Diplomarbeit

Nichtlineare numerische Verfahren zur multivariaten Dichteschätzung

angefertigt am

Institut für Numerische Simulation

vorgelegt der

Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

November 2006

von

Peter Hahnen

aus

Neuss

Inhaltsverzeichnis

1 Einleitung

Aufgrund der immer stärkeren Verschmelzung von Telekommunikation und Informatik haben wir heute eine Gesellschaft, in der in nahezu allen Bereichen unseres alltäglichen Lebens Daten in großen Mengen erfasst und gespeichert werden. Viele Unternehmen oder Institutionen sehen sich mit sehr großen Datenbanken konfrontiert, die bei geeigneter Auswertung für sie nützliche Informationen enthalten. Problematisch ist dabei, dass aufgrund des anhaltenden Wachstums des Internets und der geringen Preise für Speichermedien immer größere Datenbanken entstehen mit deren Auswertung viele Menschen überfordert sind. Daher besteht seit einigen Jahren ein steigender Bedarf an Methoden, mit denen relevante Informationen aus gegebenen Datenbanken automatisiert extrahiert werden können. Im Laufe dieser Entwicklung ist das so genannte *Data Mining* entstanden, das sich mit der Gewinnung von impliziten, bislang unbekannten Informationen aus Daten beschäftigt. Die dabei entwickelten computergestützten Verfahren werden unter dem Begriff *maschinelles Lernen* zusammengefasst.

Anwendung finden diese Techniken bereits in einer Vielzahl wissenschaftlicher Arbeiten. So werden beispielsweise Gene mittels Mikroarray-Daten analysiert, Protein-Sequenzen klassifiziert, Beobachtungsdaten der Astrophysik ausgewertet, EEG-Daten in der Medizin untersucht, seismische Messungen gruppiert oder archäologische Ausgrabungsergebnisse bewertet.

Auch in der Wirtschaft gibt es ein breites Spektrum möglicher Anwendungen des Data Minings. Besonders durch die immer stärkere Nutzung des Internets zur Abwicklung computergestützter Geschäftsprozesse entstehen bei Unternehmen riesige Datenbanken, zu deren Auswertung maschinelle Lernverfahren gewinnbringend eingesetzt werden können.

Weiter werden Verfahren des maschinellen Lernens bei der Analyse und Vorhersage von Aktien- oder Devisenkursen mittels historischer Kursdaten, zur Entscheidungshilfe bei der Vergabe von Krediten, dem Versenden von Werbebriefen an interessierte Kunden oder dem so genannten Customer Relationship Mangement angewandt. Einen guten Überblick über Anwendungen liefert [BL99], sowie [FE01].

Die allgemeine Problemstellung eines maschinellen Lernproblems lautet: Gegeben sei eine Datenmenge

$$S = \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in X \subset \mathbb{R}^d, y_i \in Y \subset \mathbb{R}\}_{i=1}^n$$

bestehend aus Datenpunkten $\boldsymbol{x}_i$ eines d-dimensionalen Merkmals- oder Datenraumes X und Antwortvariablen y_i, die den vorherzusagenden Sachverhalt beschreiben. Gesucht ist eine Funktion $f : X \to Y$, die den angenommenen funktionalen Zusammenhang zwischen X und Y möglichst gut wiedergibt. Dabei soll

f nicht nur auf den gegebenen Datenpunkten nahe an der Zielvariablen liegen, sondern vor allem bei neuen Punkten gute Vorhersagen liefern. Der Kompromiss zwischen Genauigkeit auf den gegebenen Daten und Generalisierung auf neuen Daten muss bei allen Verfahren des maschinellen Lernens gemacht werden.

Je nachdem wie der Raum Y gestaltet ist, unterscheidet man:

- *Regression*, falls Y unendlich ist,
- *Klassifikation*, für diskretes Y, sowie
- *Dichteschätzung*, falls Y nur aus einem Element besteht.

Ziel dieser Arbeit ist die Entwicklung eines neuen Verfahrens zur Dichteschätzung.

Dichteschätzung

Zur Untersuchung eines vorliegenden Datensatzes auf Strukturen, Cluster oder Anomalien ist eine Dichteschätzung bestens geeignet, da bei Massendaten die Bedeutung eines einzelnen Objektes (Punktes im hochdimensionalen Raum) gegen Null geht und die üblichen Punkte-Plots schwer lesbar und auch rechentechnisch uneffektiv werden. Will man beispielsweise einen Datensatz auf Cluster untersuchen, so kann eine Dichteschätzung vorab Gebiete mit hoher Dichte voneinander separieren und die spätere Klassifikation erleichtern. Weiterhin werden Dichteschätzungen zur Visualisierung und Auswertung großer Datenmengen verwendet. Hat man die Dichtefunktion eines Datensatzes vorliegen, so ist dieser damit vollständig charakterisiert. In [EDD03] und [MP95] werden Dichteschätzungen beispielsweise zur Auswertung von Bilddaten eingesetzt. Interessant sind derartige Methoden für Geheim- oder Sicherheitsdienste, die riesige Mengen von Bilddaten per Satellit oder Videoüberwachung aufnehmen, deren manuelle Analyse sehr zeitintensiv ist. Auf ähnliche Weise lassen sich Schrift-, Ton oder Spracherkennungen durchführen.

Problemstellung

Gegeben sei nun eine Menge von unabhängigen, identisch verteilten Daten

$$D = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n.$$

Dabei heißt identisch verteilt, dass die Daten derselben Wahrscheinlichkeitsverteilung zugrunde liegen. Gesucht ist eine möglichst gute Näherung der unbekannten Wahrscheinlichkeitsdichte.

Ohne weitere Bedingungen an die Dichte muss die Lösung nicht eindeutig sein - das Problem ist schlecht gestellt. Es ergeben sich ernsthafte numerische Schwierigkeiten, die mit dem Konzept der *Regularisierung* behandelt werden müssen [Lou].

Weiterhin besteht das Problem der Überanpassung, welches zu schlechten Vorhersageergebnissen auf neuen Daten führen kann. Oft besteht die Anwendung nicht in der Bestimmung einer Dichtefunktion auf den gegebenen Datenpunkten, sondern in der Auswertung auf vorab unbekannten Daten. Ziel ist also eine

gute Verallgemeinerung. Diese wird bei fast allen Lernalgorithmen durch spezielle Parameter gesteuert. Die optimale a-priori Bestimmung dieser Parameter ist in den meisten Fällen unklar und wird durch Ansätze wie Kreuzvalidierung, Methoden aus der strukturellen Risikominimierung oder Ähnlichem versucht (siehe [Vap00], [Sco92],[Wah90]).

Bestehende Verfahren

Generell unterscheidet man Dichteschätzungsmethoden in *parametrische* und *nichtparametrische* Verfahren. Bei *parametrischen* Ansätzen ist a-priori die Art der Verteilung bekannt. Es müssen nun aus den vorhandenen Daten die zugehörigen Parameter der Verteilung ermittelt werden. Ist beispielsweise bekannt, dass die Daten normalverteilt sind, so müssen nur noch Erwartungswert und Varianz bestimmt werden. Die klassischen statistischen Verfahren hierzu sind das *Maximum-Likelihood-* und das *Maximum-a-posteriori-Verfahren*. Das Maximum-Likelihood-Verfahren bestimmt die gesuchten Parameter θ der Verteilung als Maximum der Likelihood-Funktion

$$L(\theta, D) = \prod_{i=1}^{n} p(\boldsymbol{x}_i, \theta) = P(D|\theta),$$

die aus Punktdichten an den jeweiligen Datenpunkten besteht. Geht man zusätzlich von einer *a-priori* Verteilung der möglichen Parameter aus, so erhält man ein Maximum-*a-posteriori*-Verfahren, welches die gesuchten Parameter durch Maximierung der bedingten Wahrscheinlichkeit für eine feste Parameterwahl bei vorliegender Datenmenge ermittelt. Nach dem Satz von Bayes maximiert man also die *a-posteriori* Wahrscheinlichkeit

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}.$$

Im Gegensatz dazu machen *nichtparametrische* Verfahren keinerlei Annahmen über die Verteilung und bestimmen eine Dichtefunktion nur aus den gegebenen Daten.

Ist der Typ der Verteilung bekannt, wird ein parametrisches Verfahren die beste Wahl sein. Allerdings kann bei falscher Wahl der Verteilung der Approximationsfehler theoretisch beliebig groß werden. Daher hat sich in den letzten Jahren die Forschungsarbeit hauptsächlich auf nichtparametrische Verfahren konzentriert, welche man in *gitterbasiert* und *datenbasiert* unterscheidet.

- *Gitterbasierte* Ansätze diskretisieren den Datenraum durch ein Gitter und bestimmen die Lösung als Linearkombinationen diskreter Basisfunktionen aus einem Funktionenraum über diesem Gitter.

- *Datenbasierte* Verfahren stellen auf den vorhandenen Datenpunkten oder einem geeigneten Teil dieser Punkte so genannte *Kernfunktionen* auf. Die gesuchte Funktion ergibt sich als endliche Linearkombination dieser Kernfunktionen.

Bei allen Verfahren der Dichteschätzung wird angenommen, dass die Dichte an einem festen Punkt Einfluss auf benachbarte Punkte hat. Bei datenbasierten Ansätzen - oft *Kernschätzer* genannt - geben entsprechende Kernfunktionen

durch ihre Gestalt die Auswirkungen eines einzelnen Punktes auf die Gesamtdichte vor, während gitterbasierte Verfahren einen Funktionenraum wählen, dessen Funktionen eine gewisse Glattheit zu erfüllen haben. Umgesetzt wird dies bei vielen gitterbasierten Verfahren durch einen Regularisierungsoperator, der als Strafterm auf das Zielfunktional addiert wird und damit eine Steuerung der Regularität der Lösung ermöglicht.
Ein großer Vorteil bei der numerischen Umsetzung von Kernschätzern ist, dass diese nur mit der Datenmenge n skalieren. Die Aufstellung und Lösung der $(n \times n)$ Matrizen für Kern-basierte Ansätze erfolgt üblicherweise in $O(n^3)$ unabhängig von der Dimension des Problems.
Gitterbasierte Methoden skalieren hingegen meist nur linear in den Daten, dafür aber exponentiell in der Dimension. Bei einem äquidistanten Gitter mit N Knoten in jeder der d-Dimensionen ergibt sich somit ein Aufwand von $O(N^d)$. Dieser exponentielle Anstieg des Aufwands für die numerische Berechnung wird auch als *Fluch der Dimension* bezeichnet. Je nachdem wie die Relation von n^3 zu N^d ausfällt, sollte man sich für ein daten- bzw. gitterbasiertes Verfahren entscheiden.

Viele nichtparametrische Verfahren der Dichteschätzung lassen sich analog zu Ansätzen der Regression bzw. Klassifikation als *Regularisierungsnetzwerk* schreiben. Das über einem geeigneten Funktionenraum V zu minimerende Funktional hat dann die Form:

$$\min_{f \in V} R(f) = \sum_{i=1}^{n} C(f(\boldsymbol{x}_i), \hat{f}(\boldsymbol{x}_i)) + \lambda \Phi(f).$$

Der erste Term erzwingt mit einer Kostenfunktion C die Nähe der Funktion f zu den Daten, der zweite Term erzeugt eine gewisse Glattheit von f und der Regularisierungsparameter λ gewichtet diese beiden Terme. Für den Fall der Regression ist $\hat{f}(\boldsymbol{x}_i) = y_i$, während bei der Dichteschätzung angenommen wird, das bereits ein Schätzer vorliegt - beispielsweise in Form eines einfachen Histogramms. Mit Hilfe des *Repräsentertheorems* und der Theorie so genannter *reproduzierender Kern-Hilberträume*, lassen sich unter gewissen Voraussetzungen Äquivalenzen zwischen datenbasierten- und gitterbasierten Verfahren zeigen. Je nach Wahl der Kernfunktion resultieren spezielle Regularisierungsterme und umgekehrt. Wahba motiviert in [Wah90] Regularisierungsnetzwerke - dort *glättende Splines* genannt - Bayesianisch und zeigt Zusammenhänge zu stochastischen Prozessen. Durch spezielle Kostenfunktionen lassen sich auch *Support-Vektor-Maschinen* oder Approximationen mit radialen Basisfunktionen im Kontext von Regularisierungsnetzwerken herleiten. In [HHR99] wird für $\hat{f}$ die empirische Dichtefunktion eingesetzt, die aus Punktmassen an den Datenpunkten besteht, und daher theoretisch schwer zu handhaben ist. Vapnik und Mukherjee geben die Information der Daten in Form der empirischen Verteilungsfunktion vor und bestimmen mit Hilfe eines Support-Vektor-Ansatzes eine Approximation der Verteilungsfunktion [MV99].

Fast alle in der Literatur vorgestellten Verfahren arbeiten mit einer quadratischen Kostenfunktion C, so dass nach Minimierung ein lineares Gleichungssystem entsteht, welches direkt gelöst werden kann. Die Wahl der Kostenfunktion

und des Glättungsterms ist dabei jedoch genauso willkürlich wie die Wahl der Kernfunktionen bei datenbasierten Verfahren. Sie sind theoretisch nicht begründet.

Ein neuer Ansatz

In dieser Arbeit wird ein auf Hegland und Griebel zurückgehender Ansatz untersucht, der klassische parametrische Methoden der Statistik mit effizienten nichtparametrischen Verfahren verbindet. Von der grundlegenden Idee entspricht der Ansatz einem Maximum-a-posteriori-Verfahren, in dem so genannte *exponentielle Familien* von Verteilungen zur Lösung herangezogen werden. Diese werden durch einen Parametervektor charakterisiert. Um eine maximale Verallgemeinerung zu erzielen, werden die Parametervektoren als unendlich angesehen, weshalb sie auch als Parameterfunktionen interpretiert werden können. Die Klasse möglicher Dichtefunktionen wird hierdurch nur sehr gering eingeschränkt; im Gegensatz zu den üblichen Maximum-a-posteriori-Verfahren.

Die Wahl von *unendlichen* Parametervektoren führt dazu, dass man Schwierigkeiten mit Dichten bzw. Maßen bekommt, die für den unendlich dimensionalen Fall nicht im klassischen Sinne definiert sind. Mit dem von Hegland und Griebel entwickelten Ansatz, der so genannte *verallgemeinerte Dichten* einführt, ist trotzdem eine Maximierung des a-posteriori-Maßes möglich [GH06].

Der in dieser Arbeit vorgestellte Maximum-a-posteriori-Ansatz ist neben der Dichteschätzung auch für Regressions- und Klassifikationsaufgaben geeignet. Die Herleitung wird daher in einer sehr allgemeinen Form dargestellt, aus der sich die einzelnen Teilaufgaben ableiten lassen. Weiterhin wird in dieser Arbeit gezeigt, dass sich für eine spezielle Wahl der Exponentialfamilie Querverbindungen zu klassischen Ansätzen ziehen lassen. Wählt man die Familie der Normalverteilungen als Spezialfall einer exponentiellen Familie, so resultieren quadratische Zielfunktionale, die nach Minimierung lineare Gleichungssysteme ergeben. Es entstehen klassische Regularisierungsnetzwerke.

Konkret lautet das Zielfunktional für die Dichteschätzung:

$$\min_{u \in H} J(u) = \|u\|_H^2 + n \log \int \exp(u(x)) dx - \sum_{i=1}^{n} u(x_i).$$

Der erste Term von J gibt die Regularität der gesuchten Dichte vor, der dritte Term enthält die Informationen aus den Datenpunkten. Neu ist die Gestalt des zweiten, *nichtlinearer* Massenterms. Im Gegensatz zu *linearen* Massentermen bei klassischen Verfahren der Dichteschätzung ergibt sich dieser durch Wahl der exponentiellen Familien und ist somit statistisch begründet.

Die gesuchte Dichtefunktion f ergibt sich aus der Parameterfunktion u als

$$f(x) = \frac{\exp(u(x))}{\int_X \exp(u(z)) dz}.$$

Es kann dabei gezeigt werden, dass $J(u)$ konvex ist und ein eindeutiges Minimum besitzt.

Zur Lösung der entstehenden nichtlinearen Gleichungen werden in dieser Arbeit iterative Verfahren entwickelt. Als geeignet stellt sich dabei ein Ansatz heraus, der ähnlich zu einem Newton-Verfahren jeweils lokal eine quadratische Approximation an das Zielfunktional J bestimmt und so iterativ gegen die Lösung konvergiert. Die quadratischen Approximation resultieren nach der Minimierung in linearen Gleichungssystemen, welche direkt gelöst werden können.

