n|\mathcal{A}_{n-1}) = \mathrm{E}\left(\mathrm{E}(Z|\mathcal{A}_n)|\mathcal{A}_{n-1}\right) = \mathrm{E}(Z|\mathcal{A}_{n-1}) = S_{n-1},$$

denn $\mathcal{A}_{n-1} \subseteq \mathcal{A}_n$, und $\mathrm{E}(|S_n|) = \mathrm{E}(Z|\mathcal{A}_n) \leq \mathrm{E}(|Z|) < \infty$, weil $\mathcal{A}_n \subseteq \mathcal{A}$ ist. Daraus folgt, die Folge $\{S_n, \mathcal{A}_n\}_1^\infty$ ist ein Martingal.

Eine zentrale Eigenschaft von Martingalen ist, dass die Differenzen $S_n - S_{n-1}$ für alle n unkorreliert mit den Differenzen aus der Vergangenheit sind. Etwas allgemeiner gilt dies für jede messbare, integrierbare Abbildung der verzögerten Differenzen.

3.16 Satz
Ist $\{X_t, \mathcal{A}_t\}_{-\infty}^\infty$ eine Folge von Martingal-Differenzen, dann gilt

$$\mathrm{Cov}(X_t, \phi(X_{t-1}, X_{t-2}, \ldots)) = 0,$$

wobei ϕ eine beliebige Borel-messbare integrierbare Funktion ist.

3.17 Korollar
Ist $\{X_t, \mathcal{A}_t\}_{-\infty}^\infty$ eine Folge von Martingal-Differenzen, dann gilt

$$\mathrm{E}(X_t X_{t-k}) = 0$$

für alle t und alle $k \neq 0$.

Zum Beweis von Satz 3.16 und Korollar 3.17 sei auf Doob (1953) verwiesen. Insgesamt kann die Eigenschaft der Martingal-Differenzen in Satz 3.16 im Sinne der Abhängigkeitsstruktur stochastischer Prozesse eingeordnet werden zwischen Unkorreliertkeit und

Unabhängigkeit. Allerdings muss dabei die Asymmetrie in Hinsicht auf die Zeit beachtet werden, denn durch die Umkehrung der zeitlichen Ordnung würde zum Beispiel ein unabhängiger Prozess wieder zu einem unabhängigen Prozess. Ebenso gilt dies bei Umkehrung der zeitlichen Ordnung im Zusammenhang mit Unkorreliertheit. Aber die Umkehrung der Ordnung in einer Folge von Martingal-Differenzen führt im Allgemeinen nicht zu einer neuen Folge von Martingal-Differenzen.

Das Martingal-Konzept nimmt eine wichtige Rolle ein, um klassische Resultate in Hinsicht auf Konvergenz von Prozessen zu verifizieren, ohne Unabhängigkeit voraussetzen zu müssen. Das Martingal-Konzept ist aber eine Beschränkung bezüglich des Gedächtnisses des Prozesses. Im Folgenden wird gezeigt, dass das Gesetz der großen Zahlen und Sätze zur Schnelligkeit der Konvergenz unter der Annahme, dass ein stochastischer Prozess eine Martingal-Differenz ist, gültig sind. Die Resultate zu den Konvergenzeigenschaften entstehen aus der Tatsache, dass die Koordinaten eines stochastischen Prozesses zentriert sind, also so transformiert sind, dass $\mathrm{E}(X_t) = 0$ für alle t. Ohne Einschränkung der Allgemeingültigkeit kann der stochastische Prozess $\{X_t - E(X_t)\}_{-\infty}^{\infty}$ an Stelle von $\{X_t\}_{-\infty}^{\infty}$ betrachtet werden. Allerdings ist es mit dieser Konstruktion möglich, dass der mittlere Erwartungswert $\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}(X_t)$ für $n \to \infty$ divergiert, aber der Prozess $\{X_t - E(X_t)\}_{-\infty}^{\infty}$ trotzdem gegen Null strebt. In diesem Fall muss das Gesetz der großen Zahlen anders interpretiert werden, da es keinen Sinn ergibt, von Konvergenz der Folge der Beobachtungsmittel zu sprechen.
Starke Gesetze der großen Zahlen sind für zentrierte stochastische Prozesse über die fast sichere Konvergenz $X_n \xrightarrow{f.s.} X$ derart definiert, dass

$$\frac{1}{n}S_n = \frac{1}{n}\sum_{t=1}^{n} X_t \xrightarrow{f.s.} 0 \quad \text{für} \quad n \to \infty$$

gilt. Beim schwachen Gesetz der großen Zahlen gilt mit der Konvergenz nach Wahrscheinlichkeit $X_n \xrightarrow{Pr} X$:

$$\frac{1}{n}S_n = \frac{1}{n}\sum_{t=1}^{n} X_t \xrightarrow{Pr} 0 \quad \text{für} \quad n \to \infty.$$

Ohne jede Annahme an den stochastischen Prozess folgt aus fast sicherer Konvergenz die Konvergenz in Wahrscheinlichkeit:

$$X_n \xrightarrow{f.s.} X \quad \Longrightarrow \quad X_n \xrightarrow{Pr} X.$$

Trotzdem sind schwache Gesetze ebenfalls interessant, denn da das Konvergenz-Konzept dieser Gesetze schwächer ist, sind im Allgemeinen auch die Voraussetzungen schwächer. Im Gegensatz zu den starken und schwachen Gesetzen für unabhängige stochastische Prozesse mit hinreichenden und notwendigen Bedingungen gibt es für die Gesetze der großen Zahlen bei unterschiedlichen Abhängigkeitsstrukturen im Allgemeinen nur hinreichende Bedingungen. Der Nachweis der Notwendigkeit der Bedingungen scheitert meist an grundlegend unterschiedlichen Konzepten bei Ungleichungen für stochastische Prozesse, wie beispielsweise beim Borel-Cantelli Lemma (vgl. Hall und Heyde, 1980).

Die folgenden unterschiedlichen Erweiterungen zum schwachen Gesetz der großen Zahlen, also der Konvergenz des Beobachtungsmittels in Wahrscheinlichkeit, für Martingale $\{S_n, \mathcal{A}_n\}_1^\infty$ bzw. Martingal-Differenzen $\{X_t, \mathcal{A}_t\}_1^\infty$ mit $X_t = S_t - S_{t-1}$ sind an Hall und Heyde (1980) angelehnt.

3.18 Satz
Sei $\{S_n, \mathcal{A}_n\}_1^\infty$ ein Martingal mit $S_n = \sum_{t=1}^n X_t$ und $\mathrm{E}(X_t) = 0$, $t = 1, 2 \ldots$, und $\{b_n\}_1^\infty$ eine Folge positiver Konstanten mit $b_n \nearrow \infty$ für $n \to \infty$. Dann gilt

$$\frac{1}{b_n} S_n \xrightarrow{Pr} 0 \quad \text{für} \quad n \to \infty,$$

falls

(i) $\sum_{t=1}^{n} P(|X_t| > b_n) \underset{n\to\infty}{\longrightarrow} 0,$

(ii) $\frac{1}{b_n} \sum_{t=1}^{n} \mathrm{E}(X_t \, \mathbf{1}_{\{|X_t| \le b_n\}} | \mathcal{A}_{t-1}) \underset{n\to\infty}{\xrightarrow{Pr}} 0$ und

(iii) $\frac{1}{b_n^2} \sum_{t=1}^{n} \left[\mathrm{E}(X_t^2 \, \mathbf{1}_{\{|X_t| \le b_n\}}) - \mathrm{E}\left(\left[\mathrm{E}(X_t \, \mathbf{1}_{\{|X_t| \le b_n\}} | \mathcal{A}_{t-1}) \right]^2 \right) \right] \underset{n\to\infty}{\longrightarrow} 0$

gilt, wobei mit $\mathbf{1}$ die Indikatorfunktion bezeichnet sei.

Aus dieser allgemeinen Form des schwachen Gesetzes der großen Zahlen mit bedingten Erwartungswerten in den Annahmen (ii) und (iii) von Satz 3.18, folgt sofort die

bekannte Version von Loève (1977, S. 290) mit $b_n = n$. In dieser Form sind die Bedingungen nicht nur hinreichend, sondern auch notwendig.

3.19 Satz *Loève (1977)*
Sei $\{S_n, \mathcal{A}_n\}_1^\infty$ ein Martingal mit $S_n = \sum_{t=1}^n X_t$ und $\mathrm{E}(X_t) = 0$, $t = 1, 2 \ldots$. Dann gilt

$$\frac{1}{n} S_n \xrightarrow{Pr} 0 \quad \text{für} \quad n \to \infty$$

dann und nur dann, wenn

(i) $\sum_{t=1}^{n} P(|X_t| > n) \underset{n\to\infty}{\longrightarrow} 0,$

(ii) $\frac{1}{n} \sum_{t=1}^{n} \mathrm{E}(X_t \mathbf{1}_{\{|X_t| \le n\}}) \underset{n\to\infty}{\xrightarrow{Pr}} 0$ und

(iii) $\frac{1}{n^2} \sum_{t=1}^{n} \left[\mathrm{E}(X_t^2 \mathbf{1}_{\{|X_t| \le n\}}) - \left[\mathrm{E}(X_t \mathbf{1}_{\{|X_t| \le n\}}) \right]^2 \right] \underset{n\to\infty}{\longrightarrow} 0$

gilt, wobei mit $\mathbf{1}$ die Indikatorfunktion bezeichnet sei.

Die einfachste Version eines schwachen Gesetzes der großen Zahlen für Martingale $\{S_n, \mathcal{A}_n\}_1^\infty$ ergibt sich bei Annahme beschränkter zweiter Momente für die zugehörigen Martingal-Differenzen $\{X_t, \mathcal{A}_t\}_1^\infty$. In diesem Fall gilt die Verallgemeinerung der Tschebyschev-Ungleichung für Martingale. Mit der Markov-Ungleichung gilt für alle $\varepsilon > 0$

$$P(|S_n| \ge \varepsilon) \le \frac{1}{\varepsilon^2} \mathrm{E}(S_n^2),$$

und außerdem

$$\mathrm{E}(S_n^2) = \sum_{t=1}^{n} \mathrm{E}(X_t^2) + 2 \sum_{t>s} \mathrm{E}(X_t X_s) = \sum_{t=1}^{n} \mathrm{E}(X_t^2),$$

denn für $t > s$ gilt wegen der Martingal-Eigenschaften Korrolar 3.17. Insgesamt folgt also für quadratisch integrierbare Martingale

$$P(|S_n| \ge n\varepsilon) \le \frac{1}{n^2 \varepsilon^2} \sum_{t=1}^{n} \mathrm{E}(X_t^2),$$

und damit auch das schwache Gesetz der großen Zahlen.

3.20 Satz
Sei $\{S_n, \mathcal{A}_n\}_1^\infty$ ein Martingal mit $S_n = \sum_{t=1}^n X_t$ und $\mathrm{E}(X_t^2) < \infty$, $t = 1, 2 \ldots$. Dann gilt

$$\frac{1}{n} S_n \xrightarrow{Pr} 0 \quad \text{für} \quad n \to \infty,$$

falls $\frac{1}{n^2} \sum_{t=1}^{n} \mathrm{E}(X_t^2) \to 0$ für $n \to \infty$.

Im Unterschied zu den schwachen Gesetzen der großen Zahlen benötigen die starken Gesetze deutlich härtere Annahmen an die stochastischen Prozesse. Diese Annahmen ergeben sich im Allgemeinen aus der Tatsache, dass zum Nachweis der fast sicheren Konvergenz das Extremwertverhalten stochastischer Prozesse betrachtet werden muss. Dazu sind Ungleichungen zur Abschätzung oberer Wahrscheinlichkeitsgrenzen zum Verhalten der Prozesse von essentieller Bedeutung. Diese Aussagen können dann zum Nachweis der starken Gesetze der großen Zahlen genutzt werden.
Allerdings sind die Resultate auch in anderer Hinsicht interessant. In Anlehnung an Prozesse mit unabhängigen Koordinaten sind neben den Aussagen über das Extremwertverhalten auch Aussagen über die Geschwindigkeit der Konvergenz bei Martingalen wünschenswert. Das wichtigste Resultat zum Extremwertverhalten ist eine Anpassung der Kolmogorov-Ungleichung an das Martingal-Konzept (vgl. Vovk, 1997).

3.21 Satz *Kolmogorov-Ungleichung*
Sei $\{S_n, \mathcal{A}_n\}_1^\infty$ ein Martingal. Dann gilt für beliebiges $r \geq 1$ folgende Aussage:

$$P\left(\max_{1 \leq k \leq n} |S_k| > \varepsilon\right) \leq \frac{\mathrm{E}(|S_n|^r)}{\varepsilon^r}.$$

Als Folgerung aus dem obigen Satz ergibt sich die Aussage von Doob (1953) zur Abschätzung des erwarteten Extremwerts eines Martingals.

3.22 Satz *Doob-Ungleichung*
Sei $\{S_n, \mathcal{A}_n\}_1^\infty$ ein Martingal. Dann gilt

(i) für $r = 1$ und $|S_n| > 0$:

$$\mathrm{E}(|S_n|) \leq \mathrm{E}\left(\max_{1\leq k\leq n} |S_k|\right) \leq \frac{\exp(1)}{\exp(1)-1} + \frac{\exp(1)}{\exp(1)-1}\,\mathrm{E}\left(|S_n| \ln |S_n|\right) ,$$

(ii) für $r > 1$:

$$\mathrm{E}(|S_n|^r) \leq \mathrm{E}\left(\max_{1\leq k\leq n} |S_k|^r\right) \leq \frac{r}{r-1}\,\mathrm{E}(|S_n|^r) ,$$

bzw.

$$\|S_n\|_r \leq \left\|\max_{1\leq k\leq n} |S_k|\right\|_r \leq \frac{r}{r-1}\,\|S_n\|_r .$$

Diese Ungleichungen sind bemerkenswert, da die Anzahl n und das Verhalten der betrachteten Elemente des Martingals keine Auswirkungen auf die obere Schranke haben. Die Grenze hängt nur von dem letzten betrachteten Martingalelement S_n ab. Diese Eigenschaft ergibt sich, da alle Informationen aus der Vergangenheit in S_n enthalten ist.

Weitere wichtige Abschätzungen sind exponentielle Grenzen für die Wahrscheinlichkeit, dass ein stochastischer Prozess konvergiert. Das folgende Ergebnis für Martingale mit beschränkten Martingal-Differenzen entspricht der Hoeffding-Ungleichung (Hoeffding, 1963) für unabhängige Prozesse und ist an das Ergebnis von Azuma (1967) angelehnt.

3.23 Satz *Davidson (1994)*
Sei $\{X_t, \mathcal{A}_t\}_1^\infty$ eine Folge von Martingal-Differenzen mit $|X_t| \leq B_t < \infty$, wobei $\{B_t\}_1^\infty$ eine Folge positiver Konstanten ist. Dann gilt

$$P\left(\left|\sum_{t=1}^{n} X_t\right| > \varepsilon\right) \;\leq\; 2\exp\left\{\frac{-\varepsilon^2}{2\sum_{t=1}^{n} B_t^2}\right\} .$$

Die Aussage dieses Satzes ist, dass die Wahrscheinlichkeiten an den Rändern exponentiell abnehmen bei wachsendem $\varepsilon > 0$, die Wahrscheinlichkeit, dass die Summe über die X_t betragsmäßig groß wird, ist somit klein. Wenn die Koordinaten des Prozesses

gleichmäßig durch eine Konstante beschränkt sind, d. h. $B_t = B < \infty$ für alle t, kann die Schranke vereinfacht werden zu:

$$P\left(\left|\sum_{t=1}^{n} X_t\right| > \varepsilon\right) \leq 2\exp\left\{\frac{-\varepsilon^2}{2nB^2}\right\}.$$

Die obere Schranke für das Mittel eines Martingals entspricht damit:

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n} X_t\right| > \varepsilon\right) = P\left(\left|\sum_{t=1}^{n} X_t\right| > n\varepsilon\right) \leq 2\exp\left\{\frac{-n\varepsilon^2}{2B^2}\right\}.$$

3.24 Korollar
Sei $\{X_t, \mathcal{A}_t\}_1^\infty$ eine Folge von Martingal-Differenzen mit gleichmäßig beschränkten Koordinaten, d. h. $|X_t| \leq B < \infty$ für eine positive Konstante B. Dann gilt

$$P\left(\left|\overline{X}_n\right| > \varepsilon\right) \leq 2\exp\left\{\frac{-n\varepsilon^2}{2B^2}\right\}.$$

Für festes $\varepsilon > 0$ und wachsendes n fällt die Schranke für die Wahrscheinlichkeit, dass das Mittel $\overline{X}_n$ groß ist, exponentiell ab. Dies ist insofern interessant, als dass nicht nur wegen des Gesetzes der großen Zahlen die Konvergenz des Mittels einer Folge von Martingal-Differenzen gesichert ist, sondern dass diese Konvergenz auch schnell genug erfolgt, um auch praktisch gute Ergebnisse erwarten zu können. Eine weitere Verbesserung im Sinne engerer Schranken wird durch Laib (1999) angegeben.

In theoretischer Hinsicht sind vor allem die Kolmogorov- und die Doob-Ungleichung wichtig, da sie zentrale Bestandteile für den Nachweis starker Gesetze der großen Zahlen sind. Für Martingale können bemerkenswert starke Konvergenzresultate erzielt werden. Für Folgen von Martingal-Differenzen sind keine zusätzlichen Annahmen an die Abhängigkeitsstruktur nötig, und die Annahmen bezüglich der Momente des Prozesses sind nur geringfügig strenger als im unabhängigen Fall. Dies liegt zum größten Teil daran, dass bei den Konvergenzaussagen der Unterschied zwischen unkorrelierten Prozessen und einer Folge von Martingal-Differenzen relativ gering ist, wie bereits in Satz 3.16 und dem darauf folgenden Korollar 3.17 gezeigt wurde. Die Ursache dafür liegt

darin begründet, dass eine Vorhersage im Sinne von Ein-Schritt-Prognosen bei Martingalen nicht möglich ist.
In diesem Abschnitt sollen die beiden wichtigsten starken Gesetze der großen Zahlen vorgestellt werden, für weitergehende Resultate sei auf Pötscher und Prucha (1991a,b) und auf Davidson und de Jong (1997) verwiesen. Die vorgestellten Gesetze sind aus Stout (1974) sowie Hall und Heyde (1980) entnommen, gehen aber bereits auf Doob (1953) zurück.

3.25 Satz *Doob (1953)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ eine Folge von Martingal-Differenzen mit der zugehörigen Folge von Varianzen $\{\sigma_t^2\}_{-\infty}^{\infty}$ mit $\sigma_t^2 = \mathrm{E}(X_t^2) < \infty$ für alle t. Sei weiter $\{a_t\}_{-\infty}^{\infty}$ eine Folge positiver Konstanten mit $a_t \nearrow \infty$ für $t \to \infty$, dann gilt:

$$\frac{1}{a_n} \sum_{t=1}^{n} X_t \xrightarrow{f.s.} 0,$$

falls $\sum_{t=1}^{\infty} \frac{\sigma_t^2}{a_t^2} < \infty$.

Im Vergleich zum schwachen Gesetz der großen Zahlen in Satz 3.20 wird hier eine stärkere Bedingung bezüglich der Varianzen im stochastischen Prozess verlangt. Beim starken Gesetz muss die Summe der Varianzen langsamer wachsen als die Summe einer Folge $\{a_t\}$, für das schwache Gesetz reicht es aus, dass die Summe der Varianzen langsamer wächst als n. Bei Wahl von $a_t = n$ gilt die Aussage von Satz 3.25 ebenfalls, es folgt das klassische Resultat

$$\frac{1}{n} \sum_{t=1}^{n} X_t \xrightarrow{f.s.} 0$$

unter der Bedingung $\lim_{n\to\infty} \frac{1}{n^2} \sum_{t=1}^{n} \mathrm{E}(X_t^2) < \infty$.

Die Beschränkung auf quadratisch integrierbare Martingale im vorherigen Satz kann zu Gunsten der Klasse der Martingale fallen gelassen werden, die

$$\sum_{t=1}^{\infty} \frac{1}{a_t^r} \mathrm{E}(|X_t|^r) < \infty \quad , \; 1 \leq r \leq 2,$$

mit $\{a_t\} \nearrow \infty$ erfüllen. Es ist wichtig zu betonen, dass diese Forderung für $r < 2$ keine schwächere Forderung als für $r = 2$ ist. Aus der Bedingung für $r = 2$ kann nicht auf die Bedingung für ein $r < 2$ geschlossen werden.

3.26 Satz *Doob (1953)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ eine Folge von Martingal-Differenzen, die $\sum_{t=1}^{\infty} \frac{1}{a_t^r} \mathrm{E}(|X_t|^r) < \infty$ erfüllt, $1 \leq r \leq 2$, mit $\{a_t\}_{-\infty}^{\infty} \nearrow \infty$. Dann gilt:

$$\frac{1}{a_n} \sum_{t=1}^{n} X_t \xrightarrow{f.s.} 0.$$

Mit dieser Erweiterung des starken Gesetzes der großen Zahlen kann mit der Wahl von r entschieden werden, ob für den stochastischen Prozess entweder höhere absolute Momente existieren sollen oder der Prozess sehr gedämpft sein soll, also nur schwache Amplituden haben soll, um die Summierbarkeitsbedingung zu erhalten. Für $r = 2$ fallen die Aussagen der Sätze 3.25 und 3.26 zusammen, für $r = 1$ lautet die Voraussetzung des Satzes

$$\sum_{t=1}^{\infty} \frac{1}{a_t} \mathrm{E}(|X_t|) < \infty,$$

und es gilt mit dem Kronecker-Lemma

$$\frac{1}{a_n} \sum_{t=1}^{n} \frac{1}{a_t} \mathrm{E}(|X_t|) \longrightarrow 0 \quad \text{für} \quad n \to \infty.$$

Also muss in diesem Fall für $a_n = n$ oder a_n ungefähr gleich n der Prozess insgesamt Werte nahe bei Null besitzen, d. h. er hat ein schwaches Schwingungsverhalten. Aus diesem Satz lässt sich eine Aussage über die Konvergenz des Beobachtungsmittels ableiten. Für $a_t = n$, für alle t, folgt

$$\frac{1}{n} \sum_{t=1}^{n} X_t \xrightarrow{f.s.} 0$$

unter der Bedingung, dass $\lim_{n\to\infty} \frac{1}{n^{r-1}} \sum_{t=1}^{n} \mathrm{E}(|X_t|^r) < \infty$ gilt, $1 \leq r \leq 2$.

3.4 Mixingale

Martingal-Differenzen sind auf Grund ihrer Abhängigkeitsstrukturen recht spezielle stochastische Prozesse. Gerade die Eigenschaft, dass Ein-Schritt-Prognosen nicht vorhersagbar sind, kann in der Anwendung, z. B. bei Zeitreihen, meist nicht beobachtet werden. In diesem Abschnitt soll deshalb ein generelleres Konzept vorgestellt werden, das in gewisser Weise nur asymptotische Gedächtnislosigkeit fordert. Dieses von McLeish (1975) eingeführte Mixingal-Konzept ist von praktischer Bedeutung, denn Mixingale sind Martingal-Differenzen hinreichend ähnlich. Sie haben bei gleichen Konvergenzeigenschaften Martingalen gegenüber den Vorteil, dass die Voraussetzungen nicht so streng sind und relativ leicht zu überprüfen sind.
Die Abhängigkeitsstruktur von Mixingalen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ wird über bedingte Erwartungswerte des stochastischen Prozesses bezüglich Sub-σ-Algebren aus einer nichtfallenden Folge

$$\mathcal{A}_1 \subseteq \mathcal{A}_2 \subseteq \ldots \subseteq \mathcal{A}_m \subseteq \ldots \subseteq \mathcal{A}$$

konstruiert. Das genutzte Maß für das hier vorgestellte Konzept ist die L^r-Norm

$$\|X\|_r = (\mathrm{E}(|X|^r))^{1/r}, \quad r \geq 1,$$

wobei in den meisten Fällen die L^1- bzw. die L^2-Norm genutzt wird.