Diskretisiert wird das hier vorgestellte Verfahren mit Hilfe so genannter *dünner Gitter* [Zen91] [BG04]. Diese ermöglichen es, den *Fluch der Dimension* zu einem gewissen Teil zu umgehen. Definiert werden dünne Gitter über eine Multiskalen-Tensorproduktbasis, die aus eindimensionalen hierarchischen Basisfunktionen konstruiert wird. Die zugrunde liegende Idee ist es, Basisfunktionen der hierarchischen Darstellung, die nur geringen Einfluss auf die Lösung und somit nur geringen Anteil am Fehler haben, bei der Diskretisierung zu vernachlässigen. Am Beispiel von Interpolationsaufgaben konnte gezeigt werden, dass bei einem Aufwand von $O(N \cdot \log N^{(d-1)})$ unter gewissen Glattheitsvoraussetzungen an die zu interpolierende Funktion eine Genauigkeit der Ordnung $O(N^{-2} \cdot \log(N^{-1}))$ erreicht werden kann. Eine Darstellung auf dem vollen Gitter benötigt im Vergleich dazu $O(N^d)$ Gitterpunkte für eine Genauigkeit von $O(N^{-2})$. Der Aufwand bei der numerischen Berechnung reduziert sich damit deutlich und es wird möglich höherdimensionale Datenmengen auszuwerten. In [Gar04] werden dünne Gitter bereits erfolgreich für Regressionsaufgaben in bis zu zehn Dimensionen verwendet.

Während auf dem vollen Gitter für $O(N^d) > O(n^3)$ das kernbasierte Verfahren den geringeren Aufwand aufweist, verschiebt sich diese Relation bei Verwendung dünner Gitter zu $O(N \cdot \log N^{(d-1)})$ gegenüber $O(n^3)$ für einen datenbasierten Ansatz. Da sich in der Realität die Dimension des Datensatzes nach geeigneter Vorverarbeitung meistens auf eine moderat hohe *effektive Dimension* reduzieren lässt, während die untersuchten Datenmengen in der Regel sehr groß sind, fällt die Entscheidung zu Gunsten eines Dünngitter-basierten Verfahrens aus.

Realisiert werden dünne Gitter in dieser Arbeit mittels der so genannten *Kombinationstechnik*, die auf Griebel, Schneider und Zenger zurückgeht [GSZ92]. Dabei wird die Lösung auf dem dünnen Gitter als Summe von Teillösungen auf anisotropen Gittern, die in jeder Koordinatenrichtung äquidistant sind, aufgestellt. Diese Lösung stimmt zwar nicht unbedingt mit der Lösung auf dem eigentlichen dünnen Gitter überein, es kann aber gezeigt werden, dass die Genauigkeit von der gleichen Ordnung ist, falls gewisse Fehlerentwicklungen für die Teilprobleme existieren [GSZ92].

Anhand numerischer Experimente wird das vorgestellte Maximum-a-posteriori-Verfahren mit exponentiellen Familien auf Konvergenz analysiert und mit diversen anderen Dichteschätzern verglichen. Zum direkten Vergleich mit linearen Methoden wird ein auf Regularisierungsnetzwerken basierender Ansatz, welcher die Information der Daten in Form der empirischen Verteilungsform enthält, ein verallgemeinertes Histogramm sowie ein Kernschätzer mit Gaußkern verwendet.

Die erzielten numerischen Ergebnisse erweisen sich als viel versprechend und

rechtfertigen den durch die Nichtlinearität entstehenden Mehraufwand.

Die Beiträge dieser Arbeit lassen sich folgendermaßen zusammenfassen:

- Eine auf Regularisierungsnetzwerken basierende, lineare Methode der Dichteschätzung wird vorgestellt.
- Die Herleitung eines Maximum a-posteriori Ansatzes mit Gauß-Prior für exponentielle Familien wird dargestellt.
- Iterative Ansätze zur Lösung des nichtlinearen Problems werden entwickelt.
- Das vorgestellte Verfahren wird mittels der Kombinationstechnik für dünne Gitter diskretisiert.
- Anhand numerischer Experimente wird das Maximum-a-posteriori-Verfahren mit bestehenden Methoden verglichen.

Überblick dieser Arbeit

Zu Beginn dieser Arbeit werden in *Kapitel 2* die mathematischen Grundlagen der Dichteschätzung vorgestellt. Sie umfassen grundlegende Aussagen aus der Wahrscheinlichkeitstheorie, sowie eine kurze Einführung in die Theorie der reproduzierenden Kern-Hilberträume. Abschließend wird auf die Fehlermessung eingegangen.

In *Kapitel 3* werden die klassischen Methoden der Dichteschätzung vorgestellt, zunächst kurz die parametrischen Schätzer, anschließend nichtparametrische Schätzer, die wiederum in gitterbasiert und datenbasiert unterschieden werden. Darunter fallen klassische und verallgemeinerte Histogramme, diverse Kernschätzer, k-nächste Nachbarn-Verfahren und auf orthogonalen Reihen basierende Schätzer. Mit einer interessanten Aussage, die besagt, dass sich nahezu alle Dichteschätzer als Kernschätzer schreiben lassen, schließt dieses Kapitel.

Das *4. Kapitel* befasst sich mit der Herleitung eines neuen Schätzers, der auf Basis der empirischen Verteilungsfunktion eine Approximation der unbekannten Dichte bestimmt und als Vergleichsverfahren dient. Bevor dieses vorgestellt wird, werden zwei Techniken, aus denen dieser Ansatz entwickelt wurde, dargestellt. Zum einen ein von Hegland, Hooker und Roberts entwickeltes Verfahren, welches ähnlich zu klassischen Regularisierungsnetzwerken arbeitet, und zum anderen ein auf Support-Vector-Maschinen basierender Dichteschätzer von Vapnik und Mukherjee.

In *Kapitel 5* wird der Maximum-a-posteriori-Ansatz auf exponentiellen Familien mit Gauß-Prior zunächst für ein allgemeines Lernproblem hergeleitet und anschließend das Zielfunktional für den Spezialfall Dichteschätzung aufgestellt. Für die Herleitung wird eine verallgemeinerte Definition einer Dichtefunktion im unendlichdimensionalen Fall sowie eine Charakterisierung stationärer Punkte benötigt.

Kapitel 6 befasst sich mit möglichen Lösungsansätzen des nichtlinearen Problems. Zunächst wird ein expliziter Ansatz untersucht. Anschließend wird ein

verbessertes Iterationsverfahren entwickelt, welches das Zielfunktional lokal quadratisch approximiert. Als Diskretisierungsmethode wird die Kombinationstechnik für dünne Gitter vorgestellt.

Kapitel 7 beinhaltet einige Implementierungsdetails sowie eine Konvergenzanalyse des implementierten Algorithmus. Anhand von numerischen Experimenten wird das vorgestellte Maximum-a-posteriori-Verfahren mit bekannten Verfahren verglichen und die Qualität des resultierenden Schätzers auf simulierten Daten gezeigt. Einige Ergebnisse auf realen Daten runden dieses Kapitel ab.

Die Arbeit schließt in *Kapitel 8* mit einer Zusammenfassung der wichtigsten Ergebnisse, sowie einem Ausblick auf noch offene Fragestellungen und mögliche Erweiterungen des Verfahrens.

Danksagung

An dieser Stelle möchte ich mich bei allen, die zur Entstehung dieser Arbeit beigetragen haben, bedanken. Mein Dank gilt vor allem Prof. Dr. M. Griebel für die Überlassung des Themas, die Bereitstellung eines exzellenten Arbeitsumfeldes sowie zahlreiche hilfreiche Ideen und Vorschläge. Besonderer Dank gilt ihm und Markus Hegland von der Australian National University für frühe Einblicke in noch unveröffentlichte Manuskripte. Darüber hinaus bedanke ich mich bei meinen Zimmergenossen für die angenehme gemeinsame Zeit am Institut, insbesondere Allan Zulficar für die vielen erhellenden Diskussionen und seine Hilfen mit LaTeX. Ebenfalls danken möchte ich Prof. Dr. R. Krause für die Übernahme des Koreferates.

2 Grundlagen

Zu Beginn dieser Arbeit wollen wir eine kurze Zusammenfassung der grundlegenden mathematischen Begriffe und Aussagen geben. Leser, die mit den entsprechenden Theorien bereits vertraut sind, können dieses Kapitel getrost überspringen.

Zunächst werden die Begriffe Zufallsvariable, Verteilungsfunktion und Dichtefunktion aus der Wahrscheinlichkeitstheorie eingeführt. Da wir spätere Dichteschätzer immer im Kontext einer realen Anwendung sehen wollen, werden wir uns hier auf Radonsche Wahrscheinlichkeitsmaße über $\mathbb{R}^d$ beschränken. Darüberhinaus benötigt man für das Verständnis diverser Verfahren aus der statistischen Lerntheorie sogenannte *reproduzierende Kern-Hilberträume*. Diese werden im letzten Unterkapitel vorgestellt, bevor kurz auf die Fehlermessung eingegangen wird.

Eine vollständige Darstellung der Grundlagen der Wahrscheinlichkeitstheorie findet sich beispielsweise in [Bau02].

2.1 Zufallsvariablen und ihre Verteilung

In vielen Wissenschaften werden Beobachtungen festgehalten und Messgrößen bestimmt, die als Realisierungen von Zufallsvariablen aufgefasst werden können. Beobachtet man zum Beispiel in der Biologie eine bestimmte Tierpopulation über mehrere Jahre, so wird man von dieser diverse Merkmale wie die Gesamtstärke der Population, Gewicht, Alter und Größe einzelner Tiere, oder die Positionen der Tiere auf einem betrachteten Gebiet festhalten, die man nachher auswertet. Diese Merkmale können alle als Zufallsvariable betrachtet werden.

Zufallsvariablen

Für eine formale Schreibweise brauchen wir die folgenden Definitionen.

Definition 2.1.1 (Wahrscheinlichkeitsraum)**.** Man definiert einen *Wahrscheinlichkeitsraum* (Ω, Σ, P) über seine drei Bestandteile:

- einer nichtleeren Menge Ω, dem *Ergebnisraum* oder *Stichprobenraum*,
- einer σ-Algebra Σ auf Ω, dem *Ereignisraum*. Wobei eine *σ-Algebra* definiert ist durch:
  1. $\emptyset, \Omega \in \Sigma$
  2. $A \in \Sigma \Rightarrow \Omega \backslash A \in \Sigma$
  3. $A_1, A_2, ... \in \Sigma \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \Sigma$
- und einem Wahrscheinlichkeitsmaß P, für das gelten muss

1. $0 \leq P(A) \leq 1$ für alle $A \in \Sigma$
2. $A_1, A_2, ... \in \Sigma$ mit $A_i \cap A_j =$ für $i \neq j \Rightarrow P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
3. $P(\Omega) = 1$

Definition 2.1.2 (Messraum)**.** Ein Tupel (Ω, Σ) heißt *Messraum*, falls Σ eine σ-Algebra auf einer nichtleeren Menge Ω ist.

Definition 2.1.3 (Zufallsvariable)**.** Sei (Ω, Σ, P) ein Wahrscheinlichkeitsraum und (Ω', Σ') ein Messraum. Dann heißt die Funktion $X : \Omega \to \Omega'$ *Zufallsvariable* auf Ω.

Der Name Variable ist also etwas irreführend, da es sich eigentlich um eine Funktion handelt. In unserem Beispiel ist der Raum Ω die Tierpopulation und der Zielraum die reellen Zahlen (Gewicht, Größe), bzw. der $\mathbb{R}^2$ (Positionen der Tiere). Wir wollen im Folgenden als Messraum das Tupel $(\mathbb{R}^d, \mathcal{B})$ verwenden. Dabei ist $\mathcal{B}$ die borelsche σ-Algebra, die aus d-dimensionalen offenen Mengen gebildet wird. Wir schränken uns auf *reellwertige Zufallsvariablen* ein, da wir weniger an maßtheoretischen Aussagen interessiert sind, sondern immer auch eine Anwendung im Hintergrund sehen, bei der die vorliegenden Messgrößen aus $\mathbb{R}$ stammen.

Man unterscheidet zwischen *diskreten* und *stetigen* Zufallsvariablen. Falls X nur endlich viele oder abzählbar unendlich viele verschiedene Werte annehmen kann, so spricht man von einer *diskreten* Zufallsvariablen. Bei unseren Biologen wäre das zum Beispiel die Gesamtanzahl an Tieren, die die Population aufweist. Eine *stetige* Zufallsvariable hat entsprechend eine kontinuierliche Bildmenge. Dies ist zum Beispiel bei Größenmessungen der Fall. *Mehrdimensionale Zufallsvariablen* definiert man als Vektor von eindimensionalen Zufallsvariablen und spricht dann auch von *Zufallsvektoren*. Ordnen wir also beispielsweise jedem Tier seine Position auf der Landkarte zu, so erhalten wir kontinuierliche zweidimensionale Zufallsvariablen der Koordinaten, einen Zufallsvektor.

Verteilungsfunktion und Dichtefunktion

Charakterisiert wird eine (eindimensionale reellwertige) Zufallsvariable durch ihre *Verteilungsfunktion*.

Definition 2.1.4 (Verteilungsfunktion)**.** Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Dann heißt die Funktion $F : \Omega \to [0, 1]$, die jedem Intervall $(-\infty, x]$ die Wahrscheinlichkeit [1]

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \in (-\infty, x]\})$$

zuordnet, *Verteilungsfunktion* von X.

Oft wird eine Zufallsvariable auch als stetig definiert, falls ihre Verteilungsfunktion stetig ist. Man stellt leicht fest, dass F eine monoton steigende Funktion mit Supremum 1 und Infimum 0 ist.

[1] allgemein: Funktion, die jeder messbaren Menge A einer σ-Algebra über Ω die Wahrscheinlichkeit $P(\{\omega \in \Omega : X(\omega) \in A\})$ zuordnet

Für diskrete Zufallsvariablen ordnet die Verteilungsfunktion jedem möglichen Wert x von X eine nichtnegative Wahrscheinlichkeit $P(X = x)$ zu. Dabei muss gelten $P(X = x) \leq 1$ und $\sum_{x\in\mathbb{R}} P(X = x) = 1$. Die zugehörige *diskrete Verteilungsfunktion* lässt sich also beschreiben als

$$F(x) = P(X \leq x) = \sum_{y\leq x} P(X = y).$$

In unserem Beispiel kann man damit beispielsweise die Wahrscheinlichkeit ausdrücken, dass die Stärke der beobachteten Population eine bestimmte Grenze unterschreitet.

Eine sehr häufig auftretende diskrete Wahrscheinlichkeitsverteilung ist die *Binomialverteilung*. Hierbei gibt es nur zwei mögliche Ausgänge des Zufallsexperiments. Sei p die Wahrscheinlichkeit für Ereignis A, dann ist die Wahrscheinlichkeit bei n Durchführungen genau k mal Ereignis A zu erzielen

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

und die Verteilungsfunktion ergibt sich als Summe der einzelnen Wahrscheinlichkeiten

$$P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}.$$

Analog kann man im stetigen Fall die Verteilungsfunktion durch ein Integral über die Dichtefunktion ausdrücken.

Definition 2.1.5 (Dichtefunktion)**.** Eine integrierbare Funktion f heißt *Dichtefunktion* der Zufallsvariablen X, falls

$$\int_a^b f(t)dt = P(a \leq X \leq b)$$

gilt.

Ist die Verteilungsfunktion F einer Zufallsvariablen stetig differenzierbar, so definiert ihre Ableitung

$$f(x) = \frac{dF(x)}{dx}$$

eine Dichtefunktion. Diese erfüllt $f(x) \geq 0$ für alle $x \in \mathbb{R}$ und

$$\int_{\mathbb{R}} f(t)dt = 1.$$

Man schreibt für eine Wahrscheinlichkeitsdichte auch kurz *pdf* (probability density funktion), und für die Verteilungsfunktion *cdf* (cumulative density funktion).