3.27 Definition *L^r-Mixingal*
Auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ sei $\{X_t\}_{t=-\infty}^{\infty}$ eine Folge von Zufallsvariablen mit $\mathrm{E}(|X_t|) < \infty$ für alle t. Sei weiter $\{\mathcal{A}_t\}_{t=-\infty}^{\infty}$ eine nichtfallende Folge von Sub-σ-Algebren in $\mathcal{A}$. Die Folge von Paaren $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ist ein *L^r-Mixingal* für $r \geq 1$, falls nicht negative Folgen von Konstanten $\{c_t, -\infty \leq t \leq \infty\}$ und $\{\psi_m, m \geq 0\}$ mit $\psi_m \searrow 0$ für $m \to \infty$ existieren, so dass für alle $t \geq 1$ und $m \geq 0$ gilt:

(a) $\|\mathrm{E}(X_t|\mathcal{A}_{t-m})\|_r \leq c_t\psi_m$,

(b) $\|X_t - \mathrm{E}(X_t|\mathcal{A}_{t+m})\|_r \leq c_t\psi_{m+1}$.

Wird aus dem Zusammenhang deutlich, bezüglich welcher Norm ein L^r-Mixingal definiert ist, so wird dieses im Folgenden auch einfach als *Mixingal* bezeichnet. Bei diesem

Konzept wird für die Koordinaten X_t eines stochastischen Prozesses $\{X_t\}_{-\infty}^{\infty}$ und die Folge von Sub-σ-Algebren $\{\mathcal{A}_t\}_{-\infty}^{\infty}$ gefordert, dass der bedingte Erwartungswert einer Koordinate bezüglich der Vergangenheit, repräsentiert durch die Sub-σ-Algebren, mit steigendem Lag verschwindet. Gleichzeitig soll sichergestellt sein, dass mit steigendem Informationsgehalt, also für steigenden Index der Sub-σ-Algebren, dieser bedingte Erwartungswert asymptotisch erwartungstreu für die zugehörige Koordinate ist.

Mixingale bilden ein Konzept für schwache Abhängigkeitsstrukturen, das aber stark genug ist zum Nachweis vieler Konvergenzeigenschaften, wie beispielsweise von Gesetzen der großen Zahlen, und für Abschätzungen der Konvergenzgeschwindigkeiten. Jedoch kann im Allgemeinen nicht angenommen werden, dass die Mixingal-Eigenschaften unter Transformationen des stochastischen Prozesses erhalten bleiben (vgl. de Jong, 1998). Hall und Heyde (1980) nennen Mixingale wegen der abgeschwächten Forderung an die Abhängigkeitsstruktur auch asymptotische Martingale, obwohl dies in gewissem Sinne irreführend ist, denn Mixingale sind eher den Differenzen von Martingalen ähnlich. Eine Martingal-Differenz ist ein Mixingal mit $\psi_m = 0$ für alle $m > 0$. Das Gegenstück zu einem Martingal sind die aufsummierten Elemente eines Mixingals.

3.28 Beispiel *Martingal*
Sei $\{S_n, \mathcal{A}_n\}_1^{\infty}$ mit $S_n = \sum_{t=1}^{n} X_t$ und $\mathrm{E}(S_n^2) < \infty$ ein Martingal. Nach Definition 3.27 ist

$$\mathrm{E}(X_t|\mathcal{A}_{t-1}) = 0 \quad \text{fast sicher.}$$

Zusätzlich gilt

$$\mathrm{E}(X_t|\mathcal{A}_{t+m}) = X_t \quad \text{fast sicher, für alle } m \geq 1.$$

Damit ist $\{X_t, \mathcal{A}_t\}_1^{\infty}$ ein L^2-Mixingal mit $c_t = \|X_t\|_2$, sowie $\psi_0 = 1$ und $\psi_m = 0$ für alle $m \geq 1$. Mit der gleichen Begründung ist die Folge $\{X_t, \mathcal{A}_t\}_1^{\infty}$ auch ein beliebiges L^r-Mixingal, $r \geq 1$, solange S_n L^r-integrierbar für alle n und damit X_t L^r-integrierbar für alle t ist, also $\mathrm{E}(|X_t|^r) < \infty$ gilt.

Beachtenswert ist, dass ein Mixingal allgemein nicht als adaptierter Prozess definiert ist, denn es wird nicht vorausgesetzt, dass die X_t $\mathcal{A}_t$-meßbar sind, für alle t. Gilt

andererseits diese Annahme, so ist die Bedingung (b) aus Definition 3.27 trivial, denn

$$\mathrm{E}(X_t|\mathcal{A}_{t+m}) = \mathrm{E}(X_t|\mathcal{A}_t) = X_t$$

gilt dann für alle t und alle $m \geq 0$.
Die Bedingungen innerhalb des Mixingal-Konzepts beinhalten die Idee, dass die Folge $\{\mathcal{A}_m\}_1^\infty$ mit steigendem Index m sukzessive mehr Information für X_t enthält. Dabei ist in der Vergangenheit ($m \to \infty$) nichts über die Koordinate X_t bekannt, wie aus Bedingung (a) in Definition 3.27 ersichtlich wird, in der Zukunft ($m \to \infty$) wird dagegen wegen Bedingung (b) irgendwann alles über X_t bekannt sein. Die Folge der Konstanten $\{\psi_m\}$, die *L^r-Mixingal-Koeffizienten*, sind Zahlen, die die zeitliche Abhängigkeit im stochastischen Prozess $\{X_t\}$ angeben. Je schneller die Folge $\{\psi_m\}$ mit steigendem Index gegen Null strebt, desto kürzer ist das Gedächtnis des stochastischen Prozesses. Die Folge $\{\psi_m\}$ ist von der *Ordnung* $-\xi_0$, falls $\psi_m = O(m^{-\xi})$ für $\xi > \xi_0$.
Die Folge der Konstanten $\{c_t\}$ besteht aus einfachen Skalierungfaktoren, um die Wahl der ψ_m unabhängig von der relativen Größe der Zufallsvariablen X_t zu machen. Meist reicht es aus, die c_t als ein Vielfaches von $\|X_t\|_r$ zu wählen. Im Allgemeinen gilt für alle L^r-Mixingale mit $r \geq 1$:

$$\|X_t\|_r \leq \|\mathrm{E}(X_t|\mathcal{A}_t)\|_r + \|X_t - \mathrm{E}(X_t|\mathcal{A}_t)\|_r \leq (\psi_0 + \psi_1)c_t.$$

Mit dieser Aussage folgt auch sofort, dass ohne Einschränkung der Allgemeinheit zentrierte Prozesse betrachtet werden können, da der Erwartungswert eines Mixingals $\mathrm{E}(|X_t|) < \infty$ für alle t existiert.

Mixingale sind somit eine weit größere Klasse stochastischer Prozesse als Martingale. Viele Prozesse, für die Konvergenzeigenschaften bekannt sind, können als Mixingal beschrieben werden. Von besonderer Wichtigkeit ist dabei, dass für den Nachweis, dass ein stochastischer Prozess ein Mixingal ist, Stationarität des Prozesses nicht zwingend benötigt wird (McLeish, 1975). Allerdings ist die Mixingal-Eigenschaft eines Prozesses meist an Bedingungen bezüglich der Abhängigkeitsstruktur gebunden. Diese hängen im Allgemeinen von Annahmen an die bedingte Erwartung ab.

3.29 Beispiel *Lineare Prozesse*

(i) Sei $\{e_t\}_{-\infty}^{\infty}$ eine Folge unabhängiger Zufallsvariablen, die L^r-beschränkt sind, $r \geq 2$, d. h. es gilt $\sup_{t\geq 1} \|e_t\|_r = B < \infty$. Sei weiter $\mathrm{E}(e_t|\mathcal{A}_{t-1}) = 0$ mit $\mathcal{A}_t = \sigma\{e_j : j \leq t\}$, so dass die Martingal-Differenzen $\{e_t, \mathcal{A}_t\}$ als Innovationen mit einer Folge von Konstanten $\{a_k\}_{-\infty}^{\infty}$, $\sum_{k=-\infty}^{\infty} |a_k| < \infty$, den *heterogenen linearen Prozess*

$$X_t = \sum_{k=-\infty}^{\infty} a_k e_{t-k}$$

erzeugen.

Ein solcher Prozess mit zugehörigen σ-Algebren $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ist ein L^r-Mixingal, denn es gilt:

$$\|\mathrm{E}(X_t|\mathcal{A}_{t-m})\|_r = \left\| \sum_{k=m}^{\infty} a_k e_{t-k} \right\|_r \leq \sum_{k=m}^{\infty} \|a_k e_{t-k}\|_r \leq \sum_{k=m}^{\infty} |a_k| B$$

und

$$\|X_t - \mathrm{E}(X_t|\mathcal{A}_{t+m})\|_r = \left\| \sum_{k=-m}^{-\infty} a_k e_{t-k} \right\|_r \leq \sum_{k=m}^{\infty} |a_{-k}| B.$$

Da hier $\mathcal{A}_{t-s} \subseteq \mathcal{A}_{t-1} \subseteq \mathcal{A}$ für alle $s \geq 1$ gilt und die e_t $\mathcal{A}$-messbar sind, folgen beide Ungleichungen aus der Eigenschaft bedingter Erwartungswerte und der Bedingung an den bedingten Erwartungswert für e_t:

$$\mathrm{E}(e_t|\mathcal{A}_{t-s}) = \mathrm{E}\left(\mathrm{E}(e_t|\mathcal{A}_{t-1})|\mathcal{A}_{t-s}\right) = 0 \quad \text{fast sicher,}$$

für alle $s \geq 1$ und $\mathrm{E}(e_t|\mathcal{A}_{t-s}) = e_t$ für alle $s \leq 0$.

Als Ergebnis aus den Ungleichungen können als Mixingal-Koeffizienten

$$\psi_m = \sum_{k=m}^{\infty} (|a_k| + |a_{-k}|)$$

und als Skalierungsfaktoren

$$c_t = \sup_{i\geq 1} \|e_i\|_r$$

gewählt werden (Hansen, 1991).

(ii) Bei Annahme schwacher Stationarität wird der Nachweis, dass ein linearer Prozess ein L^2-Mixingal ist, deutlich einfacher. Sei $\{e_t\}_{-\infty}^{\infty}$ eine Folge unabhängiger

Zufallsvariablen mit $\mathrm{E}(e_t) = 0$ und beschränkter Varianz $\mathrm{Var}(e_t) < \infty$ für alle t. Mit einer Folge von Konstanten $\{a_k\}_{-\infty}^{\infty}$ mit $\sum_{k=-\infty}^{\infty} |a_k| < \infty$ ist

$$X_t = \sum_{k=-\infty}^{\infty} a_k e_{t-k}$$

ein *homogener linearer Prozess*. Die Folge von Paaren $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ist ein L^r-Mixingal für beliebiges $r \geq 1$, wobei $\mathcal{A}_t = \sigma(\{e_i\}_{-\infty}^{t})$ die durch $\{e_t, e_{t-1}, \ldots\}$ induzierte σ-Algebra ist. Dabei kann hier speziell für die L^2-Norm wegen der schwachen Stationarität

$$c_t = \sqrt{\mathrm{Var}(e_t)} \quad \text{und} \quad \psi_m = \left(\sum_{k=m}^{\infty} a_k^2 + a_{-k}^2 \right)^{1/2}$$

gewählt werden. Dies ist eine schwächere Bedingung als in (i), denn es gilt $(\sum_{k=m}^{\infty} a_k^2 + a_{-k}^2)^{1/2} \leq \sum_{k=m}^{\infty} (|a_k| + |a_{-k}|)$ (Hall und Heyde, 1980).

Die Abhängigkeitseigenschaften heterogener linearer Prozesse sind stärker, deshalb konvergieren die Mixingal-Koeffizienten langsamer gegen Null als die des homogenen linearen Prozesses, wo immerhin schwache Stationarität angenommen wird. Beide Beispiele verdeutlichen aber die Wichtigkeit der Mixingale für praktische Anwendungen. Die schwachen Voraussetzungen erlauben es, eine Vielzahl von linearen Prozessen als Mixingal aufzufassen, so insbesondere infinite MA-Prozesse und damit auch direkt alle schwach stationären AR-, MA- und ARMA-Prozesse (Andrews, 1988).

Eine äußerst wertvolle Eigenschaft von Mixingalen ist ihre asymptotische Ähnlichkeit zu Martingal-Differenzen. Die Summe von Mixingalen unterscheidet sich bis auf einen Rest von einem Martingal. Unter relativ schwachen Annahmen an die Abhängigkeitsstruktur des Prozesses ist dieser Rest asymptotisch vernachlässigbar. Für jede integrierbare Zufallsvariable X_t und jedes $m \geq 0$ gilt:

$$X_t = \sum_{k=-m}^{m} (\mathrm{E}(X_t|\mathcal{A}_{t+k}) - \mathrm{E}(X_t|\mathcal{A}_{t+k-1})) + \mathrm{E}(X_t|\mathcal{A}_{t-m-1}) + (X_t - \mathrm{E}(X_t|\mathcal{A}_{t+m})),$$

denn:

$$\sum_{k=-m}^{m} (\mathrm{E}(X_t|\mathcal{A}_{t+k}) - \mathrm{E}(X_t|\mathcal{A}_{t+k-1})) + \mathrm{E}(X_t|\mathcal{A}_{t-m-1}) + (X_t - \mathrm{E}(X_t|\mathcal{A}_{t+m}))$$

$$\begin{aligned}
= \quad & \mathrm{E}(X_t|\mathcal{A}_t) - \mathrm{E}(X_t|\mathcal{A}_{t-1}) && (k=0)\\
+ \quad & \mathrm{E}(X_t|\mathcal{A}_{t+1}) - \mathrm{E}(X_t|\mathcal{A}_t) + \mathrm{E}(X_t|\mathcal{A}_{t-1}) - \mathrm{E}(X_t|\mathcal{A}_{t-2}) && (k=\pm 1)\\
+ \quad & \mathrm{E}(X_t|\mathcal{A}_{t+2}) - \mathrm{E}(X_t|\mathcal{A}_{t+1}) + \mathrm{E}(X_t|\mathcal{A}_{t-2}) - \mathrm{E}(X_t|\mathcal{A}_{t-3}) && (k=\pm 2)\\
& \vdots\\
+ \quad & \mathrm{E}(X_t|\mathcal{A}_{t+m}) - \mathrm{E}(X_t|\mathcal{A}_{t+m-1}) + \mathrm{E}(X_t|\mathcal{A}_{t-m}) - \mathrm{E}(X_t|\mathcal{A}_{t-m-1}) && (k=\pm m)\\
+ \quad & \mathrm{E}(X_t|\mathcal{A}_{t-m-1}) + X_t - \mathrm{E}(X_t|\mathcal{A}_{t+m})\\
= \quad & X_t
\end{aligned}$$

Außerdem ist für alle k die Folge

$$\{\mathrm{E}(X_t|\mathcal{A}_{t+k}) - \mathrm{E}(X_t|\mathcal{A}_{t+k-1}), \mathcal{A}_{t+k}\}_{t=1}^{\infty}$$

eine Martingal-Differenz, weil mit dem Gesetz der iterierten Erwartung

$$\mathrm{E}(\mathrm{E}(X_t|\mathcal{A}_{t+k}) - \mathrm{E}(X_t|\mathcal{A}_{t+k-1})|\mathcal{A}_{t+k-1}) = 0$$

gilt. Mit derselben Begründung ist

$$\{\mathrm{E}(X_t|\mathcal{A}_{t+m}), \mathcal{A}_{t+m}\}_{-\infty}^{\infty}$$

ein Martingal, denn $\mathrm{E}(\mathrm{E}(X_t|\mathcal{A}_{t+m})|\mathcal{A}_{t+m-1}) = \mathrm{E}(X_t|\mathcal{A}_{t+m-1})$. Da

$$\sup_m \mathrm{E}(|\mathrm{E}(X_t|\mathcal{A}_{t+m})|) \leq \mathrm{E}(|X_t|) < \infty$$

gilt, konvergiert die Folge sowohl für $m \to \infty$ als auch $m \to -\infty$ fast sicher. Ist die Folge $\{X_t, \mathcal{A}_t\}$ ein Mixingal, gilt wegen Definition 3.27

$$\|\mathrm{E}(X_t|\mathcal{A}_{t-m})\|_r \to 0 \quad \text{und} \quad \|X_t - \mathrm{E}(X_t|\mathcal{A}_{t+m})\|_r \to 0.$$

Insgesamt folgt damit für die Grenzwerte

$$\mathrm{E}(X_t|\mathcal{A}_{t-m}) \to 0 \quad \text{und} \quad \mathrm{E}(X_t|\mathcal{A}_{t+m}) \to X_t \quad (m \to \infty).$$

Die Rest-Terme können also auf Grund der Mixingal-Eigenschaft vernachlässigt werden, falls m nur groß genug gewählt wird. Zusammenfassend ergibt sich folgender Satz.

3.30 Satz *McLeish (1975)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal. Dann ist $\{\Psi_m\}_0^{\infty}$ mit

$$\Psi_m = \mathrm{E}(X_t|\mathcal{A}_{t-m-1}) + (X_t - \mathrm{E}(X_t|\mathcal{A}_{t+m}))$$

eine Nullfolge (für $m \to \infty$) und für alle m gilt

$$X_t \;=\; \sum_{k=-m}^{m} (\mathrm{E}(X_t|\mathcal{A}_{t+k}) - \mathrm{E}(X_t|\mathcal{A}_{t+k-1})) + \Psi_m \qquad \text{fast sicher,}$$

und damit

$$X_t \;=\; \sum_{k=-\infty}^{\infty} (\mathrm{E}(X_t|\mathcal{A}_{t+k}) - \mathrm{E}(X_t|\mathcal{A}_{t+k-1})) \qquad \text{fast sicher.}$$

Dabei sind $\{\mathrm{E}(X_t|\mathcal{A}_{t+k}) - \mathrm{E}(X_t|\mathcal{A}_{t+k-1}), \mathcal{A}_{t+k}\}_{t=1}^{\infty}$ Folgen von Martingal-Differenzen für alle k.

Der Vorteil dieser Repräsentation als Teleskop-Summe plus einem Rest-Term ist die vereinfachte Übertragung der Konvergenzeigenschaften von Martingalen auf die Summe von Mixingal-Elementen. Die Schnelligkeit der Konvergenz der Folge Ψ_m hängt über die Mixingal-Eigenschaft von den Annahmen an die Abhängigkeitsstruktur des stochastischen Prozesses ab.
Insbesondere gibt es Mixingale, für die ein m_0 existiert, so daß $\Psi_m = 0$ für alle $m \geq m_0$. In diesem Fall gilt:

$$X_t \;=\; \sum_{k=-m_0}^{m_0} (\mathrm{E}(X_t|\mathcal{A}_{t+k}) - \mathrm{E}(X_t|\mathcal{A}_{t+k-1})) \qquad \text{fast sicher.}$$

Bei Stationarität des stochastischen Prozesses $\{X_t\}_{-\infty}^{\infty}$ können die Skalierungskonstanten in der Folge $\{c_t\}_{-\infty}^{\infty}$ ohne Beschränkung der Allgemeinheit gleich 1 gesetzt werden. Damit kann ein Mixingal zerlegt werden in eine Martingal-Differenz und einen Rest-Term, dessen Verhalten durch einschränkende Voraussetzungen an die Abhängigkeitsstruktur, also an das Mixingal $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$, kontrollierbar ist. In dieser Arbeit wird nur eine Reihenentwicklung für stationäre L^1-Mixingale angegeben, eine Ausweitung auf L^r-Mixingale, $r \geq 1$, ist unter ähnlichen Bedingungen ebenfalls möglich (vgl. Davidson, 1994).

3.31 Satz *Hall und Heyde (1980)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein stationäres L^1-Mixingal von der Ordnung $-\xi_0 = -1$. Dann existiert eine Zerlegung

$$X_t = W_t + Z_t - Z_{t+1},$$

wobei $\mathrm{E}(|Z_t|) < \infty$ für alle t und $\{W_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ eine Folge stationärer Martingal-Differenzen ist. Dabei ist

$$W_t = \lim_{m\to\infty} \left\{ \sum_{s=-m}^{m} (\mathrm{E}(X_{t+s}|\mathcal{A}_t) - \mathrm{E}(X_{t+s}|\mathcal{A}_{t-1})) + \mathrm{E}(X_{t+m+1}|\mathcal{A}_t) \right.$$
$$\left. + \quad X_{t-m-1} - \mathrm{E}(X_{t-m-1}|\mathcal{A}_{t-1}) \right\}$$

und

$$Z_t = \lim_{m\to\infty} \left\{ \sum_{s=0}^{m} (\mathrm{E}(X_{t+s}|\mathcal{A}_{t-1}) - X_{t-s-1} + \mathrm{E}(X_{t-s-1}|\mathcal{A}_{t-1})) \right\}$$

Die Darstellung eines Mixingals als Teleskop-Reihenentwicklung ist für Nachweise zu Resultaten in der Grenzwerttheorie von zentraler Bedeutung, was im Folgenden genutzt wird. Zuerst sei das schwache Gesetz der großen Zahlen in der Version von de Jong (1995) für L^r-Mixingale mit $1 \le r \le 2$ genannt, das direkt auf das Gesetz von Andrews (1988) für L^1-Mixingale zurückführbar ist.