Hat man für eine Zufallsvariable X eine zugehörige Dichtefunktion gegeben, drückt man die Verteilungsfunktion F als Integral

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

aus. Im diskreten Fall wird in der Literatur gelegentlich die Wahrscheinlichkeit $P(X = x)$ als *diskrete Dichtefunktion* bezeichnet. Ein wesentlicher Unterschied ist, dass die Dichte einer stetigen Zufallsvariablen an einem Punkt verschwindet, d.h.

$$P(X = x) = \int_x^x f(t)dt = 0,$$

da dies eine Nullmenge bezüglich des hier verwendeten Lebesguemaßes ist. Man kann bei kontinuierlichen Verteilungen also nicht wie im diskreten Fall von der Wahrscheinlichkeit für das Eintreten von x sprechen. Stattdessen betrachtet man die Wahrscheinlichkeit, dass X Werte aus einem bestimmten Intervall (a, b), d.h. einer messbaren Menge mit Maß ungleich Null annimmt

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(t)dt. \tag{2.1.1}$$

Normalverteilung

Bekanntestes Beispiel einer Dichtefunktion ist die Dichte der Normalverteilung

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

Diese ist vollständig charakterisiert durch Erwartungswert und Varianz

$$E[X] = \int_{-\infty}^{\infty} tf(t)dt =: \mu, \quad Var[X] = \int_{-\infty}^{\infty} (t - E[X])^2 f(t)dt =: \sigma^2.$$

In Abbildung 2.1 sind für festen Erwartungswert und verschiedene Varianzen jeweils die Dichtefunktion und die Verteilungsfunktion der zugehörigen Normalverteilungen eingezeichnet. Offensichtlich lässt sich die Dichtefunktion einer

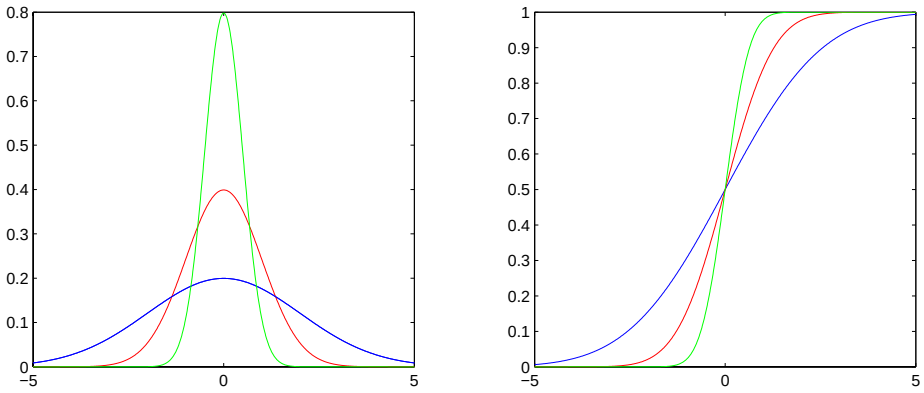

Abbildung 2.1: Normalverteilungen mit $\mu = 0$ und $\sigma^2 = 1/2, 1, 2$

Zufallsvariablen wesentlich besser interpretieren als die Verteilungsfunktion. Bei der Dichtefunktion sieht man sofort die Maximalstellen, die in der Verteilungsfunktion als Wendepunkte weniger klar zu erkennen sind. Dies ist einer der

Gründe, warum man in den meisten Fällen eher an einer Dichtefunktion, als an einer Verteilungsfunktion der vorliegenden Daten interessiert ist.

Mehrdimensionale Verteilung

Hat man nun mehrere Zufallsvariablen gegeben, so interessiert man sich mitunter für deren gemeinsame, also mehrdimensionale, Verteilungs- und Dichtefunktion.

Definition 2.1.6 (Gemeinsame Verteilungsfunktion)**.** Es seien Zufallsvariablen $X_1, .., X_N : \Omega \to \mathbb{R}$ gegeben. Dann heißt eine Funktion $F : \mathcal{B} \to [0,1]$ [2], die jeder messbaren Menge $A \in \mathcal{B}$ die Wahrscheinlichkeit

$$P((X_1, ..., X_N) \in A) = P\left(\bigcap_{i=1}^{\mathbb{N}} \{\omega \in \Omega : X_i(\omega) \in A\}\right)$$

zuordnet, *gemeinsame Verteilung* dieser Zufallsvariablen.

Definition 2.1.7 (Gemeinsame Dichte)**.** Seien $X_1, .., X_N : \Omega \to \mathbb{R}$ Zufallsvariablen mit gemeinsamer Verteilungsfunktion $F(x_1, ..., x_n)$. Dann ist eine L^1-Funktion $f : \mathbb{R}^N \to [0, \infty)$ eine *gemeinsame Dichte*, falls für alle messbaren Mengen $A \subset \mathbb{R}^N$ sich deren Wahrscheinlichkeit als Volumenintegral

$$P((X_1, ..., X_n) \in A) = \int_A f(\boldsymbol{t}) d\boldsymbol{t}$$

schreiben lässt.

Hat man die Verteilung eines Zufallsvektors gegeben, so ist oft die Verteilung einer einzelnen Komponente bei Festhalten der anderen von Interesse. Man spricht von der *Randdichte* einer Zufallsvariablen. Berechnet wird diese durch Integration über die restlichen Komponenten.

Definition 2.1.8 (Randdichte)**.** Sei $X = (X_1, ..., X_n)$ ein Zufallsvektor reelwertiger Zufallsvariablen mit Dichtefunktion f. Dann heißt die Dichte

$$f_{X_i}(x_i) = \int_{\mathbb{R}} ... \int_{\mathbb{R}} f(x_1, ..., x_n) dx_1 ... dx_{i-1} dx_{i+1} ... dx_n$$

Randdichte der Zufallsvariablen X_i.

Unabhängigkeit von Zufallsvariablen

Wie bereits in der Einleitung beschrieben, gehen wir bei den Untersuchungen in späteren Kapiteln davon aus, dass wir unabhängige, identisch verteilte Daten vorliegen haben. Man schreibt dafür auch kurz i.i.d. (independently and identically distributed).

Definition 2.1.9 (unabhängige Zufallsvariablen)**.** Sei (Ω, Σ, P) ein beliebiger Wahrscheinlichkeitsraum. Die Zufallsvariablen $X_1, ..., X_n : \Omega \to \mathbb{R}$ heißen *unabhängig*, falls

$$P(X_1 = x_1, ..., X_n = x_n) = P(X_1 = x_1) \cdot ... \cdot P(X_n = x_n), \quad \forall (x_1, ..., x_n) \in \mathbb{R}^n.$$

Damit haben wir die Begriffe Dichte- und Verteilungsfunktion theoretisch erläutert und wenden uns nun den entsprechenden empirischen Größen zu.

[2] mit $\mathcal{B}$ ist wieder die Borelsche σ-Algebra auf $\mathbb{R}^N$ gemeint

2.2 Empirische Verteilungs- und Dichtefunktion

Hat man eine Menge $D = \{x_1, ..., x_n\}$ von Datenpunkten gegeben, wobei die x_i unabhängige Resultate der gleichen Zufallsvariablen X darstellen, so kann man zunächst die *empirische Verteilungsfunktion* davon aufstellen.

Definition 2.2.1 (ecdf)**.** Die *empirische Verteilungsfunktion* (kurz ecdf) einer Datenmenge $D = \{x_1, ..., x_n\}$ ist definiert als

$$F_n(x) = \frac{\#\{x_i \leq x\}}{n} = \frac{\#x_i \in (-\infty, x]}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_{(-\infty,x]}(x_i). \tag{2.2.1}$$

Dabei ist

$$\mathcal{X}_A = \begin{cases} 1, & \text{falls} \quad x \in A \\ 0, & \text{sonst} \end{cases}$$

die sogenannte *Indikatorfunktion* auf A.

F_n hat somit die Gestalt einer Treppenfunktion, die jeweils an den Datenpunkten einen Sprung aufweist. Die Zufallsvariable $nF_n(x)$ ist binomialverteilt $B(n, p)$, mit $p = E[F_n(x)]$. Genauso wie die Binomialverteilung für $n \to \infty$ zur Normalverteilung konvergiert, konvergiert die empirische Verteilungsfunktion bei immer größer werdender Stichprobe gegen die wirkliche Verteilungsfunktion. Dies ist auch die Aussage des folgenden Satzes, der allgemein als *Fundamentalsatz der Statistik*, oder *Satz von Glivenko und Cantelli* bezeichnet wird.

Satz 2.2.2 (Fundamentalsatz der Statistik)**.** *Es sei $x_1, x_2, ...$ eine Folge von i.i.d. Auswertungen einer Zufallsvariablen zur Verteilungsfunktion F. Die zur Stichprobe $x_1, x_2, ..., x_n$ gehörige empirische Verteilungsfunktion F_n konvergiert für $n \to \infty$ mit Wahrscheinlichkeit 1 gleichmäßig gegen F. Formal ausgedrückt:*

$$P\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0\right) = 1.$$

Der theoretischen Beziehung $f(x) = F'(x)$ folgend, läßt sich die *empirische Wahrscheinlichkeitsdichte* (kurz epdf) definieren als die Ableitung der empirischen kumulativen Verteilungsfunktion

$$f_n(x) = \frac{d}{dx}F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i),$$

wobei $\delta(x)$ die Diracsche Deltafunktion ist. Die empirische Dichtefunktion ist also eine hochgradig unstetige Funtion mit Dichte $1/n$ in jedem Datenpunkt. Als Schätzer einer stetigen Dichtefunktion ist sie somit nicht zu gebrauchen, und eine Konvergenz zu einer stetigen Dichtefunktion läßt sich im Gegensatz zur ecdf nicht zeigen. Trotzdem ist man in den meisten Fällen eher an einer Dichtefunktion als an einer Verteilungsfunktion der Daten interressiert. Das liegt daran, dass man aus einer Dichtefunktion wesentlich besser wichtige Eigenschaften wie die Schiefe der Verteilung oder Stellen mit hoher Dichte ablesen kann.

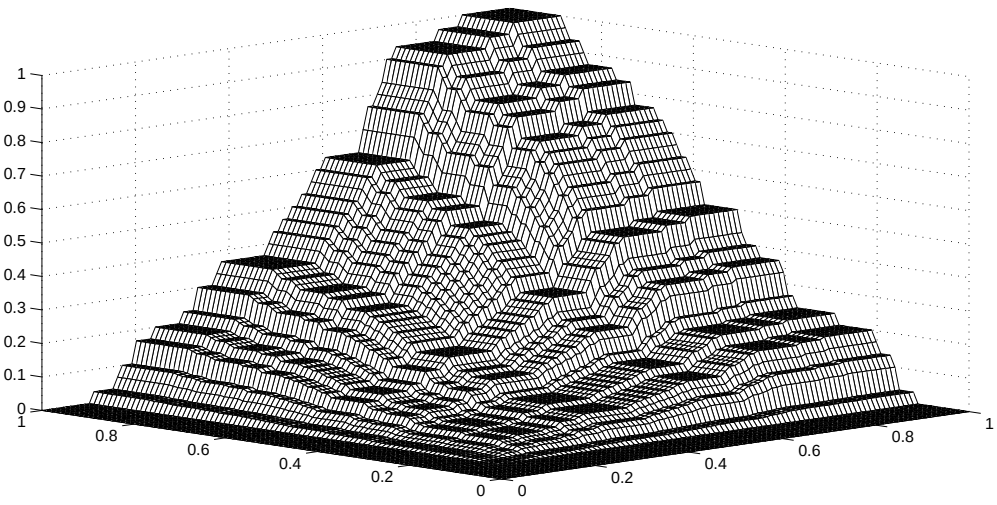

Abbildung 2.2: Empirische Verteilungsfunktion eines simulierten Datensatzes aus 30 Datenpunkten

Besonders anschaulich wird das in höheren Dimensionen. In Abbildung 2.2 ist für einen zweidimensionalen simulierten Datensatz bestehen aus 30 Datenpunkten die empirische Verteilungsfunktion gezeichnet. Wie man sieht, fällt es einem bereits in 2D schwerer, Informationen aus der Verteilungsfunktion abzulesen.

Wird die Dichteschätzung im Rahmen einer Clusteranalyse oder einer Regressionsaufgabe als Hilfsroutine verwendet, ist man meistens an einer Dichtefunktion und nicht an einer Verteilungsfunktion interessiert. Hier soll eine Dichteschätzung vorab Gebiete mit hoher Dichte voneinander separieren, um das Risiko der Missklassifikation zu verringern.

2.3 Radon-Nikodym Theorie

Im Folgenden werden Vorraussetzungen für die Existenz einer Dichtefunktion im Allgemeinen genauer untersucht. Da wir in Kapitel 5 als Stichprobenraum Ω einen Funktionenraum, z.B. den Raum $C([0,1])$ aller stetigen Funktionen auf dem Beobachtungsintervall $[0,1]$ gegeben haben, brauchen wir einen allgemeineren Begriff einer Dichtefunktion. Hilfreich ist dazu aus der Maßtheorie der Satz von Radon Nikodym, der in bestimmten Fällen die Existenz von Funktionen sichert, welche die Rolle von Dichten im klassischen Fall übernehmen. Bevor wir diesen angeben können, benötigen wir die folgende Definition:

Definition 2.3.1 (absolutstetig)**.** Seien ν und μ Maße auf einer σ-Algebra Σ. Dann heißt μ *absolutstetig* bezüglich ν, falls für alle $\mathcal{E} \in \Sigma$ mit $\nu(\mathcal{E}) = 0$ auch $\mu(\mathcal{E}) = 0$ gilt. Man schreibt dafür:

$$\mu \ll \nu.$$

Falls sowohl $\mu \ll \nu$ als auch $\nu \ll \mu$ gilt, so heißen μ und ν *äquivalent*: $\mu \equiv \nu$.

Satz 2.3.2 (Satz von Radon Nikodym)**.** *Es seien ν und μ zwei σ-finite Maße auf (Ω, Σ). Ist $\mu \ll \nu$, so existiert eine nichtnegative Funktion f auf Ω mit folgenden Eigenschaften:*

1. *f ist Σ-messbar*
2. *$\mu(A) = \int_A f(x)\nu(dx)\ \forall A \in \Sigma$.*

Die Funktion f ist ν-fast-überall eindeutig bestimmt, dass heißt für jede Σ-messbare Funktion g mit $\int_A f d\nu = \int_A g d\nu$, $\forall A \in \Sigma$, gilt $f = g$ ν-f.ü.

Beweis. Ein Beweis findet sich beispielsweise in [Bau92] Kapitel 17. □

Die Funktion f aus dem Satz von Radon Nikodym heißt *Radon-Nikodym-Ableitung* von μ nach ν und wird mit $\frac{d\mu}{d\nu}$ bezeichnet. Damit kann man die 2. Eigenschaft aus obigem Satz schreiben als:

$$\mu(A) = \int_A \frac{d\mu}{d\nu}(x)\nu(dx), \quad \forall A \in \Sigma.$$

Als Kurzschreibweise ist auch $\mu = f \cdot \nu$ gebräuchlich. Man bezeichnet dann f als die Dichte von μ bezüglich ν.

Definition 2.3.3 (absolutstetige Verteilung)**.** Eine Verteilungsfunktion F auf $\mathbb{R}$ heißt *absolutstetig bezüglich dem Lebesguemaß* μ, falls das von ihr erzeugte Wahrscheinlichkeitsmaß P absolutstetig bezüglich μ ist.

In diesem Fall liefert der Satz von Radon Nikodym eine borelmessbare Funktion f mit

$$P(A) = \int_A f d\mu, \quad A \in \mathcal{B}, \quad \text{insbesondere gilt} \quad F(x) = \int_{-\infty}^{x} f(t)dt.$$

Ist F absolutstetig, so ist F μ-fast-überall (bis auf Lebesgue-Nullmengen) differenzierbar, und es gilt

$$\frac{dF}{dx} = f(x) \quad \mu\text{-f.ü.},$$

dass heißt die Ableitung von F stimmt μ-fast-überall mit der Radon-Nikodym-Ableitung von P nach μ überein. Außerdem gilt somit für jede absolutstetige Verteilungsfunktion F

$$F(x) = \int_{-\infty}^{x} \frac{dF}{dt}dt, \quad \text{d.h. } F \text{ hat eine Dichte} \quad f(t) = \frac{dF}{dt}.$$

Die hier vorgestellten Radon-Nikodym-Dichten werden in späteren Kapiteln benötigt, um den Begriff der Dichte auf unendliche Funktionenräume zu verallgemeinern.