3.32 Satz *de Jong (1995)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal mit zugehörigen Folgen $\{c_t\}_{-\infty}^{\infty}$ und $\{\psi_m\}_1^{\infty}$, $1 \le r \le 2$, so dass für eine Folge $\{C_n\}_{n=1}^{\infty}$ mit $C_n \ge 1$ und $\dfrac{1}{n^{1/2}}\, C_n \to 0$ für $n \to \infty$ die Voraussetzungen

(i) $\lim\limits_{B\to\infty} \limsup\limits_{n\to\infty} \dfrac{1}{n} \sum\limits_{t=1}^{n} \|X_t \mathbf{1}_{\{|X_t| > BC_n\}}\|_r = 0$ mit Indikatorfunktion $\mathbf{1}$ und

(ii) für alle $K > 0$: $\lim\limits_{n\to\infty} \dfrac{1}{n} \sum\limits_{t=1}^{n} c_t \psi_m = 0$ mit $m = \lceil n^{1/2} C_n^{-1} K \rceil$

gelten. Dann folgt $\|\dfrac{1}{n} \sum\limits_{t=1}^{n} X_t\|_r \longrightarrow 0$ für $n \to \infty$ und damit

$$\left| \frac{1}{n} \sum_{t=1}^{n} X_t \right| \xrightarrow{Pr} 0.$$

Beweis:

Für beliebiges $B > 0$ und beliebiges $m \geq 1$ gilt:

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} X_t &= \frac{1}{n}\sum_{t=1}^{n}\left[X_t - \mathrm{E}\left(X_t|\mathcal{A}_{t+m-1}\right)\right] \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\left[\mathrm{E}\left(X_t\mathbf{1}_{\{|X_t|\leq BC_n\}}|\mathcal{A}_{t+m-1}\right) - \mathrm{E}\left(X_t\mathbf{1}_{\{|X_t|\leq BC_n\}}|\mathcal{A}_{t-m}\right)\right] \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\left[\mathrm{E}\left(X_t\mathbf{1}_{\{|X_t|> BC_n\}}|\mathcal{A}_{t+m-1}\right) - \mathrm{E}\left(X_t\mathbf{1}_{\{|X_t|> BC_n\}}|\mathcal{A}_{t-m}\right)\right] \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left(X_t|\mathcal{A}_{t-m}\right) \\
&= T_1 + T_2 + T_3 + T_4.
\end{aligned}$$

Sei $m = m_n$. Mit Voraussetzung (i) und der bedingten Jensen-Ungleichung kann B so groß gewählt werden, dass für alle $\varepsilon > 0$

$$\begin{aligned}
&\limsup_{n\to\infty} \|T_3\|_r \\
&\leq \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}\left[\mathrm{E}\left(\left\|X_t\mathbf{1}_{\{|X_t|>BC_n\}}\right\|_r|\mathcal{A}_{t+m_n-1}\right) - \mathrm{E}\left(\left\|X_t\mathbf{1}_{\{|X_t|>BC_n\}}\right\|_r|\mathcal{A}_{t-m_n}\right)\right] \\
&\leq \limsup_{n\to\infty} \frac{2}{n}\sum_{t=1}^{n}\left\|X_t\mathbf{1}_{\{|X_t|>BC_n\}}\right\|_r < \varepsilon
\end{aligned}$$

gilt. Mit

$$Y_{jt} = \mathrm{E}\left(X_t\mathbf{1}_{\{|X_t|\leq BC_n\}}|\mathcal{A}_{t+j}\right) - \mathrm{E}\left(X_t\mathbf{1}_{\{|X_t|\leq BC_n\}}|\mathcal{A}_{t+j-1}\right)$$

ist $\{Y_{jt}, \mathcal{A}_{t+j}\}_{j=-\infty}^{\infty}$ für $1 \leq t \leq n$ eine Folge von beschränkten Martingal-Differenzen. Dann gilt:

$$\begin{aligned}
\|T_2\|_2 &= \left\|\frac{1}{n}\sum_{t=1}^{n}\sum_{j=-m_n+1}^{m_n-1} Y_{jt}\right\|_2 \leq \sum_{j=-m_n+1}^{m_n-1}\left\|\frac{1}{n}\sum_{t=1}^{n} Y_{jt}\right\|_2 \\
&\leq 4m_n n^{-\frac{1}{2}} BC_n
\end{aligned}$$

(vgl. Andrews, 1988). Für die Wahl $m_n = \max\left\{1, \left\lceil n^{\frac{1}{2}} C_n^{-1} B^{-1} \frac{\varepsilon}{4}\right\rceil\right\}$ folgt mit der Ljapunov-Ungleichung

$$\begin{aligned}
\limsup_{n\to\infty} \|T_2\|_r &\leq \limsup_{n\to\infty} \|T_2\|_2 \\
&\leq \limsup_{n\to\infty} 4\left\lceil n^{\frac{1}{2}} C_n^{-1} B^{-1} \frac{\varepsilon}{4}\right\rceil n^{-\frac{1}{2}} C_n B \leq \varepsilon,
\end{aligned}$$

für $\varepsilon > 0$ und $r \leq 2$, wegen $\lceil x \rceil \leq x$ und $C_n = o\left(n^{1/2}\right)$.
Weiter gilt mit der Mixingal-Definition für dieselbe Wahl von m_n

$$\begin{aligned}\limsup_{n\to\infty} \|T_1 + T_4\|_r &\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left[\|X_t - \mathrm{E}\left(X_t|\mathcal{A}_{t+m_n-1}\right)\|_r + \|\mathrm{E}\left(X_t|\mathcal{A}_{t-m_n}\right)\|_r\right] \\ &\leq 2 \limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} C_t \psi_{m_n} = 0\end{aligned}$$

wegen Voraussetzung (ii). Da $\varepsilon > 0$ beliebig gewählt werden kann, folgt die Behauptung.

□

Die Voraussetzungen dieses Satzes sind für L^1-Mixingale erfüllt, wenn $\psi_m = o(m^{-1/4})$ und $C_n = n^{1/6}$ gewählt wird. Wird $C_n = 1$ für alle n gewählt, so folgt die Voraussetzung (i) aus Satz 3.32 direkt aus $\mathrm{E}(|X_t|^r) < \infty$. Dieses Ergebnis ist vergleichbar mit dem schwachen Gesetz von Gut (1992), zudem ist die Wahl einer Folge wachsender C_n schon deshalb sinnvoll, da dadurch Trend behaftete stochastische Prozesse $\{X_t\}_{-\infty}^{\infty}$ ebenfalls abgedeckt sind (vgl. Davidson, 1993).
Zudem gibt es eine erweiterte Version des schwachen Gesetzes der großen Zahlen von Chen und White (1996) für L^r-Mixingale in beliebigen Hilberträumen. Diese Version ist für reelle L^r-Mixingale mit $1 \leq r \leq 2$ identisch mit der von de Jong (1995), wichtig ist bei dieser Erweiterung aber vor allem, dass das schwache Gesetz damit allgemein für alle reellen L^r-Mixingale mit $r \geq 1$ gilt.

3.33 Satz *Chen und White (1996)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal, $r \geq 1$. Sei weiter die Folge $\{|X_t|^p\}_{-\infty}^{\infty}$ gleichmäßig integrierbar für ein p mit $1 \leq p \leq \min\{2, r\}$ und gelte

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} c_t < \infty.$$

Dann folgt $\|\frac{1}{n} \sum_{t=1}^{n} X_t\|_p \to 0$ für $n \to \infty$ und damit

$$\left| \frac{1}{n} \sum_{t=1}^{n} X_t \right| \xrightarrow{Pr} 0.$$

Neben diesen Sätzen zum schwachen Gesetz der großen Zahlen ist die Dekomposition des Mixingals $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ wie in Satz 3.30 und vor allem der Grenzwertübergang für die folgenden theoretischen Überlegungen relevant, die schließlich zu einem starken Gesetz der großen Zahlen für Mixingale führen. Zuerst seien zwei Ungleichungen zum Maximum von Summen über ein Mixingal genannt, die im Weiteren auch für die Bildung exponentieller Schranken zur Kontrolle der Konvergenzrate benötigt werden.

3.34 Lemma *Hansen (1991)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal, $r \geq 2$, dann existiert eine Konstante $K_1 < \infty$, so dass

$$\left\| \max_{j \leq n} |\sum_{t=1}^{j} X_t| \right\|_r \leq K_1 \sum_{k=-\infty}^{\infty} \left(\sum_{t=1}^{n} \|\mathrm{E}(X_t|\mathcal{A}_{t-k}) - \mathrm{E}(X_t|\mathcal{A}_{t-k-1})\|_r^2 \right)^{1/2}$$

gilt.

Beweis:
Sei $X_{kt} = \mathrm{E}(X_t|\mathcal{A}_{t-k}) - \mathrm{E}(X_t|\mathcal{A}_{t-k-1})$. Dann ist $\{X_{kt}, \mathcal{A}_{t-k}\}_{-\infty}^{\infty}$ eine Folge von Martingal-Differenzen (Satz 3.30).
Für $r \geq 2$ gelten dann folgende Ungleichungen:

$$\begin{aligned}
\left\| \max_{j \leq n} \left| \sum_{t=1}^{j} X_t \right| \right\|_r &= \left\| \max_{j \leq n} \left| \sum_{t=1}^{j} \sum_{k=-\infty}^{\infty} X_{kt} \right| \right\|_r \\
&\leq \left\| \sum_{k=-\infty}^{\infty} \max_{j \leq n} \left| \sum_{t=1}^{j} X_{kt} \right| \right\|_r \\
&\leq \sum_{k=-\infty}^{\infty} \left\| \max_{j \leq n} \left| \sum_{t=1}^{j} X_{kt} \right| \right\|_r \\
&\leq \sum_{k=-\infty}^{\infty} \frac{r}{r-1} \left\| \sum_{t=1}^{n} X_{kt} \right\|_r \\
&\leq \frac{r}{r-1} \sum_{k=-\infty}^{\infty} \left[C\, \mathrm{E} \left(\sum_{t=1}^{n} X_{kt}^2 \right)^{r/2} \right]^{1/r}, \quad \text{für ein } C < \infty.
\end{aligned}$$

Dabei gilt die Gleichung wegen Satz 3.31. Die erste Ungleichung ist die Dreiecksungleichung, die zweite die Minkowski-Ungleichung (vgl. Mitrinović, 1970, S. 55). Nach

Anwendung der Doob- (Satz 3.22) und der Burkholder-Ungleichung (vgl. Hall und Heyde, 1980, S. 23) als dritte und vierte Ungleichung ergibt die nochmalige Anwendung der Minkowski-Ungleichung:

$$\begin{aligned}\left\| \max_{j\leq n} \left| \sum_{t=1}^{j} X_t \right| \right\|_r &\leq C^{1/r} \frac{r}{r-1} \sum_{k=-\infty}^{\infty} \left(\left(\sum_{t=1}^{n} \left\| X_{kt}^2 \right\|_{r/2} \right)^{r/2} \right)^{1/r} \\ &\leq K_1 \sum_{k=-\infty}^{\infty} \left(\sum_{t=1}^{n} \|X_{kt}\|_r^2 \right)^{1/2}\end{aligned}$$

mit $K_1 = C^{1/r} \frac{r}{r-1} < \infty$.

□

Das folgende Lemma nutzt die Schranken und die Konstanten aus Lemma 3.34 und den zugehörigen Beweis für eine weitere Abschätzung.

3.35 Lemma *Hansen (1991)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal, $r \geq 2$, und gelte $\sum_{m=0}^{\infty} \psi_m < \infty$, dann existiert eine Konstante $K_2 < \infty$, so dass

$$\left\| \max_{j\leq n} |\sum_{t=1}^{j} X_t| \right\|_r \leq K_2 \left(\sum_{t=1}^{n} c_t^2 \right)^{1/2}$$

gilt. Dabei sei $K_2 = 4\, K_1 \sum_{k=0}^{\infty} \psi_k < \infty$ mit der Konstanten $K_1 < \infty$ aus Lemma 3.34.

Beweis:
Mit der Mixingal-Bedingung gilt für

$$X_{kt} = \mathrm{E}\left(X_t|\mathcal{A}_{t-k}\right) - \mathrm{E}\left(X_t|\mathcal{A}_{t-k-1}\right)$$

und $k \geq 0$:

$$\|X_{kt}\|_r \leq \|\mathrm{E}\left(X_t|\mathcal{A}_{t-k}\right)\|_r + \|\mathrm{E}\left(X_t|\mathcal{A}_{t-k-1}\right)\|_r \leq 2c_t\psi_k,$$

sowie für $k < 0$:

$$\begin{aligned}\|X_{kt}\|_r &= \|(X_t - \mathrm{E}\left(X_t|\mathcal{A}_{t-k-1}\right)) - (X_t - \mathrm{E}\left(X_t|\mathcal{A}_{t-k}\right))\|_r \\ &\leq \|X_t - \mathrm{E}\left(X_t|\mathcal{A}_{t-k-1}\right)\|_r + \|X_t - \mathrm{E}\left(X_t|\mathcal{A}_{t-k}\right)\|_r \leq 2c_t\psi_{-k}\end{aligned}$$

Daraus folgt:

$$\begin{aligned} \sum_{t=1}^{n} \|X_{kt}\|_r^2 &\leq 4 \sum_{t=1}^{n} c_t^2 \psi_k^2, \quad \text{für } k \geq 0, \\ \text{und} \qquad \sum_{t=1}^{n} \|X_{kt}\|_r^2 &\leq 4 \sum_{t=1}^{n} c_t^2 \psi_{-k}^2, \quad \text{für } k < 0. \end{aligned}$$

Mit Lemma 3.34 folgt:

$$\begin{aligned} \left\| \max_{j \leq n} |S_j| \right\|_r &\leq K_1 \sum_{k=-\infty}^{\infty} \left(\sum_{t=1}^{n} \|X_{kt}\|_r^2 \right)^{1/2} \\ &\leq K_1 \left[\sum_{k=-\infty}^{-1} \left(4 \sum_{t=1}^{n} c_t^2 \psi_{-k}^2 \right)^{1/2} + \sum_{k=0}^{\infty} \left(4 \sum_{t=1}^{n} c_t^2 \psi_k^2 \right)^{1/2} \right] \\ &\leq 2K_1 \left[\sum_{k=1}^{\infty} \left(\sum_{t=1}^{n} c_t^2 \psi_k^2 \right)^{1/2} + \sum_{k=0}^{\infty} \left(\sum_{t=1}^{n} c_t^2 \psi_k^2 \right)^{1/2} \right] \\ &\leq 4K_1 \sum_{k=0}^{\infty} \psi_k \left(\sum_{t=1}^{n} c_t^2 \right)^{1/2} \\ &= K_2 \left(\sum_{t=1}^{n} c_t^2 \right)^{1/2} \end{aligned}$$

wobei $K_2 = 4K_1 \sum_{k=0}^{\infty} \psi_k < \infty$ nach Voraussetzung gilt.

□

Mit Hilfe dieser Lemmata ergibt sich aus der Mixingal-Bedingung und unter Verwendung des Cauchy-Kriteriums, des Kronecker-Lemmas und der Markov-Ungleichung ein starkes Gesetz der großen Zahlen für beliebige L^r-Mixingale, $r \geq 2$, unter relativ schwachen Annahmen, die zudem vergleichbar sind mit den Bedingungen an das klassische Gesetz der großen Zahlen von Kolmogorov. Für den Beweis sei auf Hansen (1991) verwiesen.

3.36 Satz *Hansen (1991)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal, $r \geq 2$ mit $\mathrm{E}(X_t) = 0$, für alle t, und sei weiter $\sum_{k=0}^{\infty} \psi_k < \infty$.

(i) Sei $\sum_{t=1}^{\infty} c_t^2 < \infty$. Dann konvergiert $\sum_{t=1}^{n} X_t$ fast sicher für $n \to \infty$.

(ii) Sei $\sum_{t=1}^{\infty} t^{-2} c_t^2 < \infty$. Dann gilt:

$$\frac{1}{n} \sum_{t=1}^{n} X_t \longrightarrow 0 \quad \text{fast sicher für} \quad n \to \infty.$$

In weiteren Arbeiten vor allem von de Jong (1996) und von Chen und White (1996) gibt es weitere Abschwächungen der Voraussetzungen für starke Gesetze der großen Zahlen. Die Dämpfungsgeschwindigkeit der Folge der Mixingal-Koeffizienten $\{\psi_m\}_{m=0}^{\infty}$ kann dann deutlich langsamer gewählt werden, was gleichzeitig bedeutet, dass ein längeres Gedächtnis angenommen werden kann.
Hier soll ein einfacheres Ergebnis von Chen und White (1996) aufgeführt werden, das aber allgemein für L^r-Mixingale mit $r \geq 1$ gilt.

3.37 Satz *Chen und White (1996)*
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingale, $r \geq 1$, mit $\sup_t \mathrm{E}(|X_t|^p) < \infty$ für ein $p > 1$. Sei weiter $\{c_t\}_{-\infty}^{\infty}$ eine gleichmäßig beschränkte Folge und gelte $\psi_m = O((\ln m)^{-2-\beta})$ für ein ausreichend großes $\beta > 0$. Dann gilt:

$$\frac{\ln n}{n} \sum_{t=1}^{n} X_t \longrightarrow 0 \quad \text{fast sicher für} \quad n \to \infty.$$

Die Bedingung an die Mixingal-Koeffizienten in Satz 3.37 ist schwächer als die Bedingung $\sum_{k=0}^{\infty} \psi_k < \infty$ aus Satz 3.36. Außerdem kann die Aussage für das folgende Korollar vereinfacht werden, da $1/n$ schneller gegen Null als $\ln n/n$ für $n \to \infty$ konvergiert, so dass direkt folgendes Korollar gilt.

3.38 Korollar

Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal, $r \geq 1$, mit $\sup_t \mathrm{E}(|X_t|^p) < \infty$ für ein $p > 1$. Sei weiter $\{c_t\}_{-\infty}^{\infty}$ eine gleichmäßig beschränkte Folge und gelte $\sum_{m=0}^{\infty} \psi_m < \infty$. Dann gilt:

$$\frac{1}{n} \sum_{t=1}^{n} X_t \longrightarrow 0 \quad \text{fast sicher für} \quad n \to \infty.$$

Eine Zusammenfassung zu den starken Gesetzen der großen Zahlen für Mixingale und einer Diskussion der notwendigen Voraussetzungen findet sich bei Davidson und de Jong (1997). Unabhängig von den verschiedenen Sätzen kann die zentrale Bedingung $\sum_{k=0}^{\infty} \psi_k < \infty$ bei speziellen stochastischen Prozessen noch sehr viel stärker abgeschwächt werden, da dann die Annahmen an die Abhängigkeitsstruktur des Prozesses konkret vorgegeben sind.

3.39 Beispiel

Betrachte den heterogenen linearen Prozess aus Beispiel 3.29 (i)

$$X_t = \sum_{k=-\infty}^{\infty} a_k e_{t-k}$$

mit der Martingal-Differenzen-Folge $\{e_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ als Innovationen mit $\sup_{t \geq 1} \|e_t\|_r = B$ und $B < \infty$ für ein $r \geq 2$ und $\mathrm{E}(e_t) = 0$ für alle t sowie Konstanten $\{a_k\}_{-\infty}^{\infty}$ mit $\sum_{k=-\infty}^{\infty} |a_k| < \infty$ und sei $\mathcal{A}_t = \sigma\{e_j : j \leq t\}$. Dann ist der Prozess ein L^r-Mixingal für $r \geq 2$. Auf Grund der Linearität lässt sich hier die Konvergenz direkt beweisen, denn mit Lemma 3.34 gilt:

$$\begin{aligned}
\left\| \max_{j \leq n} \left| \sum_{t=1}^{j} \frac{1}{t} X_t \right| \right\|_r &\leq K_1 \sum_{k=-\infty}^{\infty} \left(\sum_{t=1}^{n} \left\| \frac{1}{t} \mathrm{E}(X_t | \mathcal{A}_{t-k}) - \mathrm{E}(X_t | \mathcal{A}_{t-k-1}) \right\|_r^2 \right)^{1/2} \\
&= K_1 \sum_{k=-\infty}^{\infty} \left(\sum_{t=1}^{n} \left\| \frac{1}{t} a_k e_{t-k} \right\|_r^2 \right)^{1/2} \\
&\leq B \sum_{k=-\infty}^{\infty} |a_k| \left(\sum_{t=1}^{n} \left(\frac{1}{t} \right)^2 \right)^{1/2} < \infty.
\end{aligned}$$

Mit dem Cauchy-Kriterium und dem Kronecker-Lemma (vgl. den Beweis zu Satz 3.36) sind die Voraussetzungen

$$\sup_{t\geq 1} \|e_t\|_r = B < \infty \quad \text{und} \quad \sum_{k=-\infty}^{\infty} |a_k| < \infty$$

hinreichend für das starke Gesetz der großen Zahlen.
Andererseits gilt für die Mixingal-Koeffizienten, wie in Beispiel 3.29(i) gezeigt,

$$\psi_m = \sum_{k=m}^{\infty} (|a_k| + |a_{-k}|),$$

für alle $m \geq 0$. Dies ist damit auch die Bedingung für die Gültigkeit des starken Gesetzes der großen Zahlen bei heterogenen linearen Prozessen.

Wie bei allen stochastischen Prozessen ist auch bei Mixingalen die Geschwindigkeit der Konvergenz interessant. Exponentielle Grenzen für die Wahrscheinlichkeit, dass ein stochastischer Prozess konvergiert, garantieren eine schnelle Konvergenz. In der Literatur sind solche Schranken bisher nicht hergeleitet worden.
In der statistischen Lerntheorie sind solche Aussagen allerdings essentiell. Für die folgende exponentielle Schranke für Mixingale zur Abschätzung der Konvergenzrate wird die Beschränktheit aufsummierter Mixingale unter Verwendung des Maximums der Summe genutzt. Dadurch sind die Grenzen nicht sehr scharf, allerdings haben sie ihre Vorzüge durch ihre Einfachheit und die Voraussetzungen korrespondieren zu denen der starken Gesetze der großen Zahlen.

3.40 Satz
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal, $r \geq 2$, und gelte für die Mixingal-Koeffizienten $\sum_{m=0}^{\infty} \psi_m < \infty$, dann existiert eine Konstante $K_2 < \infty$, so dass folgende Abschätzung gilt:

$$P\left(\left|\sum_{t=1}^{n} X_t\right| > \varepsilon\right) \quad \leq \quad 3\exp\left\{-\frac{\varepsilon}{K_2\left(\sum_{t=1}^{n} c_t^2\right)^{1/2}}\right\}.$$

Beweis:
Mit Lemma 3.34 und 3.35 sowie wegen der Monotonie der Exponentialfunktion und der L^r-Norm gilt folgende Ungleichung für alle $a > 0$:

$$\begin{aligned} \mathrm{E}\left(\exp\left\{\| a\,|S_n|\,\|_r\right\}\right) &= \exp\left\{\| a\,|S_n|\,\|_r\right\} \\ &\leq \exp\left\{\left\| a \max_{1\leq j\leq n}|S_j|\right\|_r\right\} \\ &\leq \exp\left\{a\,K_2\left(\sum_{t=1}^{n} c_t^2\right)^{1/2}\right\} \end{aligned}$$

mit Konstante K_2 aus Lemma 3.35.
Da die Funktion $g(x) = \exp\{\|a\,x\|_r\}$ mit $a > 0$ eingeschränkt auf die positiven reellen Zahlen die Voraussetzung der verallgemeinerten Markov-Ungleichung erfüllt, gilt für $\varepsilon > 0$:

$$\begin{aligned} P\left(|S_n| > \varepsilon\right) &\leq \frac{\mathrm{E}\left(\exp\left\{\| a\,|S_n|\,\|_r\right\}\right)}{\exp\left\{\| a\,\varepsilon\|_r\right\}} \\ &\leq \frac{\exp\left\{a\,K_2\left(\sum_{t=1}^{n} c_t^2\right)^{1/2}\right\}}{\exp\{a\,\varepsilon\}} \\ &= \exp\left\{-a\,\varepsilon + a\,K_2\left(\sum_{t=1}^{n} c_t^2\right)^{1/2}\right\} \qquad \forall\ a > 0. \end{aligned}$$

Für $a = K_2^{-1}\left(\sum_{t=1}^{n} c_t^2\right)^{-1/2}$ folgt für $\varepsilon > 0$:

$$\begin{aligned} P\left(|S_n| > \varepsilon\right) &\leq \exp\left\{-\frac{\varepsilon}{K_2\left(\sum_{t=1}^{n} c_t^2\right)^{1/2}} + 1\right\} \\ &\leq 3\exp\left\{-\frac{\varepsilon}{K_2\left(\sum_{t=1}^{n} c_t^2\right)^{1/2}}\right\}. \end{aligned}$$

□

Für die Konstante K_2 aus Lemma 3.35 gilt

$$K_2 = 4\, K_1 \sum_{k=0}^{\infty} \psi_k$$

mit Konstante $K_1 = C^{1/r} \frac{r}{r-1}$, wobei $C = C(r)$ ebenfalls eine Konstante ist, wie aus dem Beweis zu Lemma 3.34 hervorgeht. Weiterhin geht aus dem Beweis der Burkholder-Ungleichung (vgl. zum Beispiel Hall und Heyde, 1980, S. 23) hervor, dass die Konstante $C = (18r(\frac{r}{r-1})^{1/2})^r$ eine geeignete Wahl ist. Es ergibt sich insgesamt:

$$\begin{aligned} K_2 &= 4 \left(18\, r \left(\frac{r}{r-1} \right)^{1/2} \right) \frac{r}{1-r} \sum_{k=0}^{\infty} \psi_k \\ &= 72\, r \left(\frac{r}{r-1} \right)^{3/2} \sum_{k=0}^{\infty} \psi_k. \end{aligned}$$

Mit diesen Überlegungen folgt direkt ein Korollar zum vorherigen Satz.

3.41 Korollar
Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal, $r \geq 2$, und gelte $\sum_{m=0}^{\infty} \psi_m < \infty$. Dann gilt:

$$P\left(\left| \sum_{t=1}^{n} X_t \right| > \varepsilon \right) \leq 3 \exp \left\{ - \frac{\varepsilon}{72\, r \left(\frac{r}{r-1} \right)^{3/2} \sum_{k=0}^{\infty} \psi_k \left(\sum_{t=1}^{n} c_t^2 \right)^{1/2}} \right\}.$$

Die Abschätzung über das Maximum der Summen eines Mixingals ist relativ grob, denn der Informationsverlust wirkt sich stark auf die Genauigkeit der exponentiellen Abschätzungen ab. Eine engere exponentielle Wahrscheinlichkeitsschranke für Mixingale, bei der nur die Beschränktheit des Mixingals ausgenutzt wird, ähnlich wie bei der Schranke für Martingale in Satz 3.23, ergibt sich aus Abschätzungen der Summe bezüglich der L_r-Norm.

3.42 Satz

Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal, $r \geq 1$, mit $|X_t| < B_t$ fast sicher für alle t, wobei $\{B_t\}_{-\infty}^{\infty}$ eine Folge positiver Konstanten ist, und mit der zugehörigen Nullfolge $\{\psi_m\}_0^{\infty}$ und der zugehörigen beschränkten Folge $\{c_t\}_{-\infty}^{\infty}$, dann gilt für $S_n = \sum_{t=1}^{n} X_t$:

$$P\left(S_n > \varepsilon\right) \leq \exp\left\{\frac{-\varepsilon^2}{2\sum_{t=1}^{n} B_t^2}\right\} \prod_{t=1}^{n} \left(\frac{(\psi_0 + \psi_1)c_t}{B_t} + 1\right)$$

und

$$P\left(|S_n| > \varepsilon\right) \leq 2\exp\left\{\frac{-\varepsilon^2}{2\sum_{t=1}^{n} B_t^2}\right\} \prod_{t=1}^{n} \left(\frac{(\psi_0 + \psi_1)c_t}{B_t} + 1\right).$$

Beweis:

Wegen der Konvexität der Exponentialfunktion gilt für jedes $x \in [-B_t, B_t]$

$$\exp\{ax\} \leq \frac{(B_t + x)\exp\{aB_t\} + (B_t - x)\exp\{-aB_t\}}{2B_t} \quad \forall\ a > 0.$$

Mit der Mixingal-Eigenschaft $\|\mathrm{E}(X_t|\mathcal{A}_{t-1})\|_r \leq \psi_1 c_t$ ergibt sich für alle t:

$$\begin{aligned}
&\|\mathrm{E}\left(\exp\{aX_t\}\,|\mathcal{A}_{t-1}\right)\|_r \\
&\quad\leq \left\|\frac{\left(B_t + \mathrm{E}\left(X_t|\mathcal{A}_{t-1}\right)\right)\exp\{aB_t\} + \left(B_t - \mathrm{E}\left(X_t|\mathcal{A}_{t-1}\right)\right)\exp\{-aB_t\}}{2B_t}\right\|_r \\
&\quad\leq \frac{\left(\|B_t\|_r + \psi_1 c_t\right)\|\exp\{aB_t\}\|_r + \left(\|B_t\|_r + \psi_1 c_t\right)\|\exp\{-aB_t\}\|_r}{2\,\|B_t\|_r} \\
&\quad= \frac{1}{2}\left(\exp\{aB_t\} + \exp\{-aB_t\}\right) + \frac{\psi_1 c_t}{2B_t}\left(\exp\{aB_t\} + \exp\{-aB_t\}\right) \\
&\quad\leq \exp\left\{\frac{1}{2}a^2 B_t^2\right\}\left(\frac{\psi_1 c_t}{B_t} + 1\right) \\
&\quad\leq \exp\left\{\frac{1}{2}a^2 B_t^2\right\}\left(\frac{(\psi_0 + \psi_1)c_t}{B_t} + 1\right) \quad \text{fast sicher für alle } a > 0,
\end{aligned}$$

wobei die vorletzte Ungleichung aus der Reihenentwicklung der Exponentialfunktion folgt, die letzte Ungleichung gilt wegen $\psi_0 \geq 0$. Mit der gleichen Argumentation, aber

der Mixingal-Eigenschaft $\|X_t\|_r \leq (\psi_0 + \psi_1)c_t$ für $r \geq 1$ gilt für alle t:

$$\begin{aligned}
&\|\exp\{aX_t\}\|_r \\
&\quad\leq \frac{(\|B_t\|_r + (\psi_0+\psi_1)c_t)\,\|\exp\{aB_t\}\|_r + (\|B_t\|_r + (\psi_0+\psi_1)c_t)\,\|\exp\{-aB_t\}\|_r}{2\,\|B_t\|_r} \\
&\quad= \frac{1}{2}\left(\exp\{aB_t\} + \exp\{-aB_t\}\right) + \frac{(\psi_0+\psi_1)c_t}{2B_t}\left(\exp\{aB_t\} + \exp\{-aB_t\}\right) \\
&\quad\leq \exp\left\{\frac{1}{2}a^2B_t^2\right\}\left(\frac{(\psi_0+\psi_1)c_t}{B_t} + 1\right) \quad \text{fast sicher für alle } a > 0,
\end{aligned}$$

Für die Summe $S_n = \sum_{t=1}^{n} X_t$ gilt mit der Hölder-Ungleichung, sowie mit der Ljapunov-Ungleichung (vgl. Mitrinović, 1970, S. 50-54):

$$\begin{aligned}
&\mathrm{E}\left(\mathrm{E}\left(\exp\{aS_n\}\,|\mathcal{A}_{n-1}\right)\right) \\
&\quad= \mathrm{E}\left(\left|\mathrm{E}\left(\exp\{aS_n\}\,|\mathcal{A}_{n-1}\right)\right|\right) \\
&\quad= \mathrm{E}\left(\left|\mathrm{E}\left(\exp\{aX_n + aS_{n-1}\}\,|\mathcal{A}_{n-1}\right)\right|\right) \\
&\quad= \mathrm{E}\left(\left|\mathrm{E}\left(\exp\{aX_n\}\,|\mathcal{A}_{n-1}\right)\exp\{aS_{n-1}\}\right|\right) \\
&\quad\leq \left\|\mathrm{E}\left(\exp\{aX_n\}\,|\mathcal{A}_{n-1}\right)\right\|_r \mathrm{E}\left(\left|\exp\{aS_{n-1}\}\right|^{\frac{r}{r-1}}\right)^{\frac{r-1}{r}},\ r > 1 \\
&\quad\leq \left\|\mathrm{E}\left(\exp\{aX_n\}\,|\mathcal{A}_{n-1}\right)\right\|_r \mathrm{E}\left(\left|\exp\{aS_{n-1}\}\right|\right) \\
&\quad\leq \exp\left\{\frac{1}{2}a^2B_n^2\right\}\left(\frac{(\psi_0+\psi_1)c_n}{B_n} + 1\right)\mathrm{E}\left(\left|\exp\{aS_{n-1}\}\right|\right) \quad \text{fast sicher für alle } a > 0.
\end{aligned}$$

Dabei gilt die Ljapunov-Ungleichung, weil $0 < \frac{r}{r-1} < 1$. Eine Verallgemeinerung dieses Ansatzes der sukzessiven Auflösung der bedingten Erwartungswerte führt zu:

$$\begin{aligned}
&\mathrm{E}\left(\exp\{aS_n\}\right) \\
&\quad= \mathrm{E}\left(\mathrm{E}\left(\ldots\mathrm{E}\left(\mathrm{E}\left(\exp\{aS_n\}\,|\mathcal{A}_{n-1}\right)|\mathcal{A}_{n-2}\right)\ldots|\mathcal{A}_1\right)\right) \\
&\quad= \mathrm{E}\left(\,\left|\mathrm{E}\left(\ldots\mathrm{E}\left(\mathrm{E}\left(\exp\{aX_n + aS_{n-1}\}\,|\mathcal{A}_{n-1}\right)|\mathcal{A}_{n-2}\right)\ldots|\mathcal{A}_1\right)\right|\,\right) \\
&\quad\leq \left\|\mathrm{E}\left(\exp\{aX_n\}\,|\mathcal{A}_{n-1}\right)\right\|_r \mathrm{E}\left(\,\left|\mathrm{E}\left(\ldots\mathrm{E}\left(\exp\{aS_{n-1}\}\,|\mathcal{A}_{n-2}\right)\ldots|\mathcal{A}_1\right)\right|\,\right) \\
&\quad\leq \ldots
\end{aligned}$$

$$\leq \prod_{t=0}^{n-2} \left\| \mathrm{E}\left(\exp\left\{aX_{n-t}\right\} | \mathcal{A}_{n-1-t}\right)\right\|_r \mathrm{E}\left(\left|\exp\left\{aX_1\right\}\right|\right)$$

$$\leq \prod_{t=2}^{n} \left[\exp\left\{\frac{1}{2}a^2 B_t^2\right\}\left(\frac{(\psi_0+\psi_1)c_t}{B_t}+1\right)\right] \left\| \exp\left\{aX_1\right\}\right\|_r$$

fast sicher für $r \geq 1$ für alle $a > 0$.

Einen nochmalige Anwendung der Ljapunov-Ungleichung für $r \geq 1$ ergibt:

$$\mathrm{E}\left(\exp\left\{aS_n\right\}\right) \leq \prod_{t=1}^{n} \left[\exp\left\{\frac{1}{2}a^2 B_t^2\right\}\left(\frac{(\psi_0+\psi_1)c_t}{B_t}+1\right)\right]$$

$$= \exp\left\{\frac{1}{2}a^2 \sum_{t=1}^{n} B_t^2\right\} \prod_{t=1}^{n}\left(\frac{(\psi_0+\psi_1)c_t}{B_t}+1\right)$$

fast sicher für $r \geq 1$ für alle $a > 0$.

Mit der verallgemeinerten Markov-Ungleichung gilt für $\varepsilon > 0$

$$P\left(S_n > \varepsilon\right) \leq \frac{\mathrm{E}\left(\exp\left\{aS_n\right\}\right)}{\exp\left\{a\varepsilon\right\}}$$

$$\leq \exp\left\{\frac{1}{2}a^2 \sum_{t=1}^{n} B_t^2\right\} \prod_{t=1}^{n}\left(\frac{(\psi_0+\psi_1)c_t}{B_t}+1\right) \exp\left\{-a\varepsilon\right\}$$

$$\leq \exp\left\{\frac{1}{2}a^2 \sum_{t=1}^{n} B_t^2 - a\varepsilon\right\} \prod_{t=1}^{n}\left(\frac{(\psi_0+\psi_1)c_t}{B_t}+1\right)$$

für alle $a > 0$. Mit $a = \varepsilon\left(\sum_{t=1}^{n} B_t^2\right)^{-1}$ folgt:

$$P\left(S_n > \varepsilon\right) \leq \exp\left\{\frac{-\varepsilon^2}{2\sum_{t=1}^{n} B_t^2}\right\} \prod_{t=1}^{n}\left(\frac{(\psi_0+\psi_1)c_t}{B_t}+1\right).$$

Ein analoges Ergebnis erhält man für $-S_n$, so dass das Ergebnis für $|S_n|$ durch Aufsummierung der beiden resultierenden Ungleichungen erfolgt.

□

Die Voraussetzungen dieses Satzes sind in Hinsicht auf die Beschränktheit der Koordinaten des Mixingals relativ schwach, denn es reicht neben den Mixingal-Eigenschaften die fast sichere Beschränktheit. Eine nicht sehr starke Verschärfung ist demgegenüber eine gleichmäßige Beschränkung der Mixingal-Elemente $|X_t| < B < \infty$ für alle t fast sicher. Diese Annahme führt zu einer Vereinfachung des obigen Satzes.

3.43 Korollar

Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal mit $|X_t| < B < \infty$ fast sicher für alle t und mit der zugehörigen Nullfolge $\{\psi_m\}_0^{\infty}$ und beschränkter Folge $\{c_t\}_{-\infty}^{\infty}$, wobei $c_{\sup} = \sup_t(c_t)$. Dann gilt für $S_n = \sum_{t=1}^{n} X_n$:

$$P\left(S_n > \varepsilon\right) \;\leq\; \exp\left\{\frac{-\varepsilon^2}{2nB^2}\right\}\left(\frac{(\psi_0+\psi_1)c_{\sup}}{B}+1\right)^n$$

und

$$P\left(|S_n| > \varepsilon\right) \;\leq\; 2\exp\left\{\frac{-\varepsilon^2}{2nB^2}\right\}\left(\frac{(\psi_0+\psi_1)c_{\sup}}{B}+1\right)^n.$$

Das Ergebnis folgt direkt aus Satz 3.42 mit $B_t = B$ und wegen $c_t \leq c_{\sup}$ für alle t. Aus Satz 3.42 ergibt sich direkt ein weiteres Resultat für das Beobachtungsmittel.

3.44 Korollar

Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal mit $|X_t| < B < \infty$ fast sicher für alle t und mit den zugehörigen Folgen $\{\psi_m\}_0^{\infty}$ und $\{c_t\}_{-\infty}^{\infty}$. Sei weiter $c_{\sup} = \sup_t(c_t)$, dann gilt:

$$P\left(\overline{X}_n > \varepsilon\right) \;\leq\; \exp\left\{\frac{-n\varepsilon^2}{2B^2}\right\}\left(\frac{(\psi_0+\psi_1)c_{\sup}}{B}+1\right)^n$$

und

$$P\left(|\overline{X}_n| > \varepsilon\right) \;\leq\; 2\,\exp\left\{\frac{-n\varepsilon^2}{2B^2}\right\}\left(\frac{(\psi_0+\psi_1)c_{\sup}}{B}+1\right)^n.$$

Es ergibt sich mit Satz 3.42, und damit auch aus Korollar 3.43 und Korollar 3.44, das gewünschte Resultat, dass der Grenzwert für die Wahrscheinlichkeit großer Summenwerte gegen Null konvergiert für $n \to \infty$. Insbesondere ist bemerkenswert, dass die obere Schranke für $P(|\overline{X}_n| > \varepsilon)$ nur von den ersten beiden Mixingal-Koeffizienten abhängt. Die Länge des Gedächtnisses des Mixingals spielt also nur eine untergeordnete Rolle. Ein wesentlicher Grund dafür ist, dass die Folge der ψ_m nach Definition eine monoton fallende Nullfolge ist.

Trotzdem sind die Schranken in Satz 3.42 und Korollar 3.43 für die Wahrscheinlichkeit, dass die Summe eines stochastischen Prozesses mit Mixingal-Eigenschaften konvergiert, nicht so eng wie bei Martingalen (vgl. Satz 3.23 und Korollar 3.24). Dies hat seine Ursache darin, dass bei Mixingalen ein längeres Gedächtnis, je nach Wahl der Mixingal-Koeffizienten, zugelassen wird. Für Koeffizienten $\psi_m = 0$ für alle $m \geq 0$, wenn also kein Gedächtnis zugelassen wird, fallen die Schranken zusammen. Dies gilt für zentrierte Martingale stets, denn nach Beispiel 3.28 muss dann $\psi_m = 0$ für alle $m \geq 1$ gelten und auf Grund der Zentrierung kann auch $\psi_0 = 0$ gewählt werden.

Bei einer geeigneten Wahl von B können die Schranken aus Korollar 3.43 so konstruiert werden, dass B nur noch von $c_{\sup}$ und den Mixingal-Koeffizienten ψ_0 und ψ_1 abhängt. Sei $B_1 = \frac{1}{\exp\{b_1-1\}-1}(\psi_0+\psi_1)c_{\sup}$ mit einer Konstanten $b_1 > 0$, dann gilt:

$$\begin{aligned} P\left(S_n > \varepsilon\right) &\leq \exp\left\{\frac{-\varepsilon^2}{2nB_1^2}\right\}\left(\frac{(\psi_0+\psi_1)c_{\sup}}{B_1}+1\right)^n \\ &= \exp\left\{\frac{-\varepsilon^2}{2nB_1^2} + n\ln\left(\frac{(\psi_0+\psi_1)c_{\sup}}{B_1}+1\right)\right\} \\ &= \exp\left\{-\frac{\varepsilon^2}{2n}\,\frac{(\exp\{b_1-1\}-1)^2}{(\psi_0+\psi_1)^2\,c_{\sup}^2} + n(b_1-1)\right\} \\ &\leq \exp\left\{-\frac{\varepsilon^2}{2n}\,\frac{(\exp\{b_1-1\}-1)^2}{(\psi_0+\psi_1)^2\,c_{\sup}^2} + nb_1\right\} \end{aligned}$$

Die X_t werden durch die obige Wahl von B_1 weiterhin gleichmäßig beschränkt, wenn b_1 nur klein genug gewählt wird, da $c_{\sup}$ durch die Mixingal-Eigenschaften immer als ein Vielfaches von $\|X_t\|_r$ angesehen werden kann.
Wenn B_1 eine gleichmäßige Schranke für die X_t ist, so ist ein $B^\star$ mit $b^\star < b_1$ auch eine gleichmäßige obere Schranke, denn dann gilt $B_1 < B^\star$. Für die Wahl von $b_{n_0} = 1/n$ für beliebiges $n > n_0$ gilt $B_1 < B_{n_0}$ und somit $|X_t| < B_{n_0}$, wenn nur n_0 so groß gewählt wird, dass $b_{n_0} < b_1$ gilt. Da die Wahl von B beliebig ist, solange nur die X_t gleichmäßig beschränkt sind, folgt aus Korollar 3.43 für $|X_t| < B_1 = \frac{1}{\exp\{b_1-1\}-1}(\psi_0+\psi_1)c_{\sup}$ mit $b_1 > 0$ die folgende Schranke mit $b_{n_0} = 1/n$ für beliebiges $n > n_0$ mit n_0, so dass

$B_1 < B_{n_0}$ ist:

$$\begin{aligned} P\left(S_n > \varepsilon\right) &\leq \exp\left\{-\frac{\varepsilon^2}{2n}\,\frac{(\exp\{b_{n_0}-1\}-1)^2}{(\psi_0+\psi_1)^2\,c_{\sup}^2} + nb_{n_0}\right\} \\ &= \exp\left\{-\frac{\varepsilon^2}{2n}\,\frac{\left(\exp\left\{1/n-1\right\}-1\right)^2}{(\psi_0+\psi_1)^2\,c_{\sup}^2} + 1\right\} \\ &< 3\,\exp\left\{-\frac{\varepsilon^2}{2n}\frac{\left(\exp\left\{1/n-1\right\}-1\right)^2}{(\psi_0+\psi_1)^2\,c_{\sup}^2}\right\}. \end{aligned}$$

Daraus ergibt sich als abschließender Satz folgende exponentielle Schranke zur Abschätzung der Konvergenzgeschwindigkeit.

3.45 Satz

Sei $\{X_t, \mathcal{A}_t\}_{-\infty}^{\infty}$ ein L^r-Mixingal mit $|X_t| < \frac{1}{\exp\{b-1\}-1}(\psi_0+\psi_1)c_{\sup} < \infty$ fast sicher für alle t, wobei $b > 0$, und mit der zugehörigen Nullfolge $\{\psi_m\}_0^{\infty}$ sowie der beschränkten Folge $\{c_t\}_{-\infty}^{\infty}$, wobei $c_{\sup} = \sup_t(c_t)$. Dann gibt es ein $n_0 = n_0(b)$, so dass für alle $n > n_0$ gilt:

$$P\left(\sum_{t=1}^{n} X_t > \varepsilon\right) < 3\,\exp\left\{-\frac{\varepsilon^2}{2n}\frac{\left(\exp\left\{1/n-1\right\}-1\right)^2}{(\psi_0+\psi_1)^2\,c_{\sup}^2}\right\}$$

und

$$P\left(\overline{X}_n > \varepsilon\right) < 3\,\exp\left\{-\frac{n\varepsilon^2}{2}\frac{\left(\exp\left\{1/n-1\right\}-1\right)^2}{(\psi_0+\psi_1)^2\,c_{\sup}^2}\right\}.$$

Die Wahl von n_0 hängt von der benötigten Konstante b ab. Je weniger restriktiv die gleichmäßige Beschränkung des Prozesses $\{X_t\}$ gewählt werden kann, desto größer darf auch b gewählt werden. Daraus folgt wiederum, dass die Aussage des obigen Satzes bereits für relativ kleines n_0 gilt. Andererseits kann bei Wahl eines großen Wertes für $c_{\sup}$ die Konstante b ebenfalls größer gewählt werden, so dass der Satz 3.39 wiederum auch für relativ kleines n_0 gilt, wobei dieses Vorgehen allerdings zu Lasten der Schärfe der Abschätzung durch die Schranke geht.

4 Das ERM-Prinzip bei Abhängigkeitsstrukturen

Die Einführung des statistischen Lernprozesses in Kapitel 2 mit den Definitionen zum statistischen Lernen, zum Lernproblem und zum Konzept der empirischen Risiko-Minimierung ist nicht darauf aufgebaut, dass die Beobachtungen $z_1, \ldots, z_n$ unabhängig und identisch verteilt erhoben worden sind. Für den Nachweis der Konsistenz des Prinzips der empirischen Risiko-Minimierung und die Bestimmung der Konvergenzrate ist dagegen die Unabhängigkeit der Beobachtungen vorausgesetzt worden. Insbesondere werden für das Kerntheorem, der statistischen Lerntheorie, welches der Schlüssel zur Konsistenz des ERM-Prinzips ist, bereits unabhängige Beobachtungen angenommen. Dementsprechend ist es notwendig, in diesem Kapitel die statistische Lerntheorie darauf hin zu untersuchen, in welchem Maße Abhängigkeitsstrukturen in den Daten zulässig sein dürfen, damit die Konsistenz des ERM-Prinzips erhalten und eine schnelle Konvergenzrate gewährleistet bleibt.