2.4 Reproduzierende Kern-Hilberträume

Eine wichtige Rolle bei der Wahl von Funktionenräumen in der statistischen Lerntheorie, spielen so genannte *reproduzierende Kern-Hilberträume*. Auf ihnen basierende Modelle sind Grundlage für eine Vielzahl von Lernverfahren. So lassen sich sämtliche auf *Regularisierungsnetzwerken* basierende Verfahren mit Hilfe des unten angeführten Repräsentertheorems als Kern-basierte Verfahren darstellen, bei denen die Lösung sich als endliche Linearkombination der Kernfunktionen eines entsprechenden reproduzierenden Kern-Hilbertraumes darstellen lässt. Darunter fallen zum Beispiel Neuronale Netzwerke, Thin-Plate-Splines, additive Modelle und Support-Vektor-Maschinen. Für eine detailliertere Analyse sei auf [SS98],[Gir97],[SSM98] und [Wah90] hingewiesen. Erstmals vorgestellt wurden diese Räume 1950 von Aronszajn in [Aro50].

Definition 2.4.1 (Reproduzierender Kern-Hilbertraum)**.** Ein Hilbertraum H von Funktionen $f : X \to \mathbb{R}$, $X \neq \emptyset$ heißt *reproduzierender Kern-Hilbertraum* (RKHS), falls die Punktauswertungen $f \to f(x)$ für alle $x \in X$ beschränkte lineare Funktionale sind.

Sei nun H ein RKHS über einem Gebiet X gegeben, dann existiert nach dem Rieszschen Darstellungssatz für jedes $x \in X$ ein Element $\eta_x \in H$ mit der Eigenschaft

$$f(x) = \langle \eta_x, f \rangle \quad \text{für alle} \quad f \in H,$$

wobei $\langle \cdot, \cdot \rangle$ das innere Produkt von H ist. η_x wird der *Repräsenter* der Auswertung an der Stelle x genannt. Die Funktion

$$k(x, y) = \langle \eta_x, \eta_y \rangle$$

wird als *reproduzierender Kern* des Hilbertraumes bezeichnet.

Die η_x agieren hier analog zu den Diracschen Deltafunktionen im L^2. Ein wesentlicher Unterschied ist jedoch, dass die Deltafunktionen nicht in L^2 liegen. Daher ist der L^2 auch kein reproduzierender Kern-Hilbertraum.

Ein Kern $k(x, y)$ heißt positiv semidefinit, falls

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

für beliebige $n \in \mathbb{N}$, $x_1, ..., x_n \in X$ und $c_1, ..., c_n \in \mathbb{R}$. Falls $>$ gilt, spricht man entsprechend von positiv definit.

Damit ergibt sich eine grundlegende Eigenschaft eines reproduzierenden Kern-Hilbertraumes:

Satz 2.4.2. *Zu jedem RKHS gibt es eine eindeutige positiv semidefinite Funktion, den so genannten reproduzierenden Kern. Umgekehrt kann zu jeder positiv semidefiniten Funktion auf $X \times X$ ein eindeutiger RKHS von reelwertigen Funktionen auf X konstruiert werden.*

Der zu einem Kern k assoziierte Hilbertraum wird dabei konstruiert, indem er alle endlichen Linearkombinationen der Form $\sum_j c_j k(x_j, \cdot)$ und deren Limiten unter der durch das innere Produkt induzierten Norm enthält.

Anzumerken ist noch, dass sich mit Hilfe eines Tensorproduktes in analoger Weise mehrdimensionale reproduzierende Kern-Hilberträume definieren lassen.

Satz 2.4.3. *Das Tensorprodukt $H = H_1 \otimes H_2$, wobei H_1 und H_2 Hilberträume mit reproduzierendem Kern k_1 beziehungsweise k_2 sind, ist ebenfalls ein RKHS. Als reproduzierender Kern von H ergibt sich das Produkt der reproduzierenden Kerne*

$$k((x_1, x_2), (y_1, y_2)) = k_1(x_1, y_1) \cdot k_2(x_2, y_2).$$

Repräsentertheorem

Interessant für die Lerntheorie ist der Zusammenhang zwischen Regularisierungsoperatoren und Hilberträumen mit reproduzierendem Kern, der durch das so genannte *Repräsentertheorem* ausgedrückt wird. Dieses geht auf Wahba und Kimeldorf zurück (siehe [WK71]). Die wesentliche Aussage ist, dass sich unter gewissen Voraussetzungen die Lösung eines allgemeinen Lernproblems als endliche Linearkombination von Basisfunktionen ergibt, die auf den Datenpunkten positioniert sind.

Betrachten wir dazu die so genannte *Tikhonov-Regularisierung*

$$R(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|Tf\|^2 \tag{2.4.1}$$

für einen gegebenen linearen Operator T. Gesucht ist ein Minimum von $R(f)$ über alle Funktionen f aus einem Funktionenraum H. Dann wird der Lösungsraum alleine durch den Regularisierungsterm $\|Tf\|^2$ charakterisiert. Die resultierende Lösung lässt sich dann schreiben als

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) \tag{2.4.2}$$

mit reproduzierendem Kern k. Konstruiert werden kann dieser Kern mit den Eigenvektoren des linearen Operators T. Seien $\{\gamma_j\}_{j=1}^{\infty}$ die Eigenwerte und $\{\varphi_j\}_{j=1}^{\infty}$ die zugehörigen Eigenvektoren. Dann ergibt sich für den Kern die Darstellung

$$k(x, y) = \sum_{j=1}^{\infty} \frac{1}{\gamma_j} \varphi_j(x) \varphi_j(y). \tag{2.4.3}$$

Umgekehrt exisitiert auch zu jedem Kern ein entsprechender Regularisierungsoperator, wie der folgende Satz besagt.

Satz 2.4.4. *Für jeden RKHS H mit reproduzierendem Kern k existiert ein entsprechender Regularisierungsoperator $T : H \to G$, wobei G ein Hilbertraum ist, so dass für alle $f \in H$ gilt:*

$$\langle Tk(x, \cdot), Tf(\cdot) \rangle_G = f(x). \tag{2.4.4}$$

Insbesondere gilt

$$\langle Tk(x,\cdot), Tk(y,\cdot)\rangle_G = k(x,y). \tag{2.4.5}$$

Umgekehrt existiert für jeden Regularisierungsoperator $T : V \to G$, wobei V ein mit einem Skalarprodukt versehener Funktionenraum ist, ein entsprechender RKHS H mit reproduzierendem Kern k, so dass (2.4.4) *und* (2.4.5) *gelten.*

Daraus folgt insbesondere, dass G mit dem inneren Produkt $\langle T\cdot, T\cdot\rangle_G$ versehen ein RKHS ist.

Allerdings ist die Beziehung zwischen Regularisierungsoperator T und Raum H mit Kern k nicht eindeutig, da zu einem gegebenen Operator T theoretisch mehrere verschiedene Kerne gebildet werden können. Die obige Darstellung durch die Eigenvektoren existiert jedoch immer.

Mit Hilfe dieses Repräsentertheorems lässt sich also ein und dasselbe Problem auf zwei verschiedene Arten darstellen: Zum einen in einem unendlichdimensionalen Funktionenraum, zum anderen als endliche Superposition von Kernfunktionen, die auf den Datenpunkten liegen.

Verwendet man statt der quadratischen Kostenfunktion in (2.4.1) die so genannte ϵ-insensitive Kostenfunktion nach Vapnik, lassen sich auf diese Weise Support-Vektor-Maschinen darstellen. Für radial-symmetrische Kerne k resultiert ein Radiale-Basisfunktionen-Approximationsverfahren und auch im Falle einer nicht quadratischen Kostenfunktion hat die Lösung trotzdem immer die Gestalt (2.4.2), wie Smola und Schölkopf in [SS01] zeigen.

2.5 Fehlermessung

Hat man einen Dichteschätzer $\hat{f}$ konstruiert so stellt sich die Frage nach einer geeigneten Norm um dessen Güte bestimmen zu können. Wir gehen dazu im Weiteren davon aus, dass die Originaldichte f explizit gegeben ist.

Zunächst bieten sich als Normen die L_p-Normen, mit

$$\|(f-\hat{f})\|_{L_p} = \left(\int |f-\hat{f}|^p dx\right)^{1/p}, \quad \text{für} \quad 1 \le p < \infty \quad \text{und}$$
$$\|f-\hat{f}\|_{L_\infty} = \sup_x |f(x)-\hat{f}(x)|$$

an. Am häufigsten verwendet werden davon im Allgemeinen die L_1- und die L_2-Norm.

Man macht sich leicht mit Hilfe der Dreiecksungleichung klar, dass die L_1-Norm durch 2 nach oben beschränkt ist. Im Gegensatz dazu ist die L_2-Norm theoretisch unbeschränkt. Trotzdem wird sehr oft die quadrierte L_2-Norm zur Fehlermessung verwendet, insbesondere weil sie theoretisch leichter zu handhaben ist. Man bezeichnet den integrierten quadratischen Fehler meist mit *ISE (integrated squared error).*

$$ISE\{\hat{f(x)}\} = \int [\hat{f}(x) - f(x)]^2 dx$$

Für eine punktweise Auswertung wird üblicherweise analog der mittlere quadratische Fehler verwendet, der sich schreiben lässt als

$$MSE\{\hat{f}(x)\} = E[\hat{f}(x) - f(x)]^2 = Var[\hat{f}(x)] + Bias^2\{\hat{f}(x)\},$$

wobei $Bias\{\hat{f}(x)\} = E[\hat{f}(x)] - f(x)$. Das Quadrat des punktweisen Fehlers ist also die Summe aus einem Varianz-Term und einem bias-Term des Schätzers.

Im kontinuierlichen Fall ist das Analogon der so genannte *MISE (mean integrated squared error)*

$$\begin{aligned} MISE\{\hat{f}(x)\} = E[ISE] = E\left[\int (\hat{f}(x) - f(x))^2 dx\right] = \int E[\hat{f}(x) - f(x)]^2 dx \\ = \int MSE\{\hat{f}(x)\} dx = IMSE\{\hat{f(x)}\}. \end{aligned}$$

Anzumerken ist, dass die L_2-Norm Bereiche in denen f klein ist weniger stark gewichtet als die L^1-Norm. Je nachdem ob dies gewünscht ist oder nicht, sollte man sich für die entsprechende Norm entscheiden.

In späteren Kapiteln werden wir den mittleren l_2-Fehler auf dem Diskretisierungsgitter

$$MSE_{grid}\{f - \hat{f}\} = \frac{1}{N} \sum_{i=1}^{N} \left(f(t_i) - \hat{f}(t_i)\right)^2 \tag{2.5.1}$$

mit Gitterpunkten t_i und den analogen Datenfehler

$$MSE_{data}\{f - \hat{f}\} = \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - \hat{f}(x_i)\right)^2 \tag{2.5.2}$$

auf den Datenpunkten x_i betrachten.

Für eine detailliertere Analyse der verschiedenen Fehlernormen sei an dieser Stelle auf [Sco92] verwiesen.

3 Diverse Standardverfahren

Im Folgenden betrachten wir eine Menge von unabhängigen identisch verteilten (i.i.d.) Datenpunkten $\{(\boldsymbol{x}_i) \in \mathbb{R}^d\}_{i=1}^n$. Wir nehmen an, dass diesen Daten eine Dichtefunktion f zugrunde liegt, die wir mit den gegebenen Datenpunkten möglichst gut rekonstruieren wollen. Gesucht ist also eine Approximation $\hat{f}(x)$ der unbekannten Dichtefunktion f.

Grundsätzlich unterscheidet man dabei parametrische und nichtparametrische Verfahren. Im ersten Abschnitt dieses Kapitels stellen wir die *parametrische* Dichteschätzung kurz vor. Dabei geht man davon aus, dass der Verteilungstyp a-priori bekannt ist (z.B. Normalverteilung), also dass

$$f(x) \in \{f(x|\theta), \theta \in \Theta\}, \quad (\text{z.B} \quad \Theta = \mathbb{R}^k)$$

gilt. Aufgabe ist es nun, die entsprechenden Parameter θ der Verteilung auf Grundlage der Datenpunkte möglichst gut zu schätzen (z.B. Erwartungswert und Varianz einer Normalverteilung).

Im zweiten Abschnitt gehen wir auf die *nichtparametrische* Dichteschätzung ein. Hier hat man a-priori keinerlei Informationen über den Typ der Verteilung gegeben. Man ermittelt also alleine aus den gegebenen Daten eine Approximation der ursprünglichen Dichtefunktion.

Wann ist ein Schätzer nichtparametrisch?
Es hat sich als überraschend schwierig herausgestellt eine genaue Definition eines nichtparametrischen Schätzers anzugeben. Grob kann man sagen, dass ein nichtparametrischer Schätzer für eine große Klasse von zu approximierenden Dichten arbeiten können soll. Alternativ wird gesagt, ein nichtparametrischer Schätzer sollte möglichst viele, am Besten unendlich viele freie Parameter haben. Nach diesem Verständnis ist beispielsweise das in Kapitel 5 vorgestellte Verfahren ein nichtparametrischer Schätzer, auch wenn der grundlegende Ansatz, der hinter dem Verfahren steht, ein klassischer parametrischer Schätzer ist. Die genaue Klassifizierung in parametrische bzw. nicht parametrische Dichteschätzung ist bei vielen aktuellen Verfahren demnach nicht immer klar.

Wir stellen in diesem Kapitel einige Standardverfahren vor, die einen Einblick in die grundsätzliche Vorgehensweise bei der Dichteschätzung geben sollen.

3.1 Parametrische Dichteschätzer

Falls der Typ der Verteilung bekannt ist, wird eine parametrische Schätzung immer der beste Weg sein. Die prominentesten Beispiele sind das *Maximum-Likelihood-* und das *Maximum-a-posteriori-Verfahren.* Im zweiten Fall spricht man auch von *Bayesian-Parameter-Estimation.*

3.1.1 Maximum-Likelihood-Verfahren

Bei der Maximum-Likelihood-Methode fasst man die Realisierungen x_i als Funktion der gesuchten Parameter θ auf. Die so genannte *Likelihood-Funktion* lautet:

$$L(\theta, D) = \prod_{i=1}^{n} p(x_i, \theta) = p(D|\theta).$$

Dabei ist die Dichte p an jedem Punkt abhängig von dem Parametervektor θ. Für das Beispiel der Normalverteilung ist $\theta = (\mu, \sigma^2)$ mit Erwartungswert μ und Varianz σ^2. Die zugehörige Likelihood-Funktion lautet:

$$L(\mu, \sigma, D) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right).$$

Da hiervon die Ableitungen meistens relativ aufwendig zu berechnen sind, geht man über zur negativen, logarithmierten Likelihood-Funktion

$$l(\theta, D) = -\log L(\theta, D) = -\sum_{i=1}^{n} \log f(x_i, \theta)$$

und minimiert diese nach θ. Daraus erhält man bei normalverteilten x_i die Maximum-Likelihood-Schätzwerte der gesuchten Parameter:

$$l(\mu, \sigma, D) = -n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(x_i - \mu)^2.$$

Die partielle Ableitung nach μ ist

$$\frac{\partial l(\mu, \sigma, D)}{\partial \mu} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} -2(x_i - \mu).$$

Setzt man diese Null, so erhält man

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

und analog nach Minimierung nach σ

$$\sigma^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{ML})^2.$$

Es resultieren Mittelwert und mittlere quadratische Abweichung als Schätzer für Erwartungswert und Varianz der Verteilung.

3.1.2 Maximum-a-posteriori-Verfahren

Im Unterschied zur Maximum-Likelihood-Methode wird hier der Parameter θ als Zufallsvariable betrachtet, wobei das *a-priori* vorhandene Wissen über die Verteilung von θ durch die Dichtefunktion $f(\theta)$ repräsentiert wird, d.h. es wird eine *a-priori* Annahme an die Verteilung von θ gemacht. Bayes-Learning führt die ursprüngliche Wahrscheinlichkeitsdichte $f(\theta)$ nach Beobachtung von n Stichprobenwerten D in eine neue *a-posteriori* Wahrscheinlichkeitsdichte $f(\theta|D)$ über, welche das in den Datenpunkten enthaltene Wissen reflektiert. Diese Wahrscheinlichkeit berechnet sich nach der Bayesschen Regel als

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int_\Theta f(D|\theta)f(\theta)d\theta}.$$

Dabei steht im Zähler die Likelihood Funktion $f(D|\theta)$ multipliziert mit der a-priori Dichte $f(\theta)$. Der Parametervektor θ bestimmt sich dann als

$$\theta_{MAP} = \underset{\theta\in\Theta}{argmax}\{f(\theta|D)\} = \underset{\theta\in\Theta}{argmax}\{f(D|\theta)f(\theta)\},$$

d.h. als Maximum der *a-posteriori* Verteilung. Analog zur Maximum-Likelihood-Methode ist es in den meisten Fällen einfacher die Ableitung der logarithmierten Likelihood-Funktion zu berechnen. Man ändert die Aufgabe also in

$$\theta_{MAP} = \underset{\theta\in\Theta}{argmin}\{-\log(f(D|\theta)) - log(f(\theta))\}.$$

Falls die a-priori Verteilung eine Gleichverteilung ist, ergibt sich als Lösung gerade der Maximum-Likelihood-Schätzer. In dem Fall ist die Dichte $f(\theta)$ konstant und beeinflusst die Lage des Minimums nicht. Dieser Ansatz kann daher auch als Regularisierung der Maximum-Likelihood-Methode interpretiert werden. In Kapitel 5 wird eine nichttriviale Verallgemeinerung der Bayesschen-Parameterbestimmung mit Gauß Prior für unendliche Parametervektoren aus einem Funktionenraum hergeleitet.

Da in der Realität meistens nicht bekannt ist welcher Art von Verteilung die Daten zugrundeliegen, kann man mit falschen Annahmen an diese große Fehler machen. So wie die beiden Verfahren hier vorgestellt wurden, schränken sie den Lösungsraum aus dem die Approximation der Dichte gewählt wird sehr stark ein und können nur angewendet werden, wenn klar ist, um was für eine Verteilung es sich handelt. Das Maximum-a-posteriori-Verfahren, welches in Kapitel 5 vorgestellt wird, gehört vom Ansatz her zwar auch zu den parametrischen Verfahren, durch die Wahl einer Parameterfunktion statt eines endlichen Parametervektors bleibt jedoch sehr viel Spielraum für die Wahl der Lösungsfunktion.

3.2 Nichtparametrische Dichteschätzer

Im Gegensatz zu den parametrischen Verfahren, wo θ globalen Einfluss auf die approximierte Dichte hat, wird bei nichtparametrischen Schätzern ein jeweils lokal wirkender Parameter bestimmt.
Dazu stellt man folgende Überlegung an:

Sei x_0 ein beliebiger gegebener Datenpunkt und x ein neuer, nahe benachbarter Punkt. Dann kann man davon ausgehen, dass die Dichte am Punkt x einen sehr ähnlichen Wert wie in x_0 annimmt. Diese Glattheit der Dichte wird durch einen Parameter modelliert (in den meisten Verfahren h genannt), mit dem man ein Intervall um die Datenpunkte festlegt, auf dem die Dichte lokalen Einfluss hat. Wie genau dies implementiert wird, hängt vom jeweiligen Verfahren ab.

Man unterscheidet *gitterbasierte* und *datenbasierte* Verfahren. Gitterbasierte Verfahren arbeiten mit Ansatzfunktionen, die auf einem diskreten Gitter der Maschenweite h über dem betrachteten Gebiet definiert sind. Hierbei gibt man zusätzlich durch Wahl eines Funktionenraumes, aus dem die Ansatzfunktionen gewählt werden, die gewünschte Glattheit der Lösung vor. Ältestes und am häufigsten verwendetes Verfahren ist hier das Histogramm.

Im Gegensatz dazu stellen datenbasierte Dichteschätzer oder *Kernschätzer* Basisfunktionen an den gegebenen Datenpunkten auf. Die Glattheit der Lösung ergibt sich durch Wahl einer Kernfunktion. Der Parameter h steuert hier die sogenannte Bandbreite des Kernes.

Im Folgenden werden einige häufig verwendete Verfahren vorgestellt, wie sie in [Sil86] und [Sco92] zu finden sind.

3.2.1 Histogramme

Das Standardhistogramm

Die älteste und geläufigste Methode zur Dichteschätzung ist das *Histogramm*. Dabei wird das Intervall I, in dem die Werte x_i, $i = 1, ..., n$ auftreten, in gleich große Teilintervalle $I_1, ..., I_N$ der Breite h unterteilt. Sei nach geeigneter Umnummerierung $x_1 \leq x_2 \leq ... \leq x_{n-1} \leq x_n$.
Die Wahrscheinlichkeit, dass die Zufallsvariable X im Teilintervall I_j liegt, ist dann für jedes $I_j := [x_1 + (j-1)h, x_1 + jh]$:

$$P(X \in I_j) = \int_{x_1+(j-1)h}^{x_1+jh} f(x)dx.$$

Es bietet sich daher an, die relative Häufigkeit der x_i im Intervall I_j als Schätzer dieser Wahrscheinlichkeit zu verwenden, d.h.:

$$P(X \in I_j) \simeq \frac{1}{n}(\#x_i \in I_j) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_{I_j}(x_i)$$

wobei $\mathcal{X}_A$ die Indikatorfunktion auf dem Intervall A ist, d.h. 1 auf A und 0 sonst. Nach dem Mittelwertsatz der Integralrechnung folgt, falls f stetig ist:

$$\int_{x_1+(j-1)h}^{x_1+jh} f(x)dx = f(\xi) \cdot h \quad \text{für ein } \xi \in I_j.$$

Somit liegt es nahe f auf I_j durch einen konstanten Wert zu approximieren:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathcal{X}_{I_j}(x_i)$$

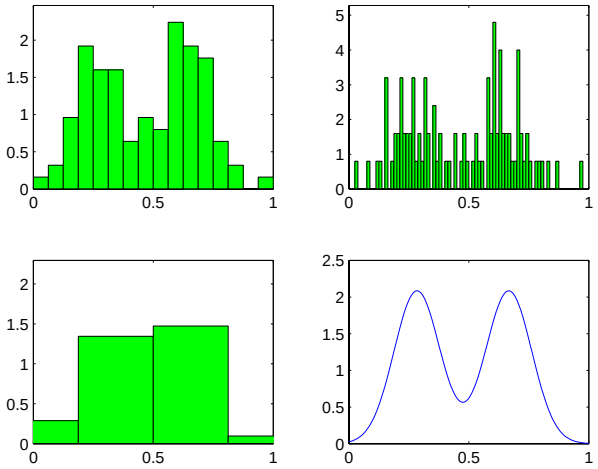

Abbildung 3.1: Histogramme mit verschiedenen Klassenbreiten und die Originaldichte

Damit erhalten wir als Approximation der Dichte das Standardhistogramm mit Ursprung x_1 und Klassenbreite h:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{N} \mathcal{X}_{I_j}(x_i) \mathcal{X}_{I_j}(x). \tag{3.2.1}$$

Die Wahl des Ursprungs ist hier beliebig. In den meisten Fällen wird jedoch der kleinste Wert x_1 der Datenmenge gewählt.

In Anlehnung an die Physik ordnet das Histogramm jedem Datenpunkt die Masse $\frac{1}{n}$ zu und misst, wieviele Datenpunkte (Masse) pro Intervall (Volumen) auftreten. Hauptgründe für die Verwendung des Histogramms als Dichteschätzer sind die einfache Berechnung und die gute Darstellung. Dabei hängt diese jedoch stark von der Wahl der Klassenbreite h und der Position des Ursprungs (hier x_1) ab, wie man in Abbildung 3.1 sehen kann. Für h gegen 0 entsteht ein „Nadeldiagramm“(oben rechts), welches aus einzelnen „Nadeln“ an den Positionen der x_i besteht. Wenn h gegen ∞ geht, wird das Histogramm dagegen zur Gleichverteilung (unten links). Liegt der Ursprung ungünstig, kann es auch passieren, dass $\hat{f}(x)$ stärker von weit entfernten Werten als von den nächsten Nachbarn von x abhängt. Die Hauptaufgabe besteht nun darin die optimale Klassenbreite h und einen geeigneten Ursprung zu wählen. In [Sco92] werden einige Versuche, die Klassenbreite h theoretisch optimal zu wählen, vorgestellt.

Ein großer Nachteil des Histogramms besteht darin, dass der resultierende Schätzer weder stetig noch differenzierbar ist. Mit einer verallgemeinerten Definition eines Histogramms lassen sich jedoch stetige und differenzierbare Dichteschätzer konstruieren, die schon wesentlich bessere Ergebnisse liefern.

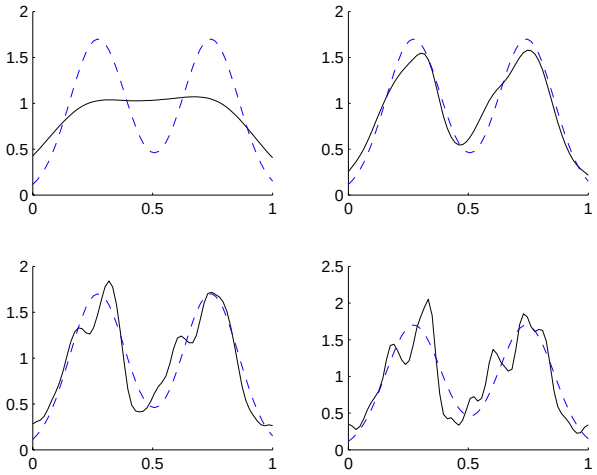

Abbildung 3.2: verallgemeinerte Histogramme mit kubischen B-Splines und unterschiedlicher Intervallbreite h. Die gestrichelte Linie stellt die Originaldichte dar.

Verallgemeinerte Histogramme

Die Idee ist nun statt der Indikatorfunktion $\mathcal{X}_A$, die konstant auf dem Intervall A ist, stückweise lineare, quadratische oder Funktionen höherer Ordnung zu verwenden. Es bietet sich hier an mit so genannten *B-Splines* zu arbeiten.
Ein B-Spline q-ten Grades über dem Intervall I ist auf jedem Teilintervall I_j, $j = 1, ..., N$ ein Polynom vom Grade q, das auf den Intervallgrenzen $q - 1$ mal stetig differenzierbar ist. Es folgt eine rekursive Definition zur Konstruktion von B-Splines beliebiger Ordnung.

Definition 3.2.1 (B-Spline)**.** Sei $I \in \mathbb{R}$ ein reelles Intervall mit Unterteilung in disjunkte Teilintervalle $I_j = [\tau_j, \tau_{j+1}], \quad j = 1, ..., N - 1$.
Dann ist der *B-Spline* der Ordnung q auf dem Intervall i definiert als

$$B_{q,i}(x) = B_{q-1,i}(x)\frac{x - \tau_i}{\tau_{i+q} - \tau_i} + B_{q-1,i+1}(x)\frac{\tau_{i+q+1} - x}{\tau_{i+q+1} - \tau_{i+1}}$$

mit

$$B_{0,i}(x) = \begin{cases} 1 & \text{falls} \quad x \in [\tau_i, \tau_{i+1}] \\ 0 & \text{sonst} \end{cases}.$$

Mit Hilfe eines Tensorproduktes lassen sich die B-Splines auch für höhere Dimensionen verwenden. Dabei kann die Unterteilung in Teilintervalle in jeder Dimension unterschiedlich sein. Die beim Standard-Histogramm verwendete Indikatorfunktion entspricht also einem B-Spline der Ordnung 0. Nun lassen sich durch B-Splines höherer Ordnung analoge Dichteschätzer formulieren, die entsprechend glatter sind.

Sei $x_1 \leq x_2 \leq ... \leq x_{n-1} \leq x_n$ eine i.i.d. Stichprobe einer stetigen Zufallsvariable X mit Dichte f. Weiterhin sei $I_j, \quad j = 1, ..., N$ eine Unterteilung des

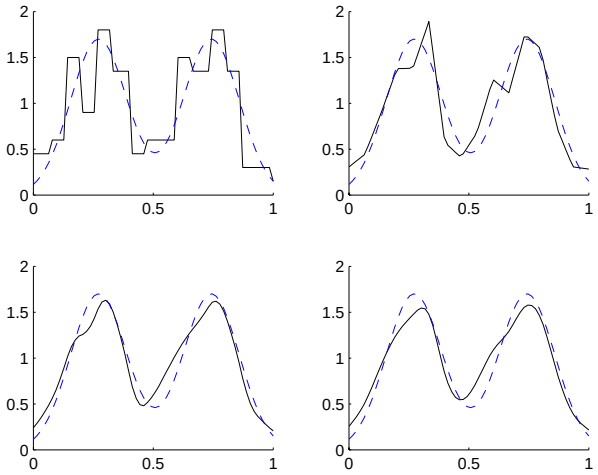

Abbildung 3.3: verallgemeinerte Histogramme mit B-Splines der Ordnung 0 (oben links) bis 3 (unten rechts). Die gestrichelte Linie stellt die Originaldichte dar.

betrachteten Gebietes. Dann definieren wir das *verallgemeinerte Histogramm* dieser Stichprobe als

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j=-q}^{N} B_{j,q}(x_i) B_{j,q}(x). \tag{3.2.2}$$

Hierbei hat jeder Datenpunkt nun nicht nur auf das Intervall, in dem er liegt, sondern auf $(q+1)$ Intervalle Einfluss.

Wie man in Abbildung 3.2 sieht, entstehen abhängig von der Wahl von h sehr unterschiedliche Schätzer, die jeweils andere Interpretationsansätze der Daten zulassen. In Abbildung 3.3 ist die Ordnung der B-Splines variiert worden. Man sieht, dass für $p = 0$ wieder das Standard-Histogramm entsteht. Für höhere Ordnungen wird der resultierende Schätzer sichtbar glatter und einzelne Ausreißer fallen weniger ins Gewicht.

Die bisher vorgestellten Histogramme sind gitterbasierte Schätzer, bei denen der lokale Einfluss der Daten auf ein festes Intervall des Gitters beschränkt wird. Im Gegensatz dazu gibt es die oben genannten datenzentrierten Ansätze, die um jeden Datenpunkt ein Intervall legen, auf welchem der Datenpunkt die Dichte beeinflusst. Im folgenden Abschnitt betrachten wir eine erste Variante davon.

3.2.2 Methode der k nächsten Nachbarn

Der Methode der k nächsten Nachbarn liegt die Annahme zugrunde, dass die Dichtefunktion an zwei nahe beieinander liegenden Punkten ähnliche Werte annimmt. Hierbei wird für jedes x die Entfernung zu allen Datenpunkten x_i bestimmt. Sei dazu $d(x, y)$ eine Abstandsfunktion (im eindimensionalen Fall

$|x-y|$). Man berechnet $d(x,x_i)$ für $i=1,...,n$ und sortiert diese aufsteigend. Nach geeigneter Umbenennung erhält man

$$d_1(x) \leq d_2(x) \leq ... \leq d_n(x).$$

Dann ist der Schätzer der k nächsten Nachbarn definiert als

$$\hat{f}(x) = \frac{k}{2nd_k(x)}.$$

Zum besseren Verständnis dieser Methode wird im Folgenden die Herleitung gezeigt.

Sei x_0 ein beliebiger Punkt und $f(x_0)$ die Dichte an dieser Stelle. Dann erwartet man bei einer Stichprobe vom Umfang n, dass $2rnf(x_0)$ Beobachtungen in das Intervall $[x_0-r, x_0+r]$ fallen. Nach Definition unserer d_k fallen gerade k Datenpunkte in das Intervall $[x_0-d_k(x_0), x_0+d_k(x_0)]$. Also liegt es nahe

$$k = 2d_k(x_0)n\hat{f}(x_0)$$

zu setzen. Nach Umformung erhält man dann obige Gleichung.

Im Gegensatz zu den bisher betrachteten Histogrammen integriert $\hat{f}$ nicht zu eins. Dies liegt daran, dass auch außerhalb des Wertebereichs der ermittelte Schätzer ungleich Null ist und nur sehr langsam abfällt. Des Weiteren ist dieses Verfahren in hohen Dimensionen nur schwer anwendbar, da eine geeignete Abstandsfunktion angegeben werden muss.

3.2.3 Schätzer mit orthogonalen Reihen

Die auf Reihen von orthogonalen Basisfunktionen basierenden Schätzer funktionieren ähnlich wie Funktionsapproximationen durch Fourierbasen. In beiden Fällen wird durch Übergang zu einer endlichen Reihe das Problem diskretisiert.