4.1 Der statistische Lernprozess bei Abhängigkeitsstrukturen

Unter der Annahme, die Beobachtungen $z_1, \ldots, z_n$ seien identisch verteilt gemäß der Verteilungsfunktion $F_Z(z)$, aber *nicht* unabhängig, muss ein Konzept aufgestellt werden, dass trotzdem die Nutzung des ERM-Prinzips erlaubt. Ein für eine Vielzahl von Fällen allgemeingültiges Konzept sind Abhängigkeitsstrukturen in stochastischen Prozessen, wie die im dritten Kapitel kurz vorgestellte Ergodentheorie oder das Mixing-Konzept. Für diese Konzepte existieren geeignete Aussagen zur Konsistenz, wie etwa Gesetze der großen Zahlen und auch exponentielle Schranken (vgl. Satz 3.10 sowie Hanson und Koopmans, 1965).

Probl ematisch ist dagegen die einfache Anwendung dieser Theorie für die Modellierung von Abhängigkeitsstrukturen in Datensätzen durch beispielsweise Zeitreihen, da, wie im dritten Kapitel beschrieben, diese Ansätze wahrscheinlichkeitstheoretisch- und ereignisorientiert sind. Im Gegensatz dazu steht der Ansatz für Martingale und Mixingale mit der Betrachtung eines stochastischen Prozesses als Folge von Paaren $\{X_i, \mathcal{A}_i\}_1^\infty$, wobei $\mathcal{A}_i = \sigma(X_1, \ldots, X_{i-1}, X_i)$ die durch $(X_1, \ldots, X_{i-1}, X_i)$ induzierte σ-Algebra ist. Durch die Nutzung dieses minimalen adaptierten Prozesses kann direkt

auf die Bildungsgesetzmäßigkeiten des Prozesses anhand der vorhergehenden Koordinaten der Folge $\{X_i\}_1^\infty$ geschlossen werden. Damit sind Abhängigkeitsstrukturen im räumlichen und zeitlichen Kontext denkbar. Allgemeingültige Beispiele sind unter anderen die heterogenen und homogenen linearen Prozesse aus Beispiel 3.29 bzw. 3.39. Durch die Einbindung dieser Prozesse wird beispielsweise auch die Anwendung auf Zeitreihen, wie MA-, AR- und ARMA-Prozesse, möglich (vgl. Andrews, 1988).

Für die Theorie des statistischen Lernens sind, wie in Kapitel 2 bereits aufgezeigt, die grundlegenden Aussagen zur Konsistenz mit zugehöriger exponentieller Konvergenzrate des ERM-Prinzips an die zentrale Frage geknüpft, ob die Folge der empirischen Risikos mindestens schwach gegen das Risiko konvergieren. Dargestellt durch die empirischen Verluste bzw. den erwarteten Verlusten, müssen dementsprechend die Ergebnisse aus den Abschnitten 3.3 und 3.4 zu Martingalen bzw. Mixingalen auf Aussagen der Form

$$\frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \xrightarrow{Pr} \int Q(z,\alpha)dF_Z(z) \quad \text{für } n\to\infty$$

angewandt werden. Das bedeutet, dass Gesetze der großen Zahlen und exponentielle Schranken für die Bestimmung der Konvergenzrate auf der Funktionenmenge $Q(z,\alpha)$, $\alpha\in\Lambda$, genutzt werden. Aus diesem Grund erweist es sich als sinnvoll, die Abhängigkeitsstruktur in den Daten direkt über die Verluste $L(y,f(x,\alpha))$, also die Funktionen $Q(z,\alpha)$, zu definieren.
Eine Konstruktion für die statistische Lerntheorie bei Abhängigkeiten in den Beobachtungen in dem Sinne, dass direkt für die Beobachtungen $z_1,\ldots,z_n$ eine Martingal- oder Mixingal-Struktur angenommen wird, hat den entscheidenden Nachteil, dass die Gesetzmäßigkeiten für Martingale oder Mixingale bei einer Transformation

$$z\to Q(z,\alpha),\quad \alpha\in\Lambda,$$

geeignet übernommen werden müssen. Die gewählte Abhängigkeitsstruktur muss also invariant gegenüber der Abbildung

$$Q(\cdot,\alpha):\mathbb{R}^{d+1}\to\mathbb{R}$$

für beliebiges $\alpha\in\Lambda$ sein. Dies ist eine starke Einschränkung für die zu wählende Menge von Funktionen $Q(z,\alpha)$, $\alpha\in\Lambda$, und damit für die Wahl sowohl der Verlustfunktion

$L(y, f(x, \alpha))$ als auch der zugelassenen Menge von Funktionen $f(x, \alpha)$, $\alpha \in \Lambda$. Bereits der einfache Fall mit euklidischer Verlustfunktion

$$Q(z, \alpha) = L(y, f(x, \alpha)) = (y - f(x, \alpha))^2, \quad \alpha \in \Lambda,$$

macht deutlich, dass die Abhängigkeitsstruktur in den Beobachtungen $z_1, \ldots, z_n$ in der Regel nicht auf die Verluste $Q(z_i, \alpha)$, $i = 1, \ldots, n$, übertragbar ist. Die Entwicklung der statistischen Lerntheorie bei Abhängigkeiten würde somit auf einzelne Spezialfälle eingeschränkt, was aber der Intention, eine möglichst generelle Theorie zu entwickeln, widerspricht.

Zur Modellierung von Abhängigkeitsstrukturen in den Verlusten, an Stelle einer Modellierung in den Beobachtungen, durch das Martingal- oder Mixingal-Konzept wird für jedes $\alpha \in \Lambda$ für die Folge von Funktionswerten $\{Q(z_i, \alpha)\}_1^\infty$ die durch $(Q(z_j, \alpha), j = 1, \ldots, i)$ induzierte σ-Algebra $\mathcal{A}_i = \sigma(Q(z_j, \alpha), j = 1, \ldots, i)$, $i = 1, 2, \ldots$, eingeführt. Für jedes $\alpha \in \Lambda$ ist das Paar damit ein adaptierter Prozess, dem die gewünschte Abhängigkeitsstruktur zugeordnet werden kann.

4.1 Definition *Abhängigkeitsstrukturen*
Seien $z_1, \ldots, z_n, \ldots$ identisch gemäß $F_Z(z)$ verteilte Beobachtungen. Weiterhin sei $Q(z, \alpha)$, $\alpha \in \Lambda$, eine Menge von Funktionen und sei $\{Q(z_i, \alpha)\}_1^\infty$ für jedes $\alpha \in \Lambda$ eine Folge von Funktionswerten bezüglich der Beobachtungen $z_1, \ldots, z_n, \ldots$. Dann wird die *Abhängigkeitsstruktur in den Daten* über die adaptierten Prozesse $\{Q(z_i, \alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, definiert:

(i) Seien die adaptierten Prozesse $\{Q(z_i, \alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, jeweils eine Folge von Martingal-Differenzen. Dann heißt die Abhängigkeitsstruktur in den Daten *Martingal-Struktur*.

(ii) Seien die adaptierten Prozesse $\{Q(z_i, \alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$ jeweils ein L^r-Mixingal, $r \geq 1$, mit der zugehörigen Nullfolge $\{\psi_m\}_0^\infty$ und der zugehörigen beschränkten Folge $\{c_i\}_1^\infty$. Dann heißt die Abhängigkeitsstruktur in den Daten *L^r-Mixingal-Struktur*.

Dieses spezielle Konzept, Abhängigkeitsstrukturen in Daten über die Verluste zu definieren, muss so aufgefasst werden, dass bei der Modellierung des funktionalen Zusam-

menhangs mittels einer Funktion $f(x, \alpha)$ aus einer vorgegebenen Menge $f(\cdot, \alpha)$, $\alpha \in \Lambda$, ein Modellierungsfehler nicht ausgeschlossen wird, wobei dieser Fehler gemäß eines stochastischen Prozesses mit Abhängigkeitsstruktur modelliert wird. Wie in Definition 4.1 eingeführt, müssen dann die empirischen Verluste für jede Beobachtung, also der Verlust $Q(z_i, \alpha)$, $i = 1, \ldots, n, \ldots$, als stochastischer Prozess mit gleicher Abhängigkeitsstruktur aufgefasst werden. Bei Annahme einer zeitlichen Abhängigkeit bedeutet dies beispielsweise, dass der empirische Verlust $Q(z_i, \alpha)$ zum Zeitpunkt i von den vorhergehenden Verlusten $Q(z_{i-1}, \alpha), Q(z_{i-2}, \alpha), \ldots, Q(z_1, \alpha)$ abhängt. Die gemachten Fehler beeinflussen somit je nach Gedächtnis des Fehler-Prozesses die neuen Fehler. Allgemeingültige Beispiele im Kontext der empirischen Risiko-Minimierung für die Nutzung stochastischer Prozesse bei der Modellierung der Abhängigkeitsstruktur sind lineare Prozesse.

4.2 Beispiel *homogener linearer Prozess*
Sei die Folge der empirischen Verluste $\{Q(z_i, \alpha)\}_1^\infty$ ein homogener linearer Prozess, wie im Beispiel 3.29 eingeführt. Das bedeutet, dass mit einer schwach stationären Folge $\{e_i\}_{-\infty}^\infty$ von unabhängigen Zufallsvariablen mit $\mathrm{E}(e_i) = 0$ und beschränkter Varianz $\mathrm{Var}(e_i) < \infty$ für alle i sowie mit Folgen von Konstanten $\{a_k(\alpha)\}_0^\infty$ mit $\sum_{k=0}^\infty |a_k(\alpha)| < \infty$ für die Funktionen $Q(z_i, \alpha)$ für alle $\alpha \in \Lambda$ gilt:

$$Q(z_i, \alpha) = \sum_{k=0}^{\infty} a_k(\alpha) e_{i-k}.$$

Die Folgen von Paaren $\{Q(z_i, \alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, sind dann jeweils ein L^2-Mixingal mit Konstanten

$$c_i = \mathrm{Var}(e_i), \quad i = 1, 2, \ldots,$$

und den Mixingal-Koeffizienten

$$\psi_m(\alpha) = \left(\sum_{k=m}^{\infty} a_k^2(\alpha) \right)^{1/2}, \quad m = 0, 1, 2, \ldots,$$

wobei $\mathcal{A}_i = \sigma(e_j, -\infty < j \leq i)$ die durch $(e_i, e_{i-1}, \ldots)$ induzierte σ-Algebra ist. Die Abhängigkeiten in den Daten haben somit eine L^2-Mixingal-Struktur.

Ein weiteres Beispiel, bei dem die Abhängigkeitsstruktur in den Daten über die Modellfehler ausgedrückt wird, ist die Theorie der seriellen Korrelation im Regressionsmodell. Bei diesem Ansatz wird ein additiver Fehler u zusätzlich zum funktionalen Zusammenhang angenommen, so dass

$$y = f(x, \alpha) + u$$

mit $\alpha \in \Lambda$ und $\mathrm{E}(u) = 0$ gilt, wobei u als linearer Prozess, beispielsweise als AR-Prozess modelliert wird (vgl. Davidson und MacKinnon, 1993).

In diesem Zusammenhang ist die Bemerkung wichtig, dass von der Abhängigkeitsstruktur in den Verlusten in der Regel nicht auf die Struktur in den Beobachtungen zurück geschlossen werden kann, da die diversen Abhängigkeitskonzepte nur in Spezialfällen invariant gegenüber Transformationen sind. Diese Einschränkung des Konzepts ist wegen des generalistischen Ansatzes der statistischen Lerntheorie im Allgemeinen kein Problem in der praktischen Anwendung, da die Folge der empirischen Verluste $\{Q(z_i, \alpha)\}_1^\infty$ bei der Anwendung des ERM-Prinzips bzw. bei der Anwendung der Algorithmen zur strukturellen Risiko-Minimierung berechenbar ist. Somit lässt sich überprüfen, ob die Abhängigkeiten in den Fehlern geeignet modelliert werden können.
Im den folgenden Abschnitten wird für die hier vorgestellten Martingal- und Mixingal-Strukturen nachgewiesen unter welchen Voraussetzungen an diese Abhängigkeiten in den Daten das Konzept der statistischen Lerntheorie anwendbar bleibt.

4.2 Das Kerntheorem der statistischen Lerntheorie unter Abhängigkeitsstrukturen

Mit der Einführung der nicht-trivialen Konsistenz für das Prinzip des empirischen Risikos im zweiten Kapitel gibt es ein Konzept, das für eine Menge von Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, Konsistenz im Sinne der generellen Eigenschaften der Menge sicher stellt (vgl. Abschnitt 2.2). Allerdings muss für den Nachweis der nicht-trivialen Konsistenz für unendlich viele Teilmengen die Konvergenz gezeigt werden, so dass das Verfahren in der Regel nicht praktikabel ist. Das Kerntheorem der statistischen Lerntheorie ist somit der Schlüssel, um das Prinzip der empirischen Risiko-Minimierung

anwenden zu können. Erst dadurch, dass in diesem Theorem die einseitige gleichmäßige Konvergenz als eine notwendige und hinreichende Bedingung für die nicht-triviale Konsistenz des ERM-Prinzips bereitgestellt wird, kann der Nachweis der Konsistenz konstruktiv geführt werden. Das Kerntheorem (Satz 2.19 in dieser Arbeit) von Vapnik und Chervonenkis (1989) gilt dabei aber nur für unabhängig und identisch verteilte Beobachtungen.
Eine Erweiterung der Aussagen des Satzes 2.19 für ein Kerntheorem der statistischen Lerntheorie für Martingal- und L^r-Mixingal-Strukturen, erlaubt, ebenso wie bei Unabhängigkeit in den Daten, die Anwendung der empirischen Risiko-Minimierung als generelles Prinzip. Dazu wird unter einfachen Annahmen nachgewiesen, dass die gleichmäßige einseitige Konvergenz weiterhin notwendige und hinreichende Bedingung für die nicht-triviale Konsistenz des ERM-Prinzips ist, so dass die Aussage des Kerntheorems der statistischen Lerntheorie unter Abhängigkeitsstrukturen dieselbe bleibt. In formaler Schreibweise bedeutet das:
Unter der Voraussetzung, dass für alle Funktionen aus der Menge $Q(z, \alpha)$, $\alpha \in \Lambda$, und für eine gegebene Verteilungsfunktion $F_Z(z)$ die Ungleichungen

$$a \leq \int Q(z, \alpha) dF_Z(z) \leq A, \quad \alpha \in \Lambda,$$

mit beliebigen, aber festen Konstanten a und A gelten, muss gezeigt werden, dass die folgenden Aussagen äquivalent sind:

(i) Für die gegebene Verteilungsfunktion $F_Z(z)$ ist die ERM-Methode strikt konsistent auf der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$.

(ii) Für die gegebene Verteilungsfunktion $F_Z(z)$ gilt gleichmäßige einseitige Konvergenz des arithmetischen Mittels gegen den Erwartungswert auf der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$:

$$\lim_{n \to \infty} P\left(\sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \varepsilon \right) = 0 \quad \text{für alle } \varepsilon > 0.$$

Seien dabei $z_1, \ldots, z_n, \ldots$ gemäß $F_Z(z)$ verteilte Beobachtungen mit geeigneter Abhängigkeitsstruktur in den Daten.

Zum Nachweis des Kerntheorems unter Abhängigkeitsstrukturen sei erst die Notwendigkeit der Bedingung gezeigt. In einem zweiten Schritt folgt dann der Beweis, dass die einseitige gleichmäßige Konvergenz auch hinreichend ist. Der Nachweis, dass die einseitige gleichmäßige Konvergenz

$$\lim_{n\to\infty} P\left(\sup_{\alpha\in\Lambda}(R(\alpha) - R_{emp}(\alpha)) > \varepsilon\right) = 0, \quad \forall\varepsilon > 0$$

eine notwendige Bedingung für die nicht-triviale Konsistenz der ERM-Prinzips ist, ergibt sich aus folgender Aussage:
Sei die ERM-Methode strikt konsistent auf der Menge $Q(z,\alpha)$, $\alpha \in \Lambda$. Nach Definition der nicht-trivialen Konsistenz heißt das, dass für alle $c \in I\!R$, so dass die Menge $\Lambda(c) = \{\alpha | \int Q(z,\alpha) dF_Z(z) \geq c\}$ nicht leer ist, die folgende Konvergenz gilt:

$$\inf_{\alpha\in\Lambda(c)} R_{emp}(\alpha) \xrightarrow[n\to\infty]{P} \inf_{\alpha\in\Lambda(c)} R(\alpha).$$

Dies ist äquivalent zu den folgenden Konvergenzaussagen:

$$\inf_{\alpha\in\Lambda(c)} \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \xrightarrow{P} \inf_{\alpha\in\Lambda(c)} \int Q(z,\alpha) dF_Z(z), \quad n\to\infty$$

$$\iff \forall\varepsilon > 0 \quad \lim_{n\to\infty} P\left(\left|\inf_{\alpha\in\Lambda(c)} \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) - \inf_{\alpha\in\Lambda(c)} \int Q(z,\alpha) dF_Z(z)\right| > \varepsilon\right) = 0$$

$$\iff \forall\varepsilon > 0 \quad \lim_{n\to\infty} P\left(\left\{\inf_{\alpha\in\Lambda(c)} \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) - \inf_{\alpha\in\Lambda(c)} \int Q(z,\alpha) dF_Z(z) > \varepsilon\right\}\right.$$
$$\left. \vee \left\{\inf_{\alpha\in\Lambda(c)} \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) - \inf_{\alpha\in\Lambda(c)} \int Q(z,\alpha) dF_Z(z) < -\varepsilon\right\}\right) = 0$$

Betrachte eine nichtfallende endliche Folge reeller Zahlen $a_1 \leq a_2 \leq \ldots \leq a_l$ mit $a_1 = a$ und $a_l = A$, so dass $|a_{i+1} - a_i| < \frac{\varepsilon}{2}$. Definiere dann für $k = 1,\ldots,l$ das Ereignis

$$\begin{aligned} T_k &= \left\{\inf_{\alpha\in\Lambda(a_k)} \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) - \inf_{\alpha\in\Lambda(a_k)} \int Q(z,\alpha) dF_Z(z) < -\frac{\varepsilon}{2}\right\} \\ &= \left\{\inf_{\alpha\in\Lambda(a_k)} \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) < \inf_{\alpha\in\Lambda(a_k)} \int Q(z,\alpha) dF_Z(z) - \frac{\varepsilon}{2}\right\}. \end{aligned}$$

Dann gilt wegen der strikten Konvergenz des ERM-Prinzips

$$\lim_{n\to\infty} P(T_k) = 0 \quad \text{für alle } k = 1, \ldots, l$$

und somit gilt auch für $T = \bigcup_{k=1}^{l} T_k$ die Konvergenz

$$P(T) \underset{n\to\infty}{\longrightarrow} 0,$$

denn es gilt die Ungleichung:

$$\lim_{n\to\infty} P(T) \leq \sum_{k=1}^{l} \lim_{n\to\infty} P(T_k) = 0.$$

Wird nun das Ereignis

$$\mathcal{T} = \left\{ \sup_{\alpha\in\Lambda} \left(\int Q(z,\alpha) dF_Z(z) - \frac{1}{n} \sum_{i=1}^{n} Q(z_i,\alpha) \right) > \varepsilon \right\}$$

definiert, dann existiert unter der Annahme, dass $\mathcal{T}$ eintritt, ein $\alpha^* \in \Lambda$, so dass die Abschätzung

$$\int Q(z,\alpha^*) dF_Z(z) - \varepsilon > \frac{1}{n} \sum_{i=1}^{n} Q(z_i,\alpha^*)$$

gilt. Denn unter der Annahme, dass für alle $\alpha \in \Lambda$ die Ungleichung

$$\int Q(z,\alpha) dF_Z(z) - \frac{1}{n} \sum_{i=1}^{n} Q(z_i,\alpha) \leq \varepsilon$$

gilt, folgt

$$\sup_{\alpha\in\Lambda} \left(\int Q(z,\alpha) dF_Z(z) - \frac{1}{n} \sum_{i=1}^{n} Q(z_i,\alpha) \right) = \varepsilon.$$

Dies steht im Widerspruch zur Annahme, dass das Ereignis $\mathcal{T}$ eintritt.

Betrachte weiter obiges $\alpha^* \in \Lambda$, dann gibt es ein k, so dass $\alpha^* \in \Lambda(a_k)$ mit $a_k \in \{a_1, \ldots, a_l\}$ und es gilt die Ungleichung

$$\int Q(z,\alpha^*) dF_Z(z) - a_k < \frac{\varepsilon}{2}.$$

Dies ergibt sich aus der folgenden Überlegung:
Unter der Annahme, dass es ein $k^\star$ gibt, so dass $a_{k^\star} \geq a$ ist und die Abschätzung

$$\int Q(z, \alpha^*) dF_Z(z) \geq a_{k^\star}$$

gilt, folgt aus der Definition von $\Lambda(c)$, $c \in I\!R$, dass es ein $k^\star$ gibt, so dass $\alpha^* \in \Lambda(a_{k^\star})$ gilt. Falls aber andererseits für $k^\star$ die Abschätzung

$$\int Q(z, \alpha^*) dF_Z(z) - a_{k^\star} \geq \frac{\varepsilon}{2}$$

gilt, kann ein $k > k^\star$ gewählt werden, so dass gerade

$$\int Q(z, \alpha^*) dF_Z(z) - a_k < \frac{\varepsilon}{2}$$

gilt, aber auch noch

$$\int Q(z, \alpha^*) dF_Z(z) \geq a_k.$$

Ein solches k muss existieren, da $|a_k - a_{k-1}| < \frac{\varepsilon}{2}$ laut Definition der Folge $a_1, \ldots, a_n$ gilt.

Für den weiteren Beweis betrachte die Ungleichung

$$\inf_{\alpha \in \Lambda(a_k)} \int Q(z, \alpha) dF_Z(z) \geq a_k,$$

aus der die Abschätzung

$$\int Q(z, \alpha^*) dF_Z(z) - \inf_{\alpha \in \Lambda(a_k)} \int Q(z, \alpha) dF_Z(z) < \frac{\varepsilon}{2}$$

folgt, und damit auch die Ungleichungen:

$$\begin{aligned} \inf_{\alpha \in \Lambda(a_k)} \int Q(z, \alpha) dF_Z(z) - \frac{\varepsilon}{2} &> \int Q(z, \alpha^*) dF_Z(z) - \varepsilon \\ &> \frac{1}{n} \sum_{i=1}^{n} Q(z_i, \alpha^*) \\ &\geq \inf_{\alpha \in \Lambda(a_k)} \frac{1}{n} \sum_{i=1}^{n} Q(z_i, \alpha). \end{aligned}$$

Die zweite Ungleichung gilt wegen der strikten Konsistenz des ERM-Prinzips, die letzte wegen des Übergangs zum Infimum. Somit tritt das Ereignis T_k für alle $k = 1, \ldots, l$

ein und damit auch das Ereignis $T = \bigcup_{k=1}^{l} T_k$. Insgesamt impliziert also das Ereignis $\mathcal{T}$ das Ereignis T, so dass

$$P(\mathcal{T}) < P(T) \underset{n\to\infty}{\longrightarrow} 0$$

und damit

$$P\left(\sup_{\alpha\in\Lambda}\left(\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right) > \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0$$

gilt. Die gleichmäßige einseitige Konvergenz ist somit eine notwendige Bedingung für das Prinzip der empirischen Risiko-Minimierung.