Sei $\{\phi_j\}_{\nu=0}^{\infty}$ eine orthogonale Basis eines Funktionenraumes über $[0,1]$, in dem die gesuchte Funktion f liegt. Dann lässt sich jede Funktion f dieses Raumes schreiben als:

$$f(x) = \sum_{\nu=0}^{\infty} f_\nu \phi_\nu(x)$$

wobei die f_ν die üblichen Koeffizienten der Faltung mit der Basis ϕ_ν sind:

$$f_\nu = \int_0^1 f(x)\phi_\nu(x)dx \tag{3.2.3}$$

Bekanntestes Beispiel dafür sind die Fourierbasen:

$$\begin{aligned} \phi_0(x) &= 1 \\ \phi_{2r-1}(x) &= \sqrt{2}cos2\pi rx \\ \phi_{2r}(x) &= \sqrt{2}sin2\pi rx \\ r &= 1,2,... \end{aligned}$$

Sei X eine Zufallsvariable, dann lässt sich (3.2.3) als Erwartungswert interpretieren

$$f_\nu = E[\phi_\nu(X)],$$

und somit schätzen durch

$$\hat{f}_\nu = \frac{1}{n}\sum_{i=1}^{n}\phi_\nu(X_i).$$

Nun schneidet man die unendliche Summe $\sum_{\nu=0}^{\infty}\hat{f}_\nu\phi_\nu$ nach k Summanden ab und erhält so den Schätzer $\hat{f}$ für die unbekannte Dichte:

$$\hat{f}(x) = \sum_{\nu=0}^{k}\hat{f}_\nu\phi_\nu(x).$$

In diesem Fall ist der Punkt k, an dem die unendliche Reihe abgeschnitten wird, der Parameter der die Regularität der Lösung vorgibt. Für $k \to \infty$ entsteht eine Summe von Diracschen-Delta-Funktionen, also gerade wieder die empirische Dichtefunktion, während für $k = 0$ die Konstante Eins resultiert.

Interessant ist hierbei ein Zusammenhang zum Histogramm. Verwendet man als orthogonale Basisfunktionen Haar-Wavelets auf einem Diskretisierungsgitter, so ergibt sich daraus als Schätzer wieder ein Standardhistogramm. Die entsprechenden Basisfunktionen sind dann für $m = 2^k + j$, $j, k \in \mathbb{N}$

$$\phi_0(x) = 1 \tag{3.2.4}$$

$$\phi_r(x) = \begin{cases} 2^{k/2} & x \in \left(\frac{j-1}{2^k}, \frac{j-1/2}{2^k}\right) \\ -2^{k/2} & x \in \left(\frac{j-1/2}{2^k}, \frac{j}{2^k}\right) \end{cases}. \tag{3.2.5}$$

3.2.4 Kerndichteschätzer

Die Schätzer, denen in jüngster Zeit wohl die meiste Forschungsarbeit gewidmet wurde, sind die *Kerndichteschätzer*. Ein Vorteil dieser Verfahren ist, dass ihre Komplexität im Wesentlichen nur von der Anzahl an Datenpunkten abhängt. Der so genannte *Fluch der Dimension* lässt sich damit zu einem gewissen Grade umgehen. Andererseits sind die entstehenden Matrizen voll besetzt, so dass eine Lösung meistens nur in $O(n^3)$ erfolgen kann. Im ersten Abschnitt führen wir den so genannten *naiven Schätzer* ein, der eine leichte Abwandlung eines Histogramms darstellt. Anschließend gehen wir auf allgemeine Kernschätzer ein.

Der naive Schätzer

Sei wieder X eine stetige Zufallsvariable mit Dichte f. Dann folgt aus der Definition der Wahrscheinlichkeitsdichte 2.1.5 und (2.1.1):

$$f(x) = \lim_{h\to 0}\frac{1}{2h}P(x-h < X < x+h). \tag{3.2.6}$$

Wir schätzen $P(x-h < X < x+h)$ durch den Anteil an Werten, der im Intervall $(x-h < X < x+h)$ liegt.
Dazu führen wir eine Gewichtungsfunktion w ein:

$$w(x) = \begin{cases} \frac{1}{2} & \text{falls} \quad |x| < 1 \\ 0 & \text{sonst} \end{cases}. \tag{3.2.7}$$

Damit lässt sich der so genannte naive Schätzer formulieren:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} w(\frac{x - x_i}{h}). \tag{3.2.8}$$

Dies entspricht der Platzierung einer konstanten Funktion der Höhe $(2nh)^{-1}$ auf einem Intervall der Breite $2h$ um jeden Datenpunkt, die dann über alle Punkte aufsummiert werden. Der naive Schätzer hat somit gewisse Ähnlichkeit mit einem Histogramm und wird daher auch als *gleitendes Histogramm* bezeichnet. Im Unterschied zum Histogramm wird nun die Konstante direkt auf den Datenpunkten und nicht auf einem Intervall aufgehängt und die Wahl des Ursprungs entfällt somit.

Man kann den naiven Kernschätzer auch als numerische Approximation der Ableitung der kumulativen Verteilungsfunktion herleiten (siehe [Sco92]). Formal ist die Ableitung der empirischen Verteilungsfunktion eine Summe von Diracschen-Deltafunktionen, die als Dichteschätzer wenig geeignet ist. Geht man aber ähnlich wie bei zentralen finiten Differenzen vor und definiert den Schätzer $\hat{f}$ als

$$\begin{aligned} \hat{f}(x) &= \frac{F_n(x - \frac{h}{2}) - F_n(x + \frac{h}{2})}{h} \\ &= \frac{1}{nh} \sum_{i=1}^{n} \mathcal{X}_{(-\frac{1}{2}, \frac{1}{2}]} \left(\frac{x - x_i}{h} \right) \end{aligned}$$

mit empirischer Verteilungsfunktion F_n und Indikatorfunktion $\mathcal{X}$, so erhält man wiederum den naiven Schätzer. Analog zum Histogramm ist der so erhaltene Schätzer aber weder stetig noch differenzierbar und es bleibt das Problem der optimalen Wahl der Bandbreite h.

Allgemeine Kernschätzer

Nutzt man statt der stückweise konstanten Gewichtungsfunktion eine glattere Funktionen K mit $\int_{-\infty}^{\infty} K(x)dx = 1$, so erhält man die allgemeine Form eines Kerndichteschätzers:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x - x_i}{h} \right). \tag{3.2.9}$$

Oft wird für K ein Gauß-Kern gewählt. Das h entspricht dann der Standardabweichung. Einige häufig verwendete Kerne sind:

- $$K_{naiv}(x) = \frac{1}{2}, |x| \leq 1$$

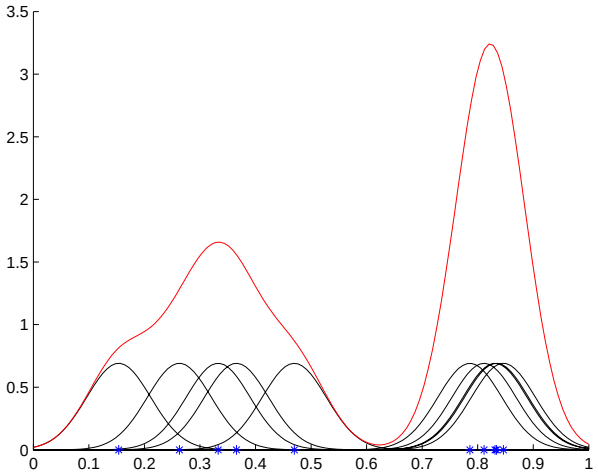

Abbildung 3.4: Beispiel eines Kernschätzers mit Gauß-Kern

- $$K_{linear}(x) = 1 - |x|, |x| \leq 1$$
- $$K_{gauss}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$
- $$K_{epanechnikov}(x) = \frac{3}{4}(1 - x^2), |x| \leq 1$$
- $$K_{cauchy}(x) = \frac{\pi}{1 + x^2}$$
- $$K_{picard}(x) = \frac{1}{2}\exp(-|x|)$$

Die Kerne unterscheiden sich vor allem auch in ihrem Träger. Beispielsweise der Gauß-Kern hat den Nachteil, dass er einen unendlichen Träger hat. Der Aufwand bei der numerischen Berechnung ist demnach sehr hoch, da jeder Datenpunkt auf die Lösung globalen Einfluss hat. Besser geeignet sind daher Kerne mit endlichem Träger. In der Praxis werden Gauß-Kerne ab einem bestimmten Schwellenwert abgeschnitten und somit eine weitere Diskretisierung des Problems vorgenommen. Nach Normierung integriert der resultierende Kernschätzer dann wieder zu Eins. Andere Methoden beschleunigen mit Hilfe der schnellen Gauß-Transformation die Berechnung des Kernschätzers für Gauß-Kerne (siehe [EDD03]).

Als Erweiterung gibt es Verfahren, welche die Bandbreite h adaptiv wählen. An Stellen mit vielen Datenpunkten werden die Kerne entsprechend schmal gewählt, wohingegen sie in Gebieten mit wenigen Datenpunkten breiter gewählt werden (siehe z.B. [KS00],[KS02],[Tur]).

3.2.5 Verallgemeinerung der Kernmethode

Mit einer erstaunlich eleganten Definition eines nichtparametrischen Schätzers, die auf TERREL und SCOTT zurück geht, lassen sich nahezu sämtliche Verfahren der Dichteschätzung als Kernschätzer interpretieren. Die Autoren zeigen in ihrer Arbeit [TS92], dass sich alle Schätzer, ob parametrisch oder nichtparametrisch, als *verallgemeinerte Kernschätzer* schreiben lassen. Bei parametrischen Schätzern haben die Datenpunkte globalen Einfluss auf die Dichte, während bei nichtparametrischen Schätzern der Einfluss lokal ist. So hat zum Beispiel das verallgemeinerte Histogramm in Gleichung (3.2.2) bereits die Form eines Kernschätzers. Auch das Verfahren der k-nächsten Nachbarn lässt sich als Kernschätzer interpretieren. Dabei ist die Bandbreite der datenzentrierten Kernfunktionen abhängig von der Lage der nächsten Nachbarn, also adaptiv zu wählen.

Analog lassen sich alle Verfahren, auch solche, die als Optimierungsproblem eines zumeist quadratischen Funktionals geschrieben werden, als verallgemeinerte Kernschätzer interpretieren. Dabei heißt verallgemeinert, dass sowohl die Bandbreite, als auch die Kernfunktionen selber adaptiv gewählt werden.

Satz 3.2.2 (General Kernel Theorem)**.** *Jeder Dichteschätzer der ein stetiges und Gâteaux-differenzierbares Funktional der empirischen Verteilungsfunktion ist, lässt sich schreiben als*

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, x_i, F_n), \tag{3.2.10}$$

wobei K die Gâteaux Ableitung von $\hat{f}$ unter Variation der x_i ist.

Beweis. Die empirische Verteilungsfunktion aus (2.2.1) lautet

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_{[x_i,\infty)}(x). \tag{3.2.11}$$

Schreibe den Dichteschätzer als Operator $\hat{f}(x) = T_x\{F_n\}$. Wir definieren

$$\begin{aligned} K(x, y, F_n) &\equiv \lim_{\epsilon\to 0}\frac{1}{\epsilon}[T_x\{(1-\epsilon)F_n + \epsilon\mathcal{X}_{[y,\infty)}\} - (1-\epsilon)T_x\{F_n\}] \\ &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}[T_x\{F_n + \epsilon(\mathcal{X}_{[y,\infty)} - F_n)\} - T_x\{F_n\}] + T_x\{F_n\} \\ &= DT_x(F_n)[\mathcal{X}_{[y,\infty)} - F_n] + \hat{f}(x), \end{aligned}$$

wobei $DT(\phi)[\eta]$ die Gateâux Ableitung von T an ϕ in Richtung η ist. Da die Gateâux Ableitung im zweiten Argument linear ist, gilt

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} K(x, x_i, F_n) &= \frac{1}{n}\sum_{i=1}^{n} DT_x(F_n)[\mathcal{X}_{[x_i,\infty)} - F_n] + \hat{f}(x) \\ &= DT_x(F_n)\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{X}_{[x_i,\infty)} - F_n\right] + \hat{f}(x) \\ &= 0 + \hat{f}(x). \end{aligned}$$

Die letzte Gleichheit ergibt sich wegen

$$DT(\phi)[0] = 0.$$

□

Abschließend lässt sich sagen, dass sich eine Vielzahl von Verfahren der Dichteschätzung als Kernschätzer schreiben lässt, wobei durch Wahl des Kerns der Funktionenraum, aus dem die Lösung stammt, vorgegeben wird.

4 Ein ecdf-basiertes Verfahren

Wie bereits in Kapitel 2 beschrieben, lässt sich die empirische Verteilungsfunktion (ecdf) theoretisch wesentlich besser handhaben als die empirische Dichtefunktion (epdf). Hauptgrund dafür ist die stückweise Stetigkeit der ecdf, die es zumindest erlaubt eine L^2-Norm anzuwenden. Diese Eigenschaft machen wir uns in dem hier vorgestellten Verfahren zunutze.

Im ersten Abschnitt stellen wir zunächst das Verfahren von Hegland, Hooker und Roberts vor, welches die Grundlage für unseren Ansatz war. Dieser nichtparametrische Schätzer bestimmt direkt aus den Daten eine Approximation der Dichtefunktion als Linearkombination einer geeigneten Funktionenraumbasis. Anschließend gehen wir auf ein von Vapnik und Mukherjee in [MV99] sowie [WGS+99] vorgeschlagenes Verfahren ein, welches mit der empirischen Verteilungsfunktion arbeitet. Die Werte der ecdf an den Datenpunkten x_i werden dabei als Zielvariable y_i interpretiert und anschließend ein Support-Vektor-Verfahren angewendet, welches eine Approximation der Verteilungsfunktion als Summe von Kernfunktionen auf den Datenpunkten bestimmt. Ein von uns ecdf-basiertes Verfahren genannter Ansatz, welches eine Kombination gewisser Anteile dieser beiden Schätzer darstellt, wird dann eingeführt und erste Ergebnisse damit präsentiert.

4.1 Der Ansatz von Hegland, Hooker & Roberts

Die Motivation des von Hegland, Hooker und Roberts entwickelten Dichteschätzers ist folgende: Angenommen wir haben einen Schätzer f_ε aus den Daten $\{x_i\}_{i=1}^n$ gegeben, welcher in L^2 liegt. Ähnlich wie bei Spline Glättern (siehe [Wah90]) oder allgemeinen Regularisierungsnetzwerken suchen wir nun eine Approximation $\hat{f}$, die das Funktional

$$J_2(u) = \int_\Omega (u(x) - f_\varepsilon(x))^2 dx + \lambda \|\mathcal{L}u(x)\|_{L^2} \tag{4.1.1}$$

für einen gegebenen Differentialoperator $\mathcal{L}$ minimiert. In welcher Form das f_ε vorliegt, wird dabei nicht genauer festgelegt. Eine Möglichkeit wäre ein einfaches Histogramm der gegebenen Daten.

Der zweite Term aus (4.1.1) kontrolliert die Glattheit der gesuchten Funktion, während der erste Teil den Fehler zum bereits gegebenen Schätzer f_ε ausdrückt. Mit dem Parameter λ kann man die Gewichtung der einzelnen Terme balancieren. Analog zur Bandbreite h bei Kernschätzern agiert das λ in dieser Darstellung also als Glättungsparameter.

Diese Form erinnert damit stark an Zielfunktionale, wie sie in der Regularisierungstheorie unter dem Namen *Tikhonov Regularisierung* auftauchen, so

zum Beispiel bei der Funktionsrekonstruktion oder Regression, sowie bei der Klassifikation oder Clusteranalyse. Allgemein geht es jeweils darum ein schlecht gestelltes Problem durch hinzufügen eines Glättungsterms zu regularisieren und damit eindeutig lösbar zu machen.