Zum Nachweis, dass die einseitige gleichmäßige Konvergenz auch eine hinreichende Bedingung für die Konsistenz des ERM-Prinzips ist wird angenommen, dass gleichmäßige einseitige Konvergenz gilt. Es ist also die strikte Konsistenz des ERM-Prinzips, also

$$\lim_{n\to\infty} P\left(\left|\inf_{\alpha\in\Lambda(c)}\int Q(z,\alpha)dF_Z(z) - \inf_{\alpha\in\Lambda(c)}\frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right) = 0,$$

für beliebiges, festes $c \in I\!R$, zu zeigen. Betrachte dazu das Ereignis

$$\mathcal{Q} := \left\{z : \left|\inf_{\alpha\in\Lambda(c)}\int Q(z,\alpha)dF_Z(z) - \inf_{\alpha\in\Lambda(c)}\frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right\}$$

und die disjunkte Aufteilung $\mathcal{Q} = \mathcal{Q}_1 \cup \mathcal{Q}_2$ mit

$$\begin{aligned}\mathcal{Q}_1 &:= \left\{z : \inf_{\alpha\in\Lambda(c)}\int Q(z,\alpha)dF_Z(z) + \varepsilon < \inf_{\alpha\in\Lambda(c)}\frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right\}\\ \text{und}\quad \mathcal{Q}_2 &:= \left\{z : \inf_{\alpha\in\Lambda(c)}\int Q(z,\alpha)dF_Z(z) - \varepsilon > \inf_{\alpha\in\Lambda(c)}\frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right\}.\end{aligned}$$

Es reicht aus die Wahrscheinlichkeiten $P(\mathcal{Q}_1)$ und $P(\mathcal{Q}_2)$ abzuschätzen, da

$$P(\mathcal{Q}) = P(\mathcal{Q}_1) + P(\mathcal{Q}_1)$$

gilt. Gelte die Annahme, dass das Ereignis $\mathcal{Q}_1$ eintritt. Um eine obere Schranke für die Wahrscheinlichkeit $P(\mathcal{Q}_1)$ zu berechnen, betrachte eine Funktion $Q(z,\alpha^*)$ mit

$\alpha^* \in \Lambda(c)$, für die

$$\int Q(z,\alpha^*)dF_Z(z) < \inf_{\alpha\in\Lambda(c)} \int Q(z,\alpha)dF_Z(z) + \frac{\varepsilon}{2}$$

gilt. Daraus folgen die Ungleichungen

$$\begin{aligned}\int Q(z,\alpha^*)dF_Z(z) + \frac{\varepsilon}{2} &< \inf_{\alpha\in\Lambda(c)} \int Q(z,\alpha)dF_Z(z) + \varepsilon \\ &< \inf_{\alpha\in\Lambda(c)} \frac{1}{n}\sum_{i=1}^n Q(z_i,\alpha) \le \frac{1}{n}\sum_{i=1}^n Q(z_i,\alpha^*),\end{aligned}$$

so dass das Ereignis

$$\left\{ z : \frac{1}{n}\sum_{i=1}^n Q(z_i,\alpha^*) - \int Q(z,\alpha^*)dF_Z(z) > \frac{\varepsilon}{2} \right\}$$

erst recht eintritt, wenn $\mathcal{Q}_1$ eintritt. Damit gilt aber

$$P(\mathcal{Q}_1) \le P\left(\frac{1}{n}\sum_{i=1}^n Q(z_i,\alpha^*) - \int Q(z,\alpha^*)dF_Z(z) > \frac{\varepsilon}{2} \right).$$

Gilt nun ein schwaches Gesetz der großen Zahlen für das Funktional $Q(\cdot,\alpha^*)$, $\alpha^* \in \Lambda(c)$, d.h. gilt

$$\frac{1}{n}\sum_{i=1}^n Q(z_i,\alpha^*) \xrightarrow[n\to\infty]{Pr} \int Q(z,\alpha^*),$$

kann dieses auf die rechte Seite der Ungleichung angewandt werden, und es folgt die Konvergenz $P(\mathcal{Q}_1) \longrightarrow 0$ für $n \to \infty$.
Tritt andererseits das Ereignis $\mathcal{Q}_2$ ein, dann gibt es eine Funktion $Q(z,\alpha^{**})$, $\alpha^{**} \in \Lambda(c)$, so dass die Ungleichungen

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n Q(z_i,\alpha^{**}) + \frac{\varepsilon}{2} &< \inf_{\alpha\in\Lambda(c)} \frac{1}{n}\sum_{i=1}^n Q(z_i,\alpha) + \varepsilon \\ &< \inf_{\alpha\in\Lambda(c)} \int Q(z,\alpha)dF_Z(z) < \int Q(z,\alpha^{**})dF_Z(z)\end{aligned}$$

gelten, woraus die Abschätzungen

$$\begin{aligned} P(\mathcal{Q}_2) &< P\left(\int Q(z,\alpha^{**})dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha^{**}) > \frac{\varepsilon}{2}\right) \\ &< P\left(\sup_{\alpha\in\Lambda}\left(\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right) > \frac{\varepsilon}{2}\right) \end{aligned}$$

folgen. Dabei gilt die zweite Ungleichung wegen $\Lambda(c) \subset \Lambda$. Mit der gleichmäßigen Konvergenz

$$P\left(\sup_{\alpha\in\Lambda}\left(\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right) > \frac{\varepsilon}{2}\right) \underset{n\to\infty}{\longrightarrow} 0$$

gilt dann ebenfalls $P(\mathcal{Q}_2) \longrightarrow 0$ für $n \to \infty$ und insgesamt

$$P(\mathcal{Q}) \leq P(\mathcal{Q}_1) + P(\mathcal{Q}_2) \underset{n\to\infty}{\longrightarrow} 0.$$

Damit ist bewiesen, dass die einseitige gleichmäßige Konvergenz auch eine hinreichende Bedingung für die nicht-triviale Konsistenz des Prinzips der empirischen Risiko-Minimierung ist, wenn für die Funktionale $Q(z,\alpha)$, $\alpha \in \Lambda$, das schwache Gesetz der großen Zahlen

$$\frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \underset{n\to\infty}{\overset{Pr}{\longrightarrow}} \int Q(z,\alpha)dF_Z(z)$$

gilt.

Aus dem oben geführten Beweis geht also hervor, dass das Kerntheorem der statistischen Lerntheorie bei Abhängigkeiten in den Daten nur dann gilt, wenn zumindest ein schwaches Gesetz der großen Zahlen gilt, andernfalls würde die gleichmäßige einseitige Konvergenz zwar notwendige, aber nicht hinreichende Bedingung sein. Damit ergibt sich auch bei Abhängigkeitsstrukturen, dass jede Analyse des ERM-Prinzips eine Worst-Case-Analyse sein muss.

Im Folgenden werden sowohl für Martingal- als auch für L^r-Mixingal-Strukturen unter Ausnutzung der Ergebnisse im dritten Kapitel Kerntheoreme hergeleitet. Das schwache Gesetz der großen Zahlen aus Satz 3.20 für Martingal-Differenzen, kann direkt auf den

stochastischen Prozess

$$\left(\frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) - \int Q(z,\alpha)dF_Z(z)\right), \quad \alpha \in \Lambda,$$

angewandt werden, unter der Voraussetzung, dass

$$\frac{1}{n^2}\sum_{i=1}^{\infty} \mathrm{E}\left(\left(Q(z_i,\alpha) - \int Q(z,\alpha)dF_Z(z)\right)^2\right) \longrightarrow 0 \quad \text{für } n \to \infty$$

für $\alpha \in \Lambda$ gilt. Damit folgt direkt das Kerntheorem für Martingal-Strukturen in den Daten.

4.3 Satz *Kerntheorem für Martingal-Strukturen*
Seien $a \in I\!R$ und $A \in I\!R$ Konstanten, so dass für alle Funktionen aus der Menge $Q(z,\alpha)$, $\alpha \in \Lambda$, und für eine gegebene Verteilungsfunktion $F_Z(z)$ die Ungleichungen

$$a \leq \int Q(z,\alpha)dF_Z(z) \leq A, \quad \alpha \in \Lambda,$$

gelten. Seien weiter $z_1, \ldots, z_n, \ldots$ gemäß $F_Z(z)$ verteilte Beobachtungen mit Martingal-Struktur, d. h. die adaptierten Prozesse $\{Q(z_i,\alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, sind jeweils eine Folge von Martingal-Differenzen. Gelte weiter

$$\frac{1}{n^2}\sum_{i=1}^{\infty} \mathrm{E}\left(\left(Q(z_i,\alpha) - \int Q(z,\alpha)dF_Z(z)\right)^2\right) \longrightarrow 0 \quad \text{für } n \to \infty$$

für alle $\alpha \in \Lambda$. Dann sind folgende Aussagen äquivalent:

(i) Für die gegebene Verteilungsfunktion $F_Z(z)$ ist die Methode der empirischen Risiko-Minimierung strikt konsistent auf der Menge der Funktionen $Q(z,\alpha)$, $\alpha \in \Lambda$.

(ii) Für die gegebene Verteilungsfunktion gilt gleichmäßige einseitige Konvergenz des Mittels gegen den Erwartungswert auf der Menge der Funktionen $Q(z,\alpha)$, $\alpha \in \Lambda$:

$$\lim_{n\to\infty} P\left(\sup_{\alpha\in\Lambda}(R(\alpha) - R_{emp}(\alpha)) > \varepsilon\right) = 0, \quad \forall \varepsilon > 0.$$

Die zusätzliche Voraussetzung

$$\frac{1}{n^2}\sum_{i=1}^{\infty}\mathrm{E}\left(\left(Q(z_i,\alpha)-\int Q(z,\alpha)dF_Z(z)\right)^2\right)\longrightarrow 0 \quad \text{für } n\to\infty$$

in Satz 4.3 für die Gültigkeit des Kerntheorems unter Martingal-Strukturen ist genügend schwach, damit in der Regel die Modellierung der Fehlerterme durch einen geeigneten stochastischen Prozess möglich ist. In diesem Zusammenhang sollte noch darauf hingewiesen werden, dass die Voraussetzungen für jede Funktion aus der Menge $Q(z,\alpha)$, $\alpha\in\Lambda$, erfüllt sein muss, denn nur dann ist die generelle Anwendung der empirischen Risiko-Minimierung als ein Prinzip sicher gestellt.

Ebenso wie für Martingal-Strukturen kann aus den Ergebnissen im Abschnitt 3.4 über Mixingale ein schwaches Gesetz der großen Zahlen verwendet werden, das mit relativ wenigen zusätzlichen Voraussetzungen für die Anwendung auf das Kerntheorem für L^r-Mixingale geeignet ist. Der Satz 3.33 von Chen und White (1996) erlaubt die Anwendung eines schwachen Gesetzes der großen Zahlen, wenn der stochastische Prozess

$$\frac{1}{n}\sum_{i=1}^{n}Q(z_i,\alpha)-\int Q(z,\alpha)dF_Z(z),\quad \alpha\in\Lambda,$$

ein L^r-Mixingal ist und die Folge der zugehörigen Konstanten $\{c_i\}_1^\infty$ die Voraussetzung

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}c_i<\infty$$

erfüllt. Somit kann das Kerntheorem der statistischen Lerntheorie ebenfalls für L^r-Mixingale bewiesen werden.

4.4 Satz *Kerntheorem für L^r-Mixingal-Strukturen*
Seien $a\in\mathbb{R}$ und $A\in\mathbb{R}$ Konstanten, so dass für alle Funktionen aus der Menge $Q(z,\alpha)$, $\alpha\in\Lambda$, und für eine gegebene Verteilungsfunktion $F_Z(z)$ die Ungleichungen

$$a\le\int Q(z,\alpha)dF_Z(z)\le A,\quad \alpha\in\Lambda,$$

gelten. Seien weiter $z_1,\ldots,z_n,\ldots$ gemäß $F_Z(z)$ verteilte Beobachtungen mit L^r-Mixingal-Struktur, $r\ge 1$, d. h. die adaptierten Prozesse $\{Q(z_i,\alpha),\mathcal{A}_i\}_1^\infty$, $\alpha\in\Lambda$, sind

jeweils ein L^r-Mixingal, $r \geq 1$ mit zugehöriger Folge von Konstanten $\{c_i\}_1^\infty$, für die zusätzlich

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} c_i < \infty$$

gelte. Dann sind folgende Aussagen äquivalent:

(i) Für die gegebene Verteilungsfunktion $F_Z(z)$ ist die Methode der empirischen Risiko-Minimierung strikt konsistent auf der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$.

(ii) Für die gegebene Verteilungsfunktion gilt gleichmäßige einseitige Konvergenz des Mittels gegen den Erwartungswert auf der Menge der Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$:

$$\lim_{n\to\infty} P\left(\sup_{\alpha\in\Lambda}(R(\alpha) - R_{emp}(\alpha)) > \varepsilon\right) = 0, \quad \forall \varepsilon > 0.$$

Das Kerntheorem für Mixingal-Strukturen in den Daten ist zur Anwendung auf eine Vielzahl von stochastischen Prozessen zugelassen, denn die Voraussetzungen für die Mixingal-Konstanten c_i, sind je nach gewählter Norm $\|\cdot\|_r$ für das Mixingal einfach zu erfüllen und weitere Forderungen, wie beispielsweise an die Ordnung der Mixingal-Koeffizienten werden nicht gestellt. Werden die empirischen Verluste beispielsweise als homogener linearer Prozess wie in Beispiel 4.2 modelliert, kann der Satz 4.4 direkt angewandt werden.

4.5 Beispiel *L^2-Mixingal-Struktur: homogener linearer Prozess*
Betrachte ein statistisches Lernproblem für eine Menge von Funktionen $f(x, \alpha)$, $\alpha \in \Lambda$. Seien $z_1, \ldots, z_n, \ldots$ mit $z_i = (x_i, y_i)$ identisch gemäß $F_Z(z)$ verteilte Beobachtungen und sei $Q(z, \alpha)$ das Verlustfunktional, wobei die empirischen Verluste als homogener linearer Prozess $\{Q(z_i, \alpha)\}_1^\infty$ mit

$$Q(z_i, \alpha) = \sum_{k=0}^{\infty} a_k(\alpha) e_{i-k}$$

aus Beispiel 4.2 modelliert sind. Seien die adaptierten Prozess $\{Q(z_i, \alpha), \mathcal{A}_i\}_1^\infty$ jeweils L^2-Mixingale mit Konstanten

$$c_i = \sqrt{\mathrm{Var}(e_i)}, \quad i = 1, 2, \ldots,$$

und Mixingal-Koeffizienten

$$\psi_m(\alpha) = \left(\sum_{k=m}^{\infty} a_k^2(\alpha) \right)^{1/2}, \quad m = 0, 1, 2, \ldots.$$

Aus der Definition des homogenen linearen Prozess folgt, dass $\mathrm{Var}(e_i) < \infty$ gilt (vgl. Beispiel 4.2), wodurch die Voraussetzung

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} c_i = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sqrt{\mathrm{Var}(e_i)} < \infty$$

immer erfüllt ist. Somit kann für dieses statistische Lernproblem mit einer durch obigen homogenen linearen Prozess modellierten Abhängigkeitsstruktur das Kerntheorem der statistischen Lerntheorie postuliert werden.

Grundsätzlich kann das Kerntheorem sowohl für Martingal- als auch für Mixingal-Strukturen mit allen Gesetzen der großen Zahlen nachgewiesen werden. Allerdings müssen dann in der Regel deutlich mehr Voraussetzungen für die jeweiligen stochastischen Prozesse erfüllt sein, beispielsweise wenn das schwache Gesetz der großen Zahlen von Loève (Satz 3.19) für Martingale oder Satz 3.32 von de Jong (1995) für Mixingale genutzt wird. Ebenso können auch die jeweiligen starken Gesetze der großen Zahlen aus Kapitel 3 verwendet werden, da aus fast sicherer Konvergenz die Konvergenz in Wahrscheinlichkeit folgt.
Eine Ausweitung der Sätze 4.3 und 4.4, die für eine beliebige, aber feste Verteilung P_Z mit Verteilungsfunktion $F_Z(z)$ formuliert sind, auf eine Familie von Verteilungen $\mathcal{P}$ ist leicht möglich, so dass die beiden Kerntheoreme der statistischen Lerntheorie bei Martingal- bzw. Mixingal-Struktur genauso wie im Fall unabhängiger Beobachtungen (vgl. Korollar 2.12) für jedes Wahrscheinlichkeitsmaß aus der Familie $\mathcal{P}$ gelten.

Wie in Kapitel 2 für unabhängige Beobachtungen ist damit auch hier das Problem, die nicht-triviale Konsistenz des Prinzips der empirischen Risiko-Minimierung unter Abhängigkeitsstrukturen zu zeigen, durch die Äquivalenz der Aussagen zu dem Worst-Case-Problem geworden, gleichmäßige einseitige Konvergenz des empirischen Risikos gegen das Risiko zu zeigen. Im Gegensatz zu den Ergebnissen im zweiten Kapitel steht die Aufgabe, notwendige und hinreichende Bedingungen für die gleichmäßige einseitige

Konvergenz unter Ausnutzung der VC-Entropie H^{Λ} zu finden, wie dies für den Fall unabhängiger Beobachtungen im zweiten Kapitel mit Satz 2.15, Korollar 2.16 sowie Satz 2.17 hergeleitet worden ist, sowohl für beschränkte reelle Funktionen als auch für Indikatorfunktionen noch aus.
Mit den exponentiellen Schranken zur Kontrolle der Konvergenzrate, die im nächsten Abschnitt entwickelt werden, können aber zumindest hinreichende Bedingungen in Abhängigkeit von der annealed VC-Entropie H^{Λ}_{ann} angegeben werden. Dadurch ist sichergestellt, dass jedes Verfahren, dass das Prinzip der empirischen Risiko-Minimierung nutzt, das zugehörige Lernproblem löst. Allerdings kann bisher nicht gezeigt werden, dass nicht auch andere Bedingungen existieren, die ebenfalls die nicht-triviale Konsistenz des ERM-Prinzip gewährleisten, da die Suche nach einer notwendigen Bedingung eine noch offene Fragestellung ist.

4.3 Konvergenzrate des statistischen Lernprozesses bei Abhängigkeitsstrukturen

In diesem Abschnitt werden ähnlich wie im Abschnitt 2.4 exponentielle Schranken zur Kontrolle der Konvergenzrate der gleichmäßigen einseitigen Konvergenz des empirischen Risikos gegen das Risiko berechnet. In Zusammenhang mit der Herleitung solcher Schranken, werden auch Bedingungen angegeben, die sicherstellen, dass diese Schranken mit schneller Rate für $n \to \infty$ konvergieren, so dass damit diese Bedingungen auch für eine schnelle Konvergenzrate der empirischen Risiko-Minimierung hinreichend sind. Gleichzeitig folgt daraus, dass dieselben Bedingungen auch hinreichend für die gleichmäßige einseitige Konvergenz für beschränkte reelle Funktionen ebenso wie für Indikatorfunktionen sind.

Für die Herleitung der Schranken bei Abhängigkeitsstrukturen in den Daten werden im Weiteren die exponentiellen Abschätzungen aus Abschnitt 3.3 und 3.4 für Martingale und Mixingale ausgenutzt. Für die Herleitung der Schranken sei zuerst der Spezialfall einer endlichen Menge reeller, beschränkter Funktionen $Q(z, \alpha)$, $\alpha \in \Lambda$, mit $|Q(z, \alpha)| \leq B$ für alle $\alpha \in \Lambda$ und $B < \infty$ betrachtet. Für eine Menge von K Funktionen

mit $\Lambda = \{\alpha_1, \ldots, \alpha_K\}$ gelten auch für Daten mit Martingal- und Mixingal-Struktur, genau wie bei unabhängigen Beobachtungen, die folgenden Ungleichungen:

$$\begin{aligned} P\left(\sup_{1\le k\le K}\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right) \\ \le \sum_{k=1}^{K} P\left(\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right) \\ \le K\ P\left(\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right). \end{aligned}$$

Seien nun $z_1, \ldots, z_n$ gemäß $F_Z(z)$ verteilte Beobachtungen mit Martingal-Struktur, d. h. die adaptierten Prozesse $\{Q(z_i,\alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, sind jeweils eine Folge von Martingal-Differenzen. Dann gilt unter Ausnutzung der Ungleichung aus Korollar 3.24 für beliebiges $\alpha_k \in \Lambda$ die Abschätzung

$$P\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k) > \varepsilon\right) \le \exp\left\{\frac{-n\varepsilon^2}{2B^2}\right\}.$$

Insgesamt kann dann die Rate der gleichmäßigen einseitigen Konvergenz bei einer endlichen Menge von Funktionen und Daten mit Martingal-Struktur durch die Abschätzung

$$\begin{aligned} P\left(\sup_{1\le k\le K}\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right) \\ \le K\ \exp\left\{\frac{-n\varepsilon^2}{2B^2}\right\} \\ \le \exp\left\{\left(\frac{\ln K}{n} - \frac{\varepsilon^2}{2B^2}\right) n\right\} \end{aligned}$$

kontrolliert werden. Die Schranke ist wegen der Gedächtnislosigkeit der Martingale identisch zu der Schranke für eine finite Funktionenmenge im Fall unabhängiger Beobachtungen.

Für Beobachtungen $z_1, \ldots, z_n$ mit L^r-Mixingal-Struktur, also mit adaptierten Prozessen $\{Q(z_i,\alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, die jeweils ein L^r-Mixingal, $r \ge 2$, mit zugehöriger Folge von Konstanten $\{c_i\}_1^\infty$ und Folgen von Mixingal-Koeffizienten $\{\psi_m(\alpha)\}_0^\infty$ sind, so dass

$\sum_{m=0}^{\infty} \psi_m(\alpha) < \infty$ gilt, kann wegen Korollar 3.41 die Ungleichung

$$P\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k) > \varepsilon\right) \leq \frac{3}{2}\exp\left\{-\frac{n\varepsilon}{72\, r\left(\frac{r}{r-1}\right)^{3/2}\sum_{m=0}^{\infty}\psi_m(\alpha_k)\left(\sum_{t=1}^{n} c_t^2\right)^{1/2}}\right\}$$

erreicht werden. Damit folgt dann ähnlich wie für Martingal-Strukturen, dass die Abschätzung

$$\begin{aligned} P\left(\sup_{1\leq k\leq K}\left(\int Q(z,\alpha_k)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha_k)\right) > \varepsilon\right) \\ \leq K\,\frac{3}{2}\,\exp\left\{-\frac{n\varepsilon}{72\, r\left(\frac{r}{r-1}\right)^{3/2}\sum_{m=0}^{\infty}\sup_k\psi_m(\alpha_k)\left(\sum_{t=1}^{n} c_t^2\right)^{1/2}}\right\} \\ \leq \frac{3}{2}\,\exp\left\{\left(\frac{\ln K}{n} - \frac{\varepsilon}{72\, r\left(\frac{r}{r-1}\right)^{3/2}\sum_{m=0}^{\infty}\sup_k\psi_m(\alpha_k)\left(\sum_{t=1}^{n} c_t^2\right)^{1/2}}\right) n\right\} \end{aligned}$$

gilt, mit der die Konvergenzrate kontrolliert werden kann. Die Bedingung

$$\frac{\ln K}{n} \underset{n\to\infty}{\longrightarrow} 0$$

ist hinreichend dafür, dass sowohl die Schranken für Martingal-Strukturen als auch für Mixingal-Strukturen mit exponentieller Rate gegen Null konvergieren. Wegen der Finitheit der Funktionenmenge ist diese Bedingung allerdings immer erfüllt. Im Gegensatz dazu müssen für infinite Mengen mit den Konzepten der annealed VC-Entropie H_{ann}^{Λ} und der Growth-Funktion G^{Λ} aus Definition 2.18 exponentielle Schranken und hinreichende Bedingungen für deren exponentielle Konvergenz hergeleitet werden.

Im Folgenden werden exponentielle Schranken für Martingal- und Mixingal-Strukturen in den Daten nur für infinite Mengen von Indikatorfunktionen $Q(z,\alpha)$, $\alpha \in \Lambda$, mit

$Q(z,\alpha) \in \{0,1\}$ unter Ausnutzung der annealed VC-Entropie vorgestellt. Diese Schranken sind mit der gleichen Begründung, wie im Fall unabhängiger Beobachtungen auf Mengen von beschränkten, reellen Funktionen übertragbar, wobei genauso wie bei Unabhängigkeit das Entropie-Konzept genutzt werden kann, welches auf der annealed VC-Entropie aufbaut. Außerdem können die hier vorgestellten Schranken direkt genutzt werden, um durch Verwendung der Growth-Funktion anstatt der annealed VC-Entropie verteilungsfreie Schranken zu konstruieren.

Zur Herleitung der Schranken für eine infinite Menge von Indikatorfunktionen bei Beobachtungen $z_1, \ldots, z_n$ mit Abhängigkeitsstrukturen kann zuerst gezeigt werden, dass es für eine beliebige, feste Stichprobe $z_1, \ldots, z_n$ vom Umfang n und für die Menge von Indikatorfunktionen $Q(z,\alpha)$, $\alpha \in \Lambda$, mit $Q(z,\alpha) \in \{0,1\}$ und $|\Lambda| = \infty$ eine feste Anzahl $K = K^{\Lambda}(z_1, \ldots, z_n)$ von Vektoren $Q(z,\alpha^\star) \in \Lambda^\star = \Lambda^\star(z_1, \ldots, z_n) \subset \Lambda$ gibt, die das minimale ε-Netz bilden. Da die Stichprobe fest aus der Menge aller Stichproben $\mathcal{Z}(n)$ vom Umfang n gewählt wurde, ist das minimale ε-Netz und damit auch die Anzahl $K^{\Lambda}(z_1, \ldots, z_n)$ der Vektoren im Netz nicht stochastisch. Deshalb folgt mit der gleichen Begründung wie für finite Mengen von Funktionen allerdings für die bedingte Wahrscheinlichkeit, dass die feste Stichprobe $z_1, \ldots, z_n$ aus $\mathcal{Z}(n)$ realisiert wurde:

$$
\begin{aligned}
& P\left(\sup_{\alpha\in\Lambda} \left| \int Q(z,\alpha) dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \right| > \varepsilon \,\middle|\, z_1, \ldots, z_n \right) \\
&\quad = P\left(\sup_{\alpha^\star\in\Lambda^\star} \left| \int Q(z,\alpha) dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha^\star) \right| > \varepsilon \,\middle|\, z_1, \ldots, z_n \right) \\
&\quad \leq \sum_{\alpha^\star\in\Lambda^\star} P\left(\left| \int Q(z,\alpha) dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha^\star) \right| > \varepsilon \,\middle|\, z_1, \ldots, z_n \right) \\
&\quad = K^{\Lambda}(z_1, \ldots, z_n) P\left(\left| \int Q(z,\alpha) dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \right| > \varepsilon \,\middle|\, z_1, \ldots, z_n \right)
\end{aligned}
$$

Um Abschätzungen für beliebige Stichproben vom Umfang n zu bekommen, ist es ausreichend, zum Erwartungswert bezüglich des Wahrscheinlichkeitsmaßes auf der Menge

der n-elementigen Stichproben $\mathcal{Z}(n)$ überzugehen:

$$
\begin{aligned}
& P\left(\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right) \\
&= E\left(P\left(\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon \,\middle|\, z_1,\ldots,z_n\right)\right) \\
&< E\left(K^{\Lambda}(z_1,\ldots,z_n)\right) E\left(P\left(\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon \,\middle|\, z_1,\ldots,z_n\right)\right) \\
&< 2\,\exp\left\{H^{\Lambda}_{ann}(n)\right\} E\left(P\left(\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon \,\middle|\, z_1,\ldots,z_n\right)\right),
\end{aligned}
$$

wegen der Definition der annealed VC-Entropie für Indikatorfunktionen:

$$H^{\Lambda}_{ann}(n) = \ln \mathrm{E}(K(z_1,\ldots,z_n)).$$

Damit die Rate der gleichmäßigen zweiseitigen Konvergenz durch eine exponentielle Schranke kontrolliert werden kann, muss damit nur noch die Wahrscheinlichkeit auf der rechten Seite der Ungleichung in Bezug auf die jeweilige Abhängigkeitsstruktur durch eine exponentielle Schranke abgeschätzt werden.
Für eine infinite Menge von Indikatorfunktionen und für Beobachtungen mit Martingal-Struktur kann dann der folgende Satz bewiesen werden.

4.6 Satz

Sei $Q(z,\alpha)$, $\alpha\in\Lambda$, eine Menge von Indikatorfunktionen mit $Q(z,\alpha)\in\{0,1\}$ und seien $z_1,\ldots,z_n$ identisch verteilte Beobachtungen gemäß der Verteilung $F_Z(z)$ mit Martingal-Struktur, d. h. die adaptierten Prozesse $\{Q(z_i,\alpha),\mathcal{A}_i\}_1^{\infty}$, $\alpha\in\Lambda$, sind jeweils eine Folge von Martingal-Differenzen. Dann gilt folgende Ungleichung:

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right\} < 2\,\exp\left\{\left(\frac{H^{\Lambda}_{ann}(n)}{n} - \frac{\varepsilon^2}{2}\right)n\right\}.$$

Hinreichend für die Nicht-Trivialität dieser exponentiellen Schranke ist die Bedingung

$$\frac{H^{\Lambda}_{ann}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

Beweis:

Für die Folgen von Martingal-Differenzen $\{Q(z_i, \alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$ kann Korollar 3.24 angewandt werden, denn Indikatorfunktionen sind durch $B = 1$ beschränkt. Damit folgt die Abschätzung

$$P\left(\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon \,\middle|\, z_1,\ldots,z_n\right) \leq 2\ \exp\left\{\frac{\varepsilon^2 n}{2}\right\}.$$

Beim Übergang zum Supremum folgt wegen der oben bewiesenen Abschätzung:

$$\begin{aligned} P\left(\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right) & \\ < \ 2\ \exp\left\{H^\Lambda_{ann}(n)\right\} E\left(\exp\left\{\frac{\varepsilon^2 n}{2}\right\}\right) & \\ = \ 2\ \exp\left\{H^\Lambda_{ann}(n)\right\} \exp\left\{\frac{\varepsilon^2 n}{2}\right\} & \\ = \ 2\ \exp\left\{\left(\frac{H^\Lambda_{ann}(n)}{n} - \frac{\varepsilon^2}{2}\right) n\right\} & \end{aligned}$$

Aus der Abschätzung folgt direkt, dass die Bedingung

$$\frac{H^\Lambda_{ann}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0$$

hinreichend für die Nicht-Trivialität der Schranke ist.

□

Ebenso wie für Daten mit Martingalstruktur gibt es auch für Beobachtungen mit Mixingal-Struktur geeignete exponentielle Schranken, die eine Kontrolle der Rate der zweiseitigen gleichmäßigen Konvergenz erlauben. Der nächste Satz gibt eine solche Schranke mit nur schwachen Voraussetzungen an die Mixingale $\{Q(z_i, \alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, an. Allerdings gilt die Aussage nur für L^r-Mixingale mit $r \geq 2$ und die exponentielle Abschätzung ist von einer etwas schwächeren Ordnung als die Schranke für Martingale, da der Abstand ε zwischen Mittel und Erwartungswert nicht quadratisch eingeht (vgl. die Korollare 3.24 und 3.41).

4.7 Satz

Sei $Q(z,\alpha)$, $\alpha \in \Lambda$, eine Menge von Indikatorfunktionen mit $Q(z,\alpha) \in \{0,1\}$ und seien $z_1,\ldots,z_n$ identisch verteilte Beobachtungen gemäß der Verteilung $F_Z(z)$ mit L^r-Mixingal-Struktur, $r \geq 2$, d. h. die adaptierten Prozessen $\{Q(z_i,\alpha),\mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, sind jeweils ein L^r-Mixingal, $r \geq 2$, mit zugehöriger Folge von Konstanten $\{c_i\}_1^\infty$ und Folgen von Mixingal-Koeffizienten $\{\psi_m(\alpha)\}_0^\infty$, so dass $\sum\limits_{m=0}^{\infty} \psi_m(\alpha) < \infty$ erfüllt ist. Dann gelten folgende Ungleichungen:

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right\}$$

$$< \;\; 3\;\exp\left\{\left(\frac{H^\Lambda_{ann}(n)}{n} - \frac{\varepsilon}{72\;r\left(\frac{r}{r-1}\right)^{3/2}\sum\limits_{k=0}^{\infty}\sup\limits_{\alpha\in\Lambda}\psi_k(\alpha)\left(\sum\limits_{t=1}^{n}c_t^2\right)^{1/2}}\right)n\right\}.$$

Hinreichend für die Nicht-Trivialität dieser exponentiellen Schranken ist die Bedingung

$$\frac{H^\Lambda_{ann}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

Beweis:

Wegen der Beschränktheit der Indikatorfunktionen kann für die Folgen von Mixingal-Differenzen $\{Q(z_i,\alpha),\mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$ das Korollar 3.41 angewandt werden. Damit folgt für alle $\alpha \in \Lambda$ die Abschätzung:

$$P\left(\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon \,\middle|\, z_1,\ldots,z_n\right)$$

$$\leq \;\; 3\;\exp\left\{-\frac{\varepsilon\, n}{72\;r\left(\frac{r}{r-1}\right)^{3/2}\sum\limits_{k=0}^{\infty}\psi_k(\alpha)\left(\sum\limits_{t=1}^{n}c_t^2\right)^{1/2}}\right\}$$

$$\leq \;\; 3\;\exp\left\{-\frac{\varepsilon\, n}{72\;r\left(\frac{r}{r-1}\right)^{3/2}\sum\limits_{k=0}^{\infty}\sup\limits_{\alpha\in\Lambda}\psi_k(\alpha)\left(\sum\limits_{t=1}^{n}c_t^2\right)^{1/2}}\right\}.$$

Mit dieser Schranke folgt der Übergang zum Supremum analog zum Beweis in Satz 4.6. Aus der Darstellung der Abschätzung folgt sofort, dass die Bedingung

$$\frac{H^{\Lambda}_{ann}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

hinreichend für die Nicht-Trivialität der Schranke ist.

□

Neben der Schranke in Satz 4.7 existiert eine weitere Schranke, die sogar für beliebige L^r-Mixingal-Strukturen, $r \geq 1$, gilt. Zusätzlich ist diese Abschätzung enger als die in Satz 4.7 und konvergiert damit schneller für $n \to \infty$. Diese exponentielle Schranke, die im Folgenden Satz vorgestellt wird, gilt ebenfalls bei schwachen Forderungen an die Mixingal-Eigenschaften.

4.8 Satz

Sei $Q(z,\alpha)$, $\alpha \in \Lambda$, eine Menge von Indikatorfunktionen mit $Q(z,\alpha) \in \{0,1\}$ und seien $z_1,\ldots,z_n$ identisch verteilte Beobachtungen gemäß der Verteilung $F_Z(z)$ mit L^r-Martingal-Struktur, $r \geq 1$, d. h. die adaptierten Prozessen $\{Q(z_i,\alpha),\mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, sind jeweils ein L^r-Mixingal, $r \geq 1$, mit zugehöriger Folge von Konstanten $\{c_i\}_1^\infty$ und einer Folge von Mixingal-Koeffizienten $\{\psi_m(\alpha)\}_0^\infty$, so dass

$$(\psi_0(\alpha)+\psi_1(\alpha))\, c_{\sup} \geq \exp\{1\}-1, \quad \alpha \in \Lambda$$

gilt. Sei weiter $\psi_0 = \sup\limits_{\alpha\in\Lambda} \psi_0(\alpha)$ und $\psi_1 = \sup\limits_{\alpha\in\Lambda} \psi_1(\alpha)$, dann gilt die folgende Ungleichung:

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF_z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon\right\}$$

$$< \; 6\, \exp\left\{\left(\frac{H^{\Lambda}_{ann}(n)}{n} - \frac{\varepsilon^2}{2}\frac{\left(\exp\{1/n-1\}-1\right)^2}{(\psi_0+\psi_1)^2\, c^2_{\sup}}\right) n\right\}.$$

Hinreichend für die Nicht-Trivialität dieser exponentiellen Schranke ist die Bedingung

$$\frac{H^{\Lambda}_{ann}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

Beweis:
Wegen der Beschränktheit der Indikatorfunktionen gilt $|Q(z,\alpha)| < 1$ für alle α, so dass für die Mixingale $\{Q(z_i,\alpha), \mathcal{A}_i\}_1^\infty$, $\alpha \in \Lambda$, der Satz 3.45 angewandt werden kann, falls ein b_{n_0} existiert (vgl. den Beweis von Satz 3.45), so dass

$$|Q(z,\alpha)| < 1 \leq \frac{1}{\exp\{b_{n_0} - 1\} - 1}(\psi_0(\alpha) + \psi_1(\alpha))\ c_{\sup},$$

wobei die rechte Ungleichung äquivalent zu

$$b_{n_0} - 1 \leq \ln[(\psi_0(\alpha) + \psi_1(\alpha))\ c_{\sup} + 1]$$

ist. Setze $b_{n_0} = 1/n_0$ (vgl. den Beweis von Satz 3.45), dann folgt

$$b_{n_0} = 1/n_0 \leq \ln[(\psi_0(\alpha) + \psi_1(\alpha))\ c_{\sup} + 1] + 1$$

und damit die Abschätzung

$$n_0 \geq \left(\ln[(\psi_0(\alpha) + \psi_1(\alpha))\ c_{\sup} + 1] + 1\right)^{-1}.$$

Mit der Voraussetzung

$$(\psi_0(\alpha) + \psi_1(\alpha))c_{\sup} \geq \exp\{1\} - 1$$

an die Mixingal-Eigenschaft gilt $n_0 \geq 2$, so dass die Aussage des Satzes 3.45 für alle $\alpha \in \Lambda$ nicht von der Wahl eines n_0 abhängt. Mit der Anwendung von Satz 3.45 gilt dann die Abschätzung

$$\begin{aligned} P\left(\left|\int Q(z,\alpha)dF_Z(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \varepsilon \,\middle|\, z_1,\ldots,z_n\right) & \\ \leq\ 6\ \exp\left\{-\frac{n\varepsilon^2}{2}\frac{\left(\exp\{1/n-1\}-1\right)^2}{(\psi_0(\alpha)+\psi_1(\alpha))^2\ c_{\sup}^2}\right\} & \\ \leq\ 6\ \exp\left\{-\frac{n\varepsilon^2}{2}\frac{\left(\exp\{1/n-1\}-1\right)^2}{(\psi_0+\psi_1)^2\ c_{\sup}^2}\right\} & \end{aligned}$$

für alle $\alpha \in \Lambda$. Mit dieser Schranke folgt der Übergang zum Supremum analog zum Beweis in Satz 4.6 bzw. 4.7. Aus der Abschätzung folgt sofort, dass für die Nicht-Trivialität der Schranke die folgende Bedingung hinreichend ist:

$$\frac{H_{ann}^{\Lambda}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

□

Die exponentiellen Schranken, die in Satz 4.6 für Martingal-Strukturen sowie in den Sätzen 4.7 und 4.8 für Mixingal-Strukturen entwickelt worden sind, werden im Folgenden zusammen mit den hinreichenden Bedingungen dafür genutzt, die Generalisierungsfähigkeit des ERM-Prinzips in Abhängigkeit der annealed VC-Entropie zu zeigen. Wie in Abschnitt 2.4 werden dazu Abschätzungen für den Wert des Risikos $R(\widetilde{\alpha}_n)$ für die durch die Minimierung des empirischen Risikos gewählte Funktion $Q(z, \widetilde{\alpha}_n)$ angegeben. Mit Wahrscheinlichkeit $(1-\eta)$, $0 < \eta \leq 1$, gilt:

$$R(\widetilde{\alpha}_n) \leq R_{emp}(\widetilde{\alpha}_n) + \sqrt{\mathcal{E}(n, \eta, H^{\Lambda}_{ann}, C)}\ ,$$

wobei die Werte für $\mathcal{E}$ im Gegensatz zum Fall unabhängiger Beobachtungen zusätzlich von den jeweiligen Abhängigkeitsstrukturen abhängen, was durch den Wert C in die Schranke einfließt.

Für *Martingal-Strukturen* gilt unter den Voraussetzungen aus Satz 4.6 für $0 < \eta \leq 1$ dann mit Wahrscheinlichkeit $(1-\eta)$ die Ungleichung

$$R(\widetilde{\alpha}_n) \leq R_{emp}(\widetilde{\alpha}_n) + \sqrt{2\frac{H^{\Lambda}_{ann}(n) - \ln(\eta/2)}{n}}.$$

Damit kann $\mathcal{E}$ durch

$$\mathcal{E} = \mathcal{E}(n, \eta, H^{\Lambda}_{ann}, C_1) = \frac{H^{\Lambda}_{ann}(n) - \ln(\eta/2)}{n\, C_1}$$

angegeben werden, wobei $C_1 = 1/2$ ist.

Für *Mixingal-Strukturen* mit den Voraussetzungen aus Satz 4.8 folgt für $0 < \eta \leq 1$ mit Wahrscheinlichkeit $(1-\eta)$ die Ungleichung

$$R(\widetilde{\alpha}_n) \leq R_{emp}(\widetilde{\alpha}_n) + \sqrt{\frac{H^{\Lambda}_{ann}(n) - \ln(\eta/6)}{nC_2}}\ ,$$

mit

$$C_2 = \frac{1}{2}\frac{\left(\exp\left\{1/n - 1\right\} - 1\right)^2}{(\psi_0 + \psi_1)^2\ c^2_{\sup}}\ .$$

Mit Satz 4.7 ergibt sich noch eine weitere Schranke für $R(\widetilde{\alpha}_n)$ bei Mixingal-Strukturen in den Daten, allerdings wird dann nur eine deutlich größere Schranke erreicht, da für die Abschätzung in diesem Satz der Abstand ε nicht quadriert auftritt. Die Angabe für diese Schranke ergibt sich analog.

Die zweite Abschätzung, die das Risiko begrenzen soll und deshalb für die Generalisierungsfähigkeit des ERM-Prinzips wichtig ist, gibt eine Schranke für die Differenz des geschätzten Risikos $R(\widetilde{\alpha}_n)$ zum minimalen Risiko $\inf_{\alpha\in\Lambda} R(\alpha)$ an. Unter Abhängigkeitsstrukturen gelten wegen Korollar 3.24 bzw. Satz 3.45 aus dem dritten Kapitel mit Wahrscheinlichkeit $(1-\eta)$, $0<\eta\leq 1$, Abschätzungen der Art

$$\inf_{\alpha\in\Lambda} R(\alpha) \geq R(\widetilde{\alpha}_n) \geq R_{emp}(\widetilde{\alpha}_n) - \frac{1}{C}\sqrt{\frac{-\ln(\eta^*(\eta))}{2n}},$$

wobei C und η^* von der jeweiligen Abhängigkeitsstruktur, also von den Schranken in Korollar 3.24 bzw. Satz 3.45, abhängen. Zusammen mit der obigen Abschätzung für das Risiko $R(\widetilde{\alpha}_n)$ ergibt sich ähnlich wie für den Fall unabhängiger Beobachtungen (vgl. die Aussagen zu Satz 2.21) mit Wahrscheinlichkeit $(1-2\eta)$ eine Abschätzung vom Typ:

$$R(\widetilde{\alpha}_n) - \inf_{\alpha\in\Lambda} R(\alpha) \leq \sqrt{\mathcal{E}(n,\eta,H^{\Lambda}_{ann},C)} - \frac{1}{C}\sqrt{\frac{-\ln(\eta^*(\eta))}{2n}}\,.$$

Für *Martingal-Strukturen* gilt unter den Voraussetzungen aus Satz 4.6 mit Wahrscheinlichkeit $(1-2\eta)$ die Ungleichung

$$R(\widetilde{\alpha}_n) - \inf_{\alpha\in\Lambda} R(\alpha) \leq \sqrt{\mathcal{E}(n,\eta,H^{\Lambda}_{ann},C_1)} - \frac{1}{C_1}\sqrt{\frac{-\ln(\eta^*(\eta))}{2n}}$$

mit $C_1 = 1/2$ und $\eta^*(\eta) = \eta$, so dass sich die Abschätzung

$$R(\widetilde{\alpha}_n) - \inf_{\alpha\in\Lambda} R(\alpha) \leq \sqrt{2\frac{H^{\Lambda}_{ann}(n) - \ln(\eta/2)}{n}} - 2\sqrt{\frac{-\ln(\eta)}{2n}}$$

ergibt. Die beiden Abschätzungen des Risikowertes $R(\widetilde{\alpha}_n)$ für das durch Minimierung des empirischen Risikos gewählte $\widetilde{\alpha}_n$ zusammen mit der Tatsache, dass die Bedingung

$$\frac{H^{\Lambda}_{ann}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0$$

mit Satz 4.6 auch hinreichend für die einseitige gleichmäßige Konvergenz des empirischen Risikos gegen das Risiko ist, ergeben insgesamt bei nur schwachen Voraussetzungen die nicht-triviale Konsistenz des Prinzips der empirischen Risiko-Minimierung bei Daten mit Martingal-Strukturen für infinite Mengen von Indikatorfunktionen $Q(z,\alpha)$, $\alpha\in\Lambda$. Die Voraussetzungen ergeben sich aus dem Kerntheorem für Mixingal-Strukturen und aus Satz 4.6.

Für *L^r-Mixingal-Strukturen*, $r \geq 1$, gilt mit den Voraussetzungen aus Satz 4.8 mit Wahrscheinlichkeit $(1-2\eta)$ die Abschätzung

$$R(\widetilde{\alpha}_n) - \inf_{\alpha\in\Lambda} R(\alpha) \leq \sqrt{\mathcal{E}(n,\eta,H^{\Lambda}_{ann},C_2)} - \frac{1}{C_2}\sqrt{\frac{-\ln(\eta^*(\eta))}{2n}}$$

mit $\eta^*(\eta) = \eta/3$ und damit die Abschätzung

$$R(\widetilde{\alpha}_n) - \inf_{\alpha\in\Lambda} R(\alpha) \leq \sqrt{\frac{H^{\Lambda}_{ann}(n) - \ln(\eta/6)}{nC_2}} - \frac{1}{C_2}\sqrt{\frac{-\ln(\eta/3)}{2n}}\,,$$

wobei für C_2 gilt:

$$C_2 = \frac{1}{2}\frac{\left(\exp\{1/n-1\}-1\right)^2}{(\psi_0+\psi_1)^2\, c^2_{\sup}}\,.$$

Analog zu Martingal-Strukturen in den Beobachtungen ergibt sich auch für L^r-Mixingal-Strukturen dadurch die nicht-triviale Konsistenz für das ERM-Prinzip auf der Menge der Indikatorfunktionen $Q(z,\alpha)$, $\alpha \in \Lambda$ unter schwachen Voraussetzungen für die Mixingal-Struktur. Auch hier ist die hinreichende Bedingung für die einseitige gleichmäßige Konvergenz des empirischen Risikos gegen das Risiko durch

$$\frac{H^{\Lambda}_{ann}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0$$

gegeben.