Die Variationsgleichung für dieses Problem lautet nun:

$$\int_\Omega s(x)(u(x) - f_\varepsilon(x))dx + \lambda C(u,s) = 0, \quad \forall s \in H^2(\Omega).$$

oder

$$\int_\Omega u(x)s(x)dx + \lambda C(u,s) = \int_\Omega s(x)f_\varepsilon(x)dx. \tag{4.1.2}$$

Dabei ist C eine stetige Bilinearform auf einem Hilbertraum $H \subset L^2(\Omega)$. Üblicherweise wählt man

$$C(u,s) = \int_\Omega \mathcal{L}u(x) \cdot \mathcal{L}s(x)dx$$

für einen Differentialoperator $\mathcal{L}$. Auf der rechten Seite von (4.1.2) steht eine Art Erwartungswertoperator. Ein naheliegender Ansatz ist daher, diesen durch eine Mittelwertbildung zu approximieren:

$$\int_\Omega s(x)f_\varepsilon(x)dx = E(s) \approx \frac{1}{M}\sum_{i=1}^{M} s(x_i).$$

Damit ergibt sich die Variationsform für die optimale Lösung $\hat{f}$:

$$\int_\Omega \hat{f}(x)s(x)dx + \lambda C(\hat{f},s) = \frac{1}{M}\sum_{i=1}^{M} s(x_i), \quad \forall s \in V \tag{4.1.3}$$

$$\Leftrightarrow \langle \hat{f}, s\rangle_{L_2} + \lambda C(\hat{f},s) = \frac{1}{M}\sum_{i=1}^{M} s(x_i), \quad \forall s \in V. \tag{4.1.4}$$

Hierbei stellt man eine grundlegende Unsauberkeit des Verfahrens fest. So wie $E(s)$ definiert wurde, ist gerade die empirische Dichtefunktion, also die Summe von Dirac'schen Deltafunktionen, als f_ε gewählt worden, für die eine L^2-Norm nicht existiert. Daher wird diese erst nach Minimierung in der schwachen Form eingesetzt. Um dieses Problem zu umgehen, werden wir später mit der empirischen Verteilungsfunktion arbeiten, die stückweise stetig und integrierbar ist.

Die Autoren erwähnen außerdem in ihrer Arbeit, dass man diesen Ansatz auch als Kernschätzer interpretieren kann. Wählt man zum Beispiel als Grundgebiet $\Omega = \mathbb{R}$ und als Differentialoperator die 2. Ableitung, so kann man aus Gleichung (4.1.4) die Euler Gleichungen ableiten als:

$$u(x) + \lambda u^{(4)}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x).$$

Falls man eine Greensche Funktion g hat, die

$$g(x) + g^{(4)}(x) = \delta_0(x)$$

erfüllt, kann man die Fundamentallösungen schreiben als:

$$\hat{f}(x) = \frac{1}{n\lambda^{1/4}} \sum_{i=1}^{n} g\left(\frac{x - x_i}{\lambda^{1/4}}\right).$$

Als Approximation erhält man dann gerade einen Kernschätzer mit Kern

$$g(x) = \frac{1}{\sqrt{2}} \exp\left(\frac{-|x|}{\sqrt{2}}\right)\left[cos\left(\frac{|x|}{\sqrt{2}}\right) + sin\left(\frac{|x|}{\sqrt{2}}\right)\right].$$

Dies lässt sich auch für allgemeinere Differentialoperatoren und beschränkte Gebiete Ω zeigen. Grundlage ist hier die Beziehung zwischen Regularisierungsoperatoren und Hilberträumen mit reproduzierendem Kern, wie sie in Kapitel 2.4 bereits vorgestellt wurde.
Zur numerischen Lösung des obigen Systems wird eine Finite-Elemente-Approximation angewendet. Sei $\{\varphi_i\}_{i=1}^{N}$ eine Basis eines Teilraumes $V_N \subset V$. Wir schreiben die gesuchte Lösung als $\hat{f}(x) = \sum_{j=1}^{N} \alpha_j \varphi_j(x)$. Damit ergeben sich die Gleichungen $k = 1, ..., N$:

$$\sum_{j=1}^{N} \alpha_j \langle \varphi_k, \varphi_j \rangle_{L^2} + \lambda \sum_{j=1}^{N} \alpha_j C(\varphi_k, \varphi_j) = \frac{1}{M} \sum_{i=1}^{M} \varphi_j(\boldsymbol{x}_i)$$

bzw. in Matrixschreibweise:

$$(\mathcal{M} + \lambda \mathcal{C})\boldsymbol{\alpha} = \frac{1}{M} \mathcal{B}^T \mathbf{1},$$

wobei

$$\mathcal{M}_{i,j} = \langle \varphi_i, \varphi_j \rangle_{L^2}, \quad i, j = 1, ..., N$$

Massenmatrix,

$$\mathcal{C}_{i,j} = \langle \mathcal{L}\varphi_i, \mathcal{L}\varphi_j \rangle_{L^2}, \quad i, j = 1, ..., N$$

Regularisierungsmatrix und

$$\mathcal{B}_{i,j} = \varphi_j(\boldsymbol{x}_i), \quad i = 1, ..., M \quad j = 1, ..., N$$

Datenmatrix genannt werden. Es ensteht also nach Minimerung des quadratischen Funktionals $J(u)$ ein lineares Gleichungssystem, welches mit geeigneten Verfahren zu lösen ist. Dieses ist dünn besetzt und kann daher, falls V die Dimension $d << n$ hat, in $O(n + d)$ Operationen gelöst werden. Als Basisfunktionen werden B-Splines vorgeschlagen.

4.2 Der Ansatz von Vapnik und Mukherjee

Der in [MV99] beschriebene Ansatz basiert auf Support-Vektor-Maschinen, wie sie von Vapnik in [Vap00] vorgestellt wurden. Daher wollen wir zu Beginn eine kleine Einführung in die Theorie der Support-Vektor-Maschinen (SVM) geben, wobei wir nicht die Notation von Vapnik verwenden werden, sondern stattdessen einen auf Regularisierungsnetzwerken basierenden Ansatz zur Motivation der SVM anführen.

4.2.1 Support-Vektor-Maschinen

Support-Vektor-Maschinen werden seit geraumer Zeit für die Lösung diverser Probleme des maschinellen Lernens eingesetzt.

Das zu lösende Problem ist hier ein klassisches Regressionsproblem. Es soll also aus gegebenen Trainingsdaten

$$D = (x_1, y_1), ..., (x_n, y_n) \quad x_i \in X, y_i \in Y$$

eine Funktion bestimmt werden, die den angenommenen funktionalen Zusammenhang zwischen X und Y möglichst gut wiedergibt. Es wird dabei vorausgesetzt, dass sich f schreiben lässt als

$$f(x) = \sum_{j=1}^{\infty} c_j \phi_j(x) + b, \tag{4.2.1}$$

wobei $\{\phi_j(x)\}_{j=1}^{\infty}$ eine Menge von linear unabhängigen Basisfunktionen ist. Die Parameter c_j und b sind aus den Daten zu bestimmen. Ohne weitere Voraussetzungen an f ist das Problem schlecht gestellt, sodass der übliche Ansatz der Regularisierungstheorie, einen zusätzlichen Glättungsterm auf das Funktional zu addieren (*Tikhonov-Regularisierung*), verwendet wird. Es entsteht ein zu minimierendes Funktional der Form

$$\min_{f \in \mathcal{H}} R(f) = C \sum_{i=1}^{n} V_\epsilon(y_i - f(x_i)) + \frac{1}{2}\Phi(f). \tag{4.2.2}$$

Vapnik schlägt an dieser Stelle die Wahl der so genannten *ϵ-insensitiven Kostenfunktion*

$$V_\epsilon(x) = \begin{cases} 0, & \text{falls} \quad |x| < \epsilon \\ |x| - \epsilon & \text{sonst} \end{cases} \tag{4.2.3}$$

vor. Damit erhält man ein Approximations-Schema, welches *Support-Vektor-Maschine* genannt wird:

$$f(x, \alpha, \alpha^*) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(x, x_i) + b, \tag{4.2.4}$$

wobei α_i^* und α_i positive Koeffizienten sind, die das folgende quadratische Programm lösen:

$$\min_{\alpha^*, \alpha} R(\alpha^*, \alpha) = \epsilon \sum_{i=1}^{n} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{n} y_i (\alpha_i^* - \alpha_i) + \frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) \tag{4.2.5}$$

mit den Nebenbedingungen

$$\begin{aligned} & 0 \leq \alpha^*, \alpha \leq C \\ & \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) = 0 \\ & \alpha_i \alpha_i^* = 0 \quad \forall i = 1, ..., n \end{aligned}$$

und einer symmetrischen Kernfunktion

$$K(x,y) = \sum_{j=1}^{\infty} \gamma_j \phi_j(x) \phi_j(y).$$

Dies ist gerade der reproduzierende Kern eines reproduzierenden Kern-Hilbertraumes. Es fällt dabei auf, dass der Parameter b nicht mehr auftaucht, da er nach Kenntnis der α_i und α_i^* ermittelt werden kann. Aufgrund der Struktur des Minimierungsproblems werden nur einige der Koeffizienten $(\alpha_i^* - \alpha_i)$ ungleich Null sein. Die zugehörigen Datenpunkte x_i sind so genannte *Support-Vektoren*. Wie viele Support-Vektoren es gibt, hängt vom ϵ in der Kostenfunktion und dem Regularisierungsparameter C ab. Falls kein Rauschen über den Daten liegt, ist das optimale C unendlich.

Für den Fall der Klassifikation kann man sich den Support-Vektor-Ansatz sehr anschaulich als eine Hyperebene vorstellen, welche die beiden Klassen trennt. Die Support-Vektoren sind dabei die Datenpunkte, welche die Hyperebene charakterisieren. Mit Hilfe der Kernfunktion kann man allgemeine nichtlineare Trennflächen erzeugen. Dieses Vorgehen wird auch als *Kerntrick* bezeichnet.

4.2.2 Dichteschätzung mit SVM

Um die Methodik der Support-Vektor-Maschinen auf die Dichteschätzung anwenden zu können, muss das Problem zunächst in eine Regressionsaufgabe umformuliert werden. Dazu betrachtet man die Verteilungsfunktion

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{4.2.6}$$

der gesuchten Dichtefunktion f.

Eine gute Approximation der Verteilungsfunktion ist bekanntlich die empirische Verteilungsfunktion

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_{(-\infty,x]}(x_i),$$

mit Indikatorfunktion $\mathcal{X}$.

Die Konvergenz von F_n gegen F für $n \to \infty$ ist asymptotisch bekannt (siehe Satz 2.2.2). Insbesondere hat die Zufallsvariable F_n an einem festen Punkt x die Standardabweichung

$$\sigma = \sqrt{\frac{1}{n} F_n(x)(1 - F_n(x))}.$$

Also verwendet man $(x_1, F_n(x_1)), ..., (x_n, F_n(x_n))$ als Eingangswerte für das Regressionsproblem. Darüber hinaus kann statt eines festen ϵ für die Kostenfunktion V_ϵ, an jedem Datenpunkt ein gesondertes ϵ_i verwendet werden. Dies hat die Gestalt

$$\epsilon_i = \lambda \sigma_i = \lambda \sqrt{\frac{1}{n} F_n(x_i)(1 - F_n(x_i))},$$

wobei σ_i gerade die vorab bekannte Standardabweichung der Verteilungsfunktion an dem Punkt x_i ist.

Es werden also die Tripel

$$(x_1, F_n(x_1), \epsilon_1), ..., (x_n, F_n(x_n), \epsilon_n)$$

aufgestellt und anschließend wird damit ein Standard-SVM-Verfahren durchgeführt. Analog zu den bereits vorgestellten Kernschätzern ist hier noch ein geeigneter Kern zu wählen, der wiederum die Glattheit der Lösung vorgibt.

Grob gesagt ist das Ziel eine möglichst glatte Funktion zu finden, die innerhalb eines „ϵ_i-Schlauches“ um die empirische Verteilungsfunktion liegt.

4.3 Das ecdf-basierte Verfahren

Im Folgenden betrachten wir das Verfahren aus dem ersten Unterkapitel in einer analog zum gerade beschriebenen Verfahren einmal „aufintegrierten“ Variante. Wir ermitteln also zunächst eine Approximation an die Verteilungsfunktion der gegebenen Daten. Da wir diese als Linearkombination von Basisfunktionen eines geeigneten Funktionenraumes schreiben, ergibt sich die Dichtefunktion als Linearkombination der abgeleiteten Basisfunktionen. Es handelt sich also hier analog zum Verfahren von Hegland, Hooker und Roberts um ein Verfahren, welches die Lösung aus einem diskretisierten Funktionenraum ermittelt. Der Ansatz aus dem vorherigen Abschnitt bestimmt im Gegensatz dazu die Koeffizienten einer Kerndarstellung auf einer Teilmenge der Datenpunkte.

Mit gegebener Datenmenge $D = \{x_i\}_{i=1}^n$ stellen wir zunächst die empirische Verteilungsfunktion

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_{(-\infty,x]}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_{[x_i,\infty)}(x)$$

auf. Diese ist Ausgangspunkt für unser Zielfunktional

$$\min_{u\in V} R(u) = \|u(x) - F_n(x)\|_{L^2}^2 + \lambda\|\mathcal{L}u(x)\|_{L^2}^2, \tag{4.3.1}$$

welches über alle u aus einem geeigneten Funktionenraum V definiert ist. Auch hier haben wir wieder einen Regularisierungsparameter $\lambda > 0$, der die Balance zwischen Glättung mittels Differentialoperator $\mathcal{L}$ und Genauigkeit an die Daten steuert. Wir suchen also eine Funktion, die zum einen möglichst nah an der empirischen Verteilungsfunktion liegt und dabei gewisse Glattheitsvorraussetzungen erfüllt. In den meisten Fällen werden diese durch $\mathcal{L} = \nabla$ oder $\mathcal{L} = \Delta$ gegeben.

Wir stellen u bezüglich einer unendlichen Funktionenraumbasis $\{\varphi_j\}_{j=1}^{\infty}$ dar

$$u(x) = \sum_{j=1}^{\infty} \alpha_j \varphi_j(x),$$

setzen dies in Gleichung (4.3.1) ein, und erhalten

$$\begin{aligned} R(u) &= \int (u(x) - F_n(x))^2 dx + \lambda \int (\mathcal{L}u(x))^2 dx \\ &= \int \left(\sum_{j=1}^{\infty} \alpha_j \varphi_j(x) - \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_{[x_i,\infty)}(x) \right)^2 dx + \lambda \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha_i \alpha_j \langle \mathcal{L}\varphi_i, \mathcal{L}\varphi_j \rangle. \end{aligned}$$

Differentiation in den Koeffizienten α_k, $(k = 1, ..., \infty)$ ergibt jeweils

$$\begin{aligned} \frac{\partial R}{\partial \alpha_k} &= 2 \int \left(\sum_{j=1}^{\infty} \alpha_j \varphi_j(x) - \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_{[x_i,\infty)}(x) \right) \varphi_k(x) dx + 2\lambda \sum_{j=1}^{\infty} \alpha_j \langle \mathcal{L}\varphi_j, \mathcal{L}\varphi_k \rangle \\ &= 2 \sum_{j=1}^{\infty} \alpha_j \langle \varphi_j, \varphi_k \rangle - \frac{2}{n} \sum_{i=1}^{n} \int \varphi_k(x) \mathcal{X}_{[x_i,\infty)}(x) dx + 2\lambda \sum_{j=1}^{\infty} \alpha_j \langle \mathcal{L}\varphi_j, \mathcal{L}\varphi_k \rangle \\ &= 2 \sum_{j=1}^{\infty} \alpha_j \Big(\langle \varphi_j, \varphi_k \rangle + \lambda \langle \mathcal{L}\varphi_j, \mathcal{L}\varphi_k \rangle \Big) - \frac{2}{n} \sum_{i=1}^{n} \int_{x_i}^{\infty} \varphi_k(x) dx. \end{aligned}$$

Nach einigen Umformungen resultiert das Gleichungssystem $(k = 1, ..., \infty)$

$$\sum_{j=1}^{\infty} \alpha_j \Big(\langle \varphi_j, \varphi_k \rangle + \lambda \langle \mathcal{L}\varphi_j, \mathcal{L}\varphi_k \rangle \Big) = \frac{1}{n} \sum_{i=1}^{n} \int_{x_i}^{\infty} \varphi_k(x) dx.$$

In Matrixschreibweise mit unendlich dimensionalen Matrizen und Vektoren erhält man:

$$(\mathcal{M} + \lambda \mathcal{C}) \boldsymbol{\alpha} = \boldsymbol{r}.$$

Hierbei ist $\mathcal{M}$ wieder die *Massenmatrix* $\mathcal{M}_{j,k} = \langle \varphi_j, \varphi_k \rangle, \quad j,k = 1, ..., \infty$ und $\mathcal{C}$ die positiv semidefinite *Regularisierungsmatrix* $\mathcal{C}_{j,k} = \langle \mathcal{L}\varphi_j, \mathcal{L}\varphi_k \rangle$ für $j,k = 1, ..., \infty$. Der Vektor auf der rechten Seite beinhaltet die Information aus den Daten und besteht aus den Komponenten

$$\boldsymbol{r}_k = \frac{1}{n} \sum_{i=1}^{n} \int_{x_i}^{\infty} \varphi_k(x) dx \quad k = 1, ..., \infty.$$

Wir haben also ein lineares Gleichungssystem für die Koeffizienten der gesuchten Funktion u aufgestellt. Dieses stimmt auf der linken Seite mit den Matrizen aus dem Verfahren von Hegland, Hooker und Roberts überein. Lediglich die rechte Seite enthält nun die Information aus den Daten nicht mehr in Form einer empirischen Dichtefunktion, sondern einer empirischen Verteilungsfunktion.