Eine verteilungsfreie Analyse mit Hilfe von Abschätzungen unter Ausnutzung der Growth-Funktion kann wegen der Gültigkeit der Abschätzung

$$H^{\Lambda}_{ann}(n) \leq G^{\Lambda}(n)$$

durch einfaches Einsetzen von $G^{\Lambda}(n)$ statt der annealed VC-Entropie $H^{\Lambda}_{ann}(n)$ in die obigen Ergebnisse erfolgen. Durch den Einsatz der VC-Dimension können dann auch konstruktive und verteilungsfreie Schranken angegeben werden (vgl. Abschnitt 2.4 und Vapnik, 1998). Damit ist die Generalisierungsfähigkeit des Prinzips der empirischen Risiko-Minimierung auch für Martingal- und Mixingal-Strukturen in den Beobachtungen nachgewiesen.

Wie zu Beginn dieses Abschnitts bereits erläutert wurde, lassen sich die hier entwickelten Schranken auch auf beschränkte, reelle Funktionen anwenden, wobei dazu

ein weiteres Entropie-Konzept eingeführt werden muss, das im zweiten Kapitel bereits kurz erläutert wurde. Zusammen mit den Kerntheoremen und den Sätzen zur Konvergenzrate aus diesem Kapitel kann so aber auch die Generalisierungsfähigkeit des ERM-Prinzips für beschränkte, reelle Funktionen nachgewiesen werden.

Hier sei nochmals darauf hingewiesen, dass die Bedingungen für die Generalisierungsfähigkeit im Gegensatz zum Fall unabhängiger Beobachtungen nur hinreichend sind. Der Nachweis einer notwendigen Bedingung oder der Beweis, dass die Bedingung

$$\frac{G^{\Lambda}(n)}{n} \underset{n\to\infty}{\longrightarrow} 0$$

auch notwendig für die gleichmäßige einseitige Konvergenz ist, bleibt eine offene Frage für das theoretische Konzept des statistischen Lernens bei Abhängigkeitsstrukturen.

5 Ausblick

Die statistische Lerntheorie, die auf den seit 1960 entwickelten Ergebnissen von Vapnik und Chervonenkis aufbaut, wurde maßgeblich von Vapnik (1995) formuliert. Grundlegende Idee dieser Theorie ist die Nutzung der empirischen Risiko-Minimierung an Stelle der Minimierung des wahren Risikos. Für eine generelle Anwendung dieses Vorgehens als ein Prinzip muss gewährleistet sein, dass sich das Verfahren der empirischen Risiko-Minimierung konsistent verhält, also unter einem geeigneten allgemeingültigen Konvergenz-Konzept gute Konsistenzeigenschaften besitzt, so dass das wahre durch das empirische Risiko ersetzt werden kann. Im zweiten Kapitel wurde die Theorie, die zur Generalisierungsfähigkeit des ERM-Prinzips führt, vorgestellt. Dieses von Vapnik (1995, 1998) entwickelte einheitliche Konzept gilt allerdings nur für unabhängige Beobachtungen.
Unter Ausnutzung der strukturellen Risiko-Minimierung hat Vapnik (1998), aufbauend auf dem grundlegenden theoretischen Prinzip der empirischen Risiko-Minimierung, auch den prinzipiellen Aufbau des Algorithmus vorgestellt, der eine schnelle und effiziente Möglichkeit bietet, Zusammenhänge in den Daten zu erkennen. Eine große Klasse solcher Algorithmen sind die Support Vector Machines (SVM), die auch für große und vor allem hochdimensionale Datensätze in den Bereichen Mustererkennung (vgl. z. B. Heisele et al., 2003, und Belousov, Verzakov und von Frese, 2002) aber auch bei Regressionmodellen (z. B. Ma, Theiler und Perkins, 2003, und Mangasarian und Musicant, 2000) gute Ergebnisse liefern.

Es existieren zahlreiche Arbeiten zur Anwendbarkeit und Verbesserung der SVM-Algorithmen (z. B. Joachims, 1998, und Vapnik und Chapelle, 2000) und auch theoretisch motivierte Arbeiten im Bereich der statistischen Lerntheorie (z. B. Chen, Wang und Dong, 2003, und Schölkopf, 2003). Jedoch befassen sich diese Arbeiten grundsätzlich nur mit der Anwendung der empirischen Risiko-Minimierung oder der Support Vector Machine auf unabhängige Beobachtungen (bzw. Trainingsbeispiele). Für die empirische Anwendung von SVM-Algorithmen auf zeitlich-dynamische Daten gibt es allerdings bereits seit längerer Zeit Beispiele in der Literatur (Mukherjee, Osuna und Girosi, 1997, und Müller et al., 1997) mit teilweise exzellenten Ergebnissen. Aber auch

jüngste Arbeiten zur Anwendung von Algorithmen, die auf der Idee der Support Vector Machine aufbauen (z. B. Cao, 2003, oder Cao und Tay, 2001), beschreiben zwar Anwendungen auf zeitabhängige Daten, ohne allerdings auf die Abhängigkeitsstrukturen in den Daten im Zusammenhang mit der dann in Frage gestellten Gültigkeit des ERM-Prinzips einzugehen.

In dieser Arbeit wird, ausgehend von der Idee, die Allgemeingültigkeit des Prinzips der empirischen Risiko-Minimierung möglichst weitgehend zu erhalten, ein geeignetes Konzept zur Modellierung von Abhängigkeitsstrukturen in Daten entwickelt. Dabei wird davon ausgegangen, dass für die empirischen Verluste an Stelle der Daten eine konkrete Abhängigkeitsstruktur angenommen werden kann. Dadurch sind die Beobachtungen zwar ebenfalls abhängig modelliert, allerdings kann mit diesem Konzept in der Regel die konkrete Struktur in der Stichprobe nicht spezifiziert werden. Da aber die hier betrachteten Theorien zur Abhängigkeit sehr allgemeingültig sind und mit nur wenigen Voraussetzungen bezüglich der Strukturen in den empirischen Verlusten auskommen, kann davon ausgegangen werden, dass für eine Vielzahl verschiedener Abhängigkeitsstrukturen in den Beobachtungen ein geeigneter stochastischer Prozess zur Modellierung der Verluste angepasst werden kann. Zur Modellierung als zeitabhängigen stochastischen Prozess kann vor allem das von McLeish (1975, 1977) entwickelte Konzept der Mixingale herangezogen werden. Durch dieses Konzept wird eine große Klasse von dynamischen Prozessen, wie unter anderen die MA-, AR- und ARMA-Prozesse, abgedeckt (Andrews, 1988).

Die Verknüpfung der Mixingal- und auch der Martingal-Theorie mit der Theorie zum Prinzip der empirischen Risiko-Minimierung basiert vor allem auf der Tatsache, dass zum Nachweis der Generalisierungsfähigkeit des ERM-Prinzips schwache Gesetze der großen Zahlen benötigt werden. Dabei haben sich die in Abschnitt 3.4 vorgestellten Ergebnisse für Mixingale als genauso tragfähig erwiesen wie die Ergebnisse für Martingale, obwohl bei Mixingalen nicht die Gedächtnislosigkeit des Prozesses gefordert werden muss. Mit der Annahme einer Mixingal- oder auch Martingal-Strukturen als Abhängigkeitsstruktur in den empirischen Verlusten kann das von Vapnik postulierte Kerntheorem der statistischen Lerntheorie verallgemeinert werden, so dass die

gleichmäßige einseitige Konvergenz des empirischen Risikos gegen das wahre Risiko weiterhin hinreichend und notwendig für die Konsistenz des ERM-Prinzips ist. Dabei müssen sowohl für Martingal-Strukturen als auch für Mixingal-Strukturen in den empirischen Verlusten nur schwache, in der Regel einfach erfüllbare, zusätzliche Voraussetzungen angenommen werden.

Der zweite Schritt, um für das Prinzip der empirischen Risiko-Minimierung Konsistenz nachzuweisen, ist die Bestimmung von Bedingungen für die gleichmäßige einseitige Konvergenz eines stochastischen Prozesses. Im Gegensatz zum Fall der Unabhängigkeit in den Daten können für dynamische Datenstrukturen, wie sie in dieser Arbeit betrachtet werden, keine hinreichenden und notwendigen, sondern nur hinreichende Bedingungen angegeben werden. Somit ist nicht ausgeschlossen, dass noch weitere hinreichende Bedingungen mit geringeren Anforderungen existieren. Sowohl im Fall der Unabhängigkeit als auch bei den betrachteten Abhängigkeitsstrukturen hängen diese Bedingungen hauptsächlich von der Mannigfaltigkeit der Menge von Funktionen, die für das Lernproblem zugelassen sind, und von der vorgegebenen Verlustfunktion ab. Die verschiedenen, dazu von Vapnik und Chervonenkis entwickelten Entropien, wie annealed VC-Entropie, Growth-Funktion bzw. VC-Dimension, sind so konzipiert, dass die Komplexität der Funktionenmenge in Bezug zu den beobachteten Daten bewertet wird (Vapnik, 1993). Die Abhängigkeitsstruktur in den empirischen Verlusten wird von diesen Entropien nicht zusätzlich berücksichtigt, um die Funktionenmenge zu charakterisieren. Mit dieser Argumentation ist durch die Nutzung anderer Entropie-Konzepte, welche die Abhängigkeitsstrukturen in den Daten berücksichtigen, eine Abschwächung der Bedingungen für die gleichmäßige einseitige Konvergenz denkbar.

Ein wichtiger Punkt, um das ERM-Prinzip für die statistische Lerntheorie auch praktisch nutzbar zu machen, ist die Kontrolle der Konvergenzrate des empirischen Risikos gegen das wahre Risiko, denn nur bei schneller Konvergenz kann erwartet werden, dass Algorithmen entwickelt werden können, die für eine praktische Anwendung effizient sind. In dieser Arbeit sind dazu exponentielle Schranken für Mixingale und auch Martingale entwickelt worden, die die schnelle Konvergenz stochastischer Prozesse sicherstellen. Die Anwendung dieser Schranken im Kontext der empirischen Risiko-

Minimierung wird in dieser Arbeit ausführlich für Indikatorfunktionen, also insbesondere für die Klassifikation, gezeigt. Die entwickelten Schranken in Abhängigkeit von der zu Grunde liegenden Entropie sind größer als die Schranken im Fall unabhängiger Beobachtungen. Dies hat seine Ursache darin, dass in der Information aus einer Beobachtung in Daten mit Abhängigkeitsstrukturen auch schon Information aus anderen Datenpunkten enthalten ist. Unter Berücksichtigung der oben beschriebenen Entropie-Konzepte, die Abhängigkeitsstrukturen ausnutzen, sind hier Verbesserungen vorstellbar.
Eine Verallgemeinerung der Ergebnisse zur Konvergenzrate des empirischen Risikos auf reelle Funktionen wird in dieser Arbeit nicht konkret durchgeführt. Diese ist allerdings direkt möglich, denn, wie aus der Herleitung der Schranken für reelle Funktionen bei unabhängigen Beobachtungen deutlich wird, kann dieser Fall auf Indikatorfunktionen zurückgeführt werden.

Das generelle Ergebnis dieser Arbeit ist der Nachweis, dass die empirische Risiko-Minimierung auch bei Mixingal- und auch Martingal-Strukturen in den Daten ein allgemeingültiges Prinzip darstellt. Diese Resultate beziehen sich ausschließlich auf den konzeptionell theoretischen Teil der statistischen Lerntheorie. Dennoch geht daraus hervor, dass in einer Datensituation mit dynamischen Strukturen bei der Anwendung von Verfahren bzw. Algorithmen, die das ERM-Prinzip nutzen, überprüft werden sollte, ob die empirischen Verluste den Voraussetzungen und Bedingungen der in dieser Arbeit untersuchten Abhängigkeitsstrukturen genügen. Insbesondere müssen die Verluste also als Mixingal oder Martingal modelliert werden können.

Literaturverzeichnis

Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D. (1997). Scale-Sensitive Dimensions, Uniform Convergence, and Learnability, *Journal of the Association for Computer Machinery*, Vol. 44, S. 615-631.

Andrews, D. W. K. (1988). Laws of Large Numbers for Dependent Non-Identically Distributed Random Variables, *Econometric Theory*, Vol. 4, S. 458-467.

Azuma, K. (1967). Weighted Sums of Certain Dependent Random Variables, *Tohoku Mathematical Journal*, Vol. 19, S. 357-367.

Bachelier, L. (1900). *Théorie de la speculation*. Gauthier-Villars, Paris.

Belousov, A. I., Verzakov, S. A., von Frese, J. (2002). A Flexible Classification Approach with Optimal Generalisation Performance: Support Vector Machines, *Chemometrics and Intelligent Laboratory Systems*, Vol. 64, S. 15-25.

Bernstein, S. (1927). Sur l'extension du théoréme limite du calcul des probabilitiés aux sommes de quantités dépendantes, *Mathematische Annalen*, Vol. 85, S. 1-59.

Bernstein, S. (1940). New Applications of Almost Independent Quantities, *Izvestiya Akademii Nauk, Ser. Math.*, Vol. 4, S. 137-150 (in russisch).

Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis Dimension, *Journal of the Association for Computer Machinery*, Vol. 36, S. 929-965.

Brockwell, P. J., Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer Verlag, New York.

Cao, L. (2003). Support Vector Machines Experts for Time Series Forecasting, *Neurocomputing*, Vol. 51, S. 321-339.

Cao, L., Tay, F. E. H. (2001). Financial Forecasting Using Support Vector Machines, *Neural Computing & Applications*, Vol. 10, S. 184-192.

Chen, Y., Wang, G., Dong, S. (2003). Learning with Progressive Transductive Support Vector Machine, *Pattern Recognition Letters*, Vol. 24, S. 1845-1855.

Chen, X., White, H. (1996). Laws of Large Numbers for Hilbert Space-Valued Mixingales with Applications, *Econometric Theory*, Vol. 12, S. 284-304.

Cherkassky, V., Mulier, F. (1998). *Learning from Data – Concepts, Theory and Methods*. Wiley & Sons, New York.

Cherkassky, V., Shao, X. (1999). Model Complexity Control for Regression Using VC Generalization Bounds, *IEEE Transactions on Neural Networks*, Vol. 10, S. 1075-1089.

Cherkassky, V., Shao, X. (2001). Signal Estimation and Denoising Using VC-Theory, *Neural Networks*, Vol. 14, S. 37-52.

Cherkassky, V., Shao, X., Mulier, F., Vapnik, V. (1999). Model Complexity Control for Regression Using VC Generalization Bounds, *IEEE Transactions on Neural Networks*, Vol. 10, S. 1075-1089.

Chernoff, H. (1952). A Measure of Asymptotic Efficiency for Tests Based on the Sum of Observations. *Annals of Mathematical Statistics*, Vol. 23, S. 493-509.

Davidson, J. (1993). An L_1-Convergence Theorem for Heterogeneous Mixingale Arrays with Trending Moments, *Statistics & Probability Letters*, Vol. 16, S. 301-304.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press, Oxford.

Davidson, J., de Jong, R. (1997). Strong Laws of Large Numbers for Dependent Heterogeneous Processes: A Synthesis of Recent and New Results, *Econometric Reviews*, Vol. 16, S. 251-279.

Davidson, R., MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York.

de Jong, R. M. (1995). Laws of large numbers for dependent heterogeneous processes, *Econometric Theory*, Vol. 11, S. 347-358.

de Jong, R. M. (1996). A Strong Law of Large Numbers for Triangular Mixingale Arrays, *Statistics & Probability Letters*, Vol. 27. S. 1-9.

de Jong, R. M. (1998). Weak Laws of Large Numbers for Dependent Random Variables, *Annales d'économie et de statistique*, Vol. 51, S. 209-225.

Doob, J. L. (1953). *Stochastic Processes*. Wiley & Sons, New York.

Evgeniou, T., Poggio, T., Pontil, M., Verri, A. (2002). Regularization and statistical learning theory for data analysis, *Computational Statistics & Data Analysis*, Vol. 38, S. 421-432.

Friedman, J. H. (1994). An Overview of Predictive Learning and Function Approximation, in: Cherkassky, V., Friedman, J. H., Wechsler, H. (Hrsg.), *From Statistics to Neural Networks*, S. 1-61.

Gray, R. M. (1988). *Probability, Random Processes, and Ergodic Properties*. Springer Verlag, New York.

Gut, A. (1992). The weak law of large numbers for arrays, *Statistics & Probability Letters*, Vol. 14, S. 49-52.

Hall, P., Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.

Hansen, B. E. (1991). Strong laws for dependent heterogeneous processes, *Econometric Theory*, Vol. 7, S. 213-221.

Hansen, B. E. (1992). Erratum, *Econometric Theory*, Vol. 8, S. 421-422.

Hansen, B. E. (1992). Convergence to stochastic integrals for dependent heterogeneous processes, *Econometric Theory*, Vol. 8, S. 489-500.

Hanson, D. L., Koopmans, L. H. (1965). Convergence rates for the law of large numbers for the linear combinations of exchangeable and $*$-mixing stochastic processes, *Annals of Mathematical Statistics*, Vol. 36, S. 1840-1852.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Verlag, New York.

Heisele, B., Serre, T., Prentice, S., Poggio, T. (2003). Hierarchical classification and feature reduction for fast face detection with support vector machines, *Pattern Recognition*, Vol. 36, S. 2007-2017.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *American Statistical Association Journal*, Vol. 58, S. 13-30.

Joachims, T. (1998). Making Large-Scale SVM Learning Practical, *Research Reports of the unit no. VIII (AI) Computer Science Department of the University of Dortmund*, Vol. 24, S. 1-13.

Kohler, M., Krzyżak, A., Schäfer, D. (2002). Application of structural risk minimization multivariate smoothing spline regression estimates, *Bernoulli*, Vol. 8, S. 475-489.

Laib, N. (1999). Exponential-type inequalities for martingale difference sequences. Application to nonparametic regression estimation, *Communications in Statistics – Theory and Methods*, Vol. 28, S. 1565-1576.

Lévy, S. (1935). Propriétés asymptotiques des sommes de variables aléatoires enchainées, *Bulletin Sciénce Mathématique*, Vol. 59, S. 84-96 & S. 109-128.

Lévy, S. (1937). *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, Paris.

Li, S. Z., Zhu, L., Zhang, Z. Q., Blake, A., Zhang, H. J., Shum, H. (2002). Statistical Learning of Multi-view Face Detection, *Proceedings of ECCV 2002*, S. 67-81.

Loève, M. (1977). *Probability Theory I*. Springer Verlag.

Ma, J. S., Theiler, J., Perkins, S. (2003). Accurate On-line Support Vector Regression, *Neural Computation*, Vol. 15, S. 2683-2703.

Mangasarian, O. L., Musicant, D. R. (2000). Robust Linear and Support Vector Regression, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, S. 950-955.

McLeish, D. L. (1975). A maximal inequality and dependent strong laws, *The Annals of Probability*, Vol. 3, S. 829-839.

McLeish, D. L. (1977). On the invariance principle for nonstationary mixingales, *The Annals of Probability*, Vol. 5, S. 616-621.

Mendelson, S. (2003). A few Notes on Statistical Learning Theory, *in: Mendelson, A., Smola, A. J. (eds.) Advanced Lectures on Machine Learning*, S. 1-40.

Mitrinović, D. S. (1970). *Analytic Inequalities*. Springer Verlag, Berlin.

Müller, K. R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., Vapnik, V. (1997). Predicting Time Series with Support Vector Machines, *Proceedings of ICANN'97*, Springer Lecture Notes in Computer Science, S. 999-1005.

Mukherjee, S., Osuna, E., Girosi, F. (1997). Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines, *Proceedings of IEEE NNSP'97*, S. 1-10.

Parrondo, J. M. R., van den Broek, C. (1993). Vapnik-Chervonenkis Bounds for Generalization, *Journal of Physics A: Mathematical and General*, Vol. 26, S. 2211-2223.

Philipp, W., Stout, W. (1975). *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables*. American Mathematical Society, Providence, Rhode Island.

Pötscher, B. M., Prucha, I. R. (1991a). Basic Structure of the Asymptotic Theory in Dynamic Nonlinear Econometric Models, Part I: Consistency and Approximation Concepts, *Econometric Reviews*, Vol. 10, S. 125-216.

Pötscher, B. M., Prucha, I. R. (1991b). Basic Structure of the Asymptotic Theory in Dynamic Nonlinear Econometric Models, Part II: Asymptotic Normality, *Econometric Reviews*, Vol. 10, S. 253-325.

Rao, M. M. (1995). *Stochastic Processes: General Theory*. Kluwer Academic Publishers, Dordrecht.

Schölkopf, B. (1998). SVMs - A Practical Consequence of Learning Theory. *IEEE Intelligent Systems*, Vol. 13, S. 18-21.

Schölkopf, B. (2003). Statistical Learning Theory, Capacity and Complexity. *Complexity*, Vol. 8, S. 87-94.

Stout, W. F. (1974). *Almost Sure Convergence*. Academic Press, New York.

Valiant, L. G. (1984). A Theory of the Learnable, *Communications of the ACM*, Vol. 27, S. 1134-1142.

Vapnik, V., Levin, E., Cun, Y. L. (1994). Measuring the VC-Dimension of a Learning Machine, *Neural Computation*, Vol. 6, S. 851-876.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer Verlag, New York.

Vapnik, V. (1993). Three Fundamental Concepts of the Capacity of Learning Machines, *Physica A*, Vol. 200, S. 538-544.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley & Sons, New York.

Vapnik, V. (1999). An Overview of Statistical Learning Theory, *IEEE Transactions on Neural Networks*, Vol. 10, S. 988-999.

Vapnik, V., Chapelle, O. (2000). Bounds on Error Expectation for Support Vector Machines, *Neural Computation*, Vol. 12, S. 2013-2036.

Vapnik, V., Chervonenkis, A. Y. (1968). Uniform Convergence of Frequencies of Occurence of Events to their Probabilities, *Soviet Mathematics – Doklady*, Vol. 9, S. 915-918.

Vapnik, V., Chervonenkis, A. Y. (1971). On the Uniform Convergence of Relative Frequencies of Events to their Probabilities, *Theory of Probability and its Applications*, Vol. 16, S. 264-280.

Vapnik, V., Chervonenkis, A. Y. (1979). *Theorie der Zeichenerkennung*. Akademie Verlag, Berlin.

Vapnik, V., Chervonenkis, A. Y. (1981). Necessary and Sufficient Conditions for the Uniform Convergence of Means to their Expectations, *Theory of Probability and its Applications*, Vol. 26, S. 532-553.

Vapnik, V., Chervonenkis, A. Y. (1991). The Necessary and Sufficient Conditions for Consistency in the Empirical Risk Minimization Method, *Pattern Recognition and Image Analysis*, Vol. 1, S. 283-305.

Vidyasagar, M. (1997). *A Theory of Learning and Generalization*. Springer Verlag, Berlin.

Ville, J. (1939). *Etudes critique de la notion de collectif*. Gauthier-Villars, Paris.

Vovk, V. G. (1997). A Strictly Martingale Version of Kolmogorov's Strong Law of Large Numbers, *Theory and its Applications*, Vol. 41, S. 605-608.