Zur numerischen Lösung des Systems ist eine diskrete Darstellung im Rechner nötig. Man wählt dazu einen endlich dimensionalen Funktionenraum $V_N \subset V$, der durch einen endlichen Satz von Funktionen dargestellt wird. Das heißt jede Funktion aus V_N, und damit auch die gesuchte Lösung u, lassen sich in der Form $u(x) = \sum_{i=j}^{N} \alpha_j \varphi_j(x)$ darstellen. Mit dieser endlichen Basis erhalten wir daraus ein Gleichungssystem mit endlichen Matrizen und Vektoren, welches wir

mit einem iterativen Verfahren - beispielsweise dem CG-Verfahren - lösen. Als Resultat erhalten wir die Koeffizienten $\boldsymbol{\alpha}$ einer Approximation der Verteilungsfunktion der gegebenen Daten. Da wir aus dieser Approximation eine zugehörige Dichtefunktion ableiten wollen, setzen wir voraus, dass die Basisfunktionen $\{\varphi_j\}_{j=1}^N$ differenzierbar sind. Damit schreiben wir unseren Dichteschätzer als Linearkombination der errechneten Koeffizienten mit den abgeleiteten Basisfunktionen:

$$\hat{f}(x) = \sum_{i=j}^{N} \alpha_j \varphi_j'(x).$$

4.4 Numerische Ergebnisse

In unseren Experimenten haben wir als Basisfunktionen B-Splines verschiedener Ordnung verwendet, wie sie in Kapitel 3.2.1 vorgestellt wurden. Diese werden auf einem äqudistanten Gitter mit (2^l+1) Punkten aufgestellt, wobei l das Diskretisierungslevel bezeichnet. Um eine Partition der Eins zu erhalten, sind am Rand je nach Ordnung der B-Splines zusätzliche Punkte nötig. Wenn p die Ordnung der B-Splines bezeichnet, benötigt man also insgesamt $(2^l + p + 1)$ Basisfunktionen. Gleichzeitig gibt uns das die Größe der entstehenden Matrix vor, welche $((2^l+p+1)^d)^2$ Einträge hat, falls d die Dimension des Problems bezeichnet. Damit die resultierende Dichte stetig ist, sind mindestens quadratische B-Splines für das Aufstellen der Verteilungsfunktion erforderlich. Der gleiche Effekt tritt bei den Glattheitsvoraussetzungen auf. Man hat diese für das Ausgangsproblem, die Bestimmung einer Verteilungsfunktion, eine Ordnung höher zu wählen.

Bevor erste Ergebnisse mit dem beschriebenen Verfahren präsentiert werden, sollte man sich Gedanken über geeignete Testdatensätze machen.

4.4.1 Simulieren von Daten

Wenn die Güte von Funktionsrekonstruktionen untersucht werden soll, gibt man in der Regel eine zu rekonstruierende Funktion f vor und wählt als Datenmenge ein äquidistantes Gitter $(x_1, ..., x_N)$ mit zugehörigen Funktionswerten $f(x_1), ..., f(x_N)$. Die Aufgabe ist eine passende Interpolierende zu finden. Ein ähnliches Verfahren verwenden wir hier auch, nur dass bei der Dichteschätzung die Information über die Funktion aus der Lage der Datenpunkte hervorgehen muss. Dazu geben wir eine Verteilungsfunktion vor, dessen Dichtefunktion rekonstruiert werden soll. Wir wählen ein äquidistantes Gitter auf dem Intervall $(0, 1)$ und bilden die Inverse der Verteilungsfunktion auf diesen Punkten.

Analog arbeitet die so genannte *Inversionsmethode*, mit der aus gleichverteilten Daten Zufallszahlen erzeugt werden, die bezüglich einer beliebigen Verteilung verteilt sind. Im Unterschied dazu geben wir die gleichverteilten Zufallszahlen als äquidistantes Gitter vor. Veranschaulicht wird dies in Abbildung 4.4.1 für den Fall der Standardnormalverteilung. Auf der Ordinate sind gleichverteilte Punkte zwischen 0 und 1 abgebildet. Die Punkte auf der Abszisse sind bezüglich der Standardnormalverteilung gleichmäßig verteilte Daten. Solche bezüglich einer vorgegebenen Verteilung gleichmäßig verteilte Daten werden für die folgen-

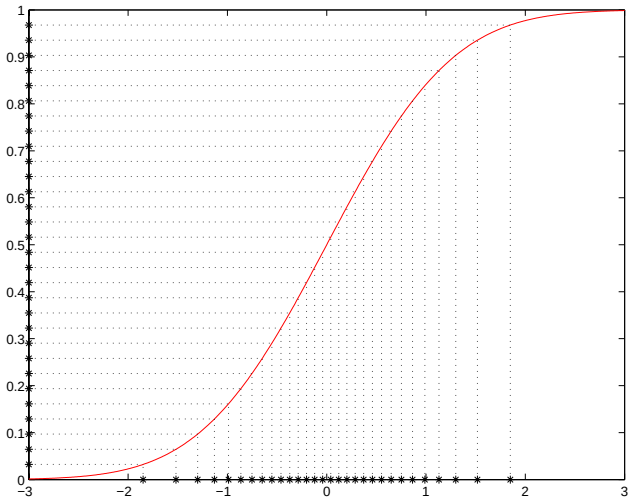

Abbildung 4.1: bzgl. der Normalverteilung gleichverteilte Daten

den Untersuchungen verwendet. Auf diese Weise erhält man Datenpunkte, aus denen ein Dichteschätzungsverfahren die Originaldichte sehr gut rekonstruieren können sollte, da sie keinen stochastischen Fehler enthalten.

4.4.2 Synthetischer Datensatz

Um ein erstes Ergebnis mit der hier vorgestellten Methode zu bekommen, wurde ein Datensatz aus zwei überlagerten Normalverteilungen erzeugt. In Abbildung 4.3 sind auf Level 5 mit kubischen B-Splines als Basis und Regularisierungsparameter λ für vier verschieden große Datensätze die resultierenden Dichteschätzer sowie die Originaldichte abgebildet. Als Regularisierungsoperator wählen wir die zweite Ableitung. Zusätzlich sind auf der Abszisse jeweils die Datenpunkte abgebildet, die gemäß der gewählten Verteilung gleichmäßig verteilt sind. Die resultierenden Gleichungssysteme werden mit Hilfe eines CG-Verfahrens gelöst.

Abbildung 4.2 stellt den mittleren quadratischen Fehler (MSE) auf den Gitter- bzw. Datenpunkten in Abhängigkeit vom Diskretisierungslevel für verschieden große Datenmengen dar. In der oberen Reihe ist $\lambda = 0.1$ gewählt worden, in der unteren $\lambda = 1$. Man sieht hier recht gut das Wechselspiel von Regularisierungsparameter und Diskretisierungslevel. Je höher das Level gewählt wird, desto höher ist auch der Regularisierungsparameter zu wählen. Begründet ist dies durch die veränderten Ansatzfunktionen, die ebenfalls deutlichen Einfluss auf die Regularität der Lösung haben. Während zum Beispiel auf 800 Datenpunkten mit $\lambda = 0.1$ der Fehler bei einer Diskretisierungstiefe vom Level 7 am geringsten ist, verschiebt sich dies zu Level 8 falls $\lambda = 1$ gesetzt wird. Das Problem ist, dass bei höherem Level die Träger der Ansatzfunktionen schmaler werden und damit die resultierende Dichte stärker oszilliert. Daher muss diesem Effekt durch eine entsprechend stärkere Glättung entgegengewirkt werden. Andererseits darf bei geringerem Level λ nicht zu groß sein. Dies sieht man recht gut am Fehler auf dem Gitter für die verschiedenen Glättungsparameter bei Level 3. Für $\lambda = 0.1$ fällt hier der Fehler wesentlich kleiner aus.

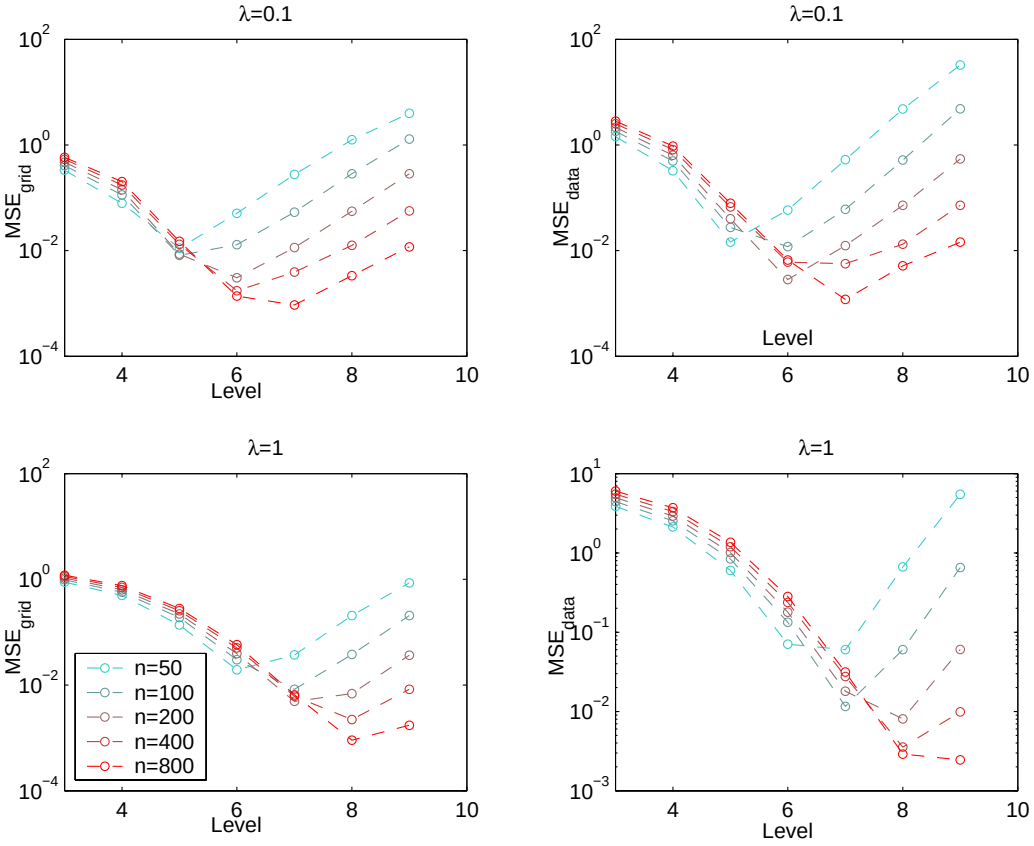

Abbildung 4.2: mittlerer quadratischer Fehler auf den Gitter- bzw. Datenpunkten in Abhängigkeit vom Diskretisierungslevel mit verschiedener Anzahl an Datenpunkten.

Level l	$\sqrt{\frac{MSE^{l-1}_{grid}}{MSE^{l}_{grid}}}$	$\sqrt{\frac{MSE^{l-1}_{data}}{MSE^{l}_{data}}}$	$\sqrt{\frac{MSE^{l-1}_{grid}}{MSE^{l}_{grid}}}$	$\sqrt{\frac{MSE^{l-1}_{data}}{MSE^{l}_{data}}}$
4	1.6973	1.7247	1.2574	1.2747
5	3.6734	3.4815	1.6376	1.6498
6	3.3011	3.4409	2.2055	2.2016
7	1.2106	2.3603	2.9836	2.9957
8	0.5288	0.4818	2.6795	3.2847
9	0.5332	0.5960	0.7213	1.0878

Tabelle 4.1: Konvergenzverhalten des ecdf-basierten Verfahrens bei 800 Datenpunkten und $\lambda = 0.1$ (links) bzw. $\lambda = 1$ (rechts). Dargestellt ist die relative Änderung des L^2-Fehlers.

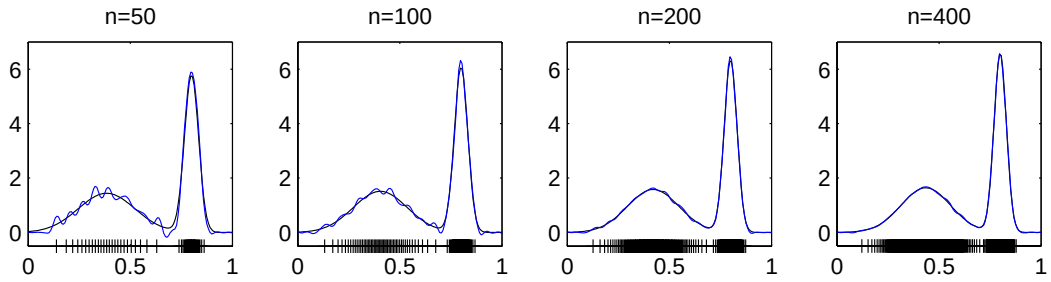

Abbildung 4.3: Dichteschätzung auf verschieden großen simulierten Datensätzen mit Diskretisierungslevel 5 und $\lambda = 0.01$. Die Originaldichte ist jeweils gestrichelt gezeichnet. Auf der Abszisse sind die Datenpunkte markiert.

In Tabelle 4.1 ist die relative Änderung des L^2-Fehlers wiederum sowohl auf den Datenpunkten, als auch auf dem Diskretisierungsgitter berechnet worden. Die bei Interpolationsaufgaben mit quadratischen Splines zu erwartende Ordnung von h^3 ist dabei bis zu einem gewissen Level zu beobachten. Anzumerken ist noch, dass die Verteilungsfunktion zwar mit kubischen Splines aufgestellt wird, die Basisfunktionen der Dichtefunktion aber dementsprechend nur quadratische Ordnung haben. Eine höhere Konvergenzrate kann hier also nicht erreicht werden.

Zusätzlich hat noch die Datenmenge n einen Einfluss auf die optimale Wahl von λ. Je mehr Daten man hat, desto geringer sollte λ gewählt werden. Dies liegt daran, dass bei größer werdender Datenmenge die Lücken zwischen benachbarten Datenpunkten, über die man sonst hinweg glättet, immer geringer werden. Dadurch bekommt man eine höhere Informationsdichte, die man sich nicht durch eine zu starke Glättung zerstören sollte.

4.4.3 Old Faithful Geyser Datensatz

Als nächstes Beispiel ist in Abbildung 4.4 eine eindimensionale Dichteschätzung des bekannten Old Faithful Geyser Datensatzes gemacht worden. Der hier verwendete Datensatz besteht aus 107 Werten, die jeweils die Wartezeit in Minuten zwischen zwei Ausbrüchen des Geysirs im Yellowstone Nationalpark in den USA darstellen. Um die Vorgehensweise noch einmal besser zu verdeutlichen, ist in der oberen Grafik von Abbildung 4.4 die empirische Verteilungsfunktion und die daraus erstellte Approximation abgebildet. Unten sind die einzelnen Datenpunkte und die approximierte Dichtefunktion eingezeichnet. Der Regularisierungsparameter wurde hier als $\lambda = 0.01$ gewählt mit der zweiten Ableitung als Differentialoperator. Als Ansatzfunktionen des Funktionenraums wurden kubische B-Splines über einem äquidistanten Gitter der Diskretierungsstufe 5 gewählt.

In Kapitel 7 werden zum Vergleich des Maximum a-posteriori-Ansatz mit diesem und anderen Verfahren weitere Ergebnisse dargestellt.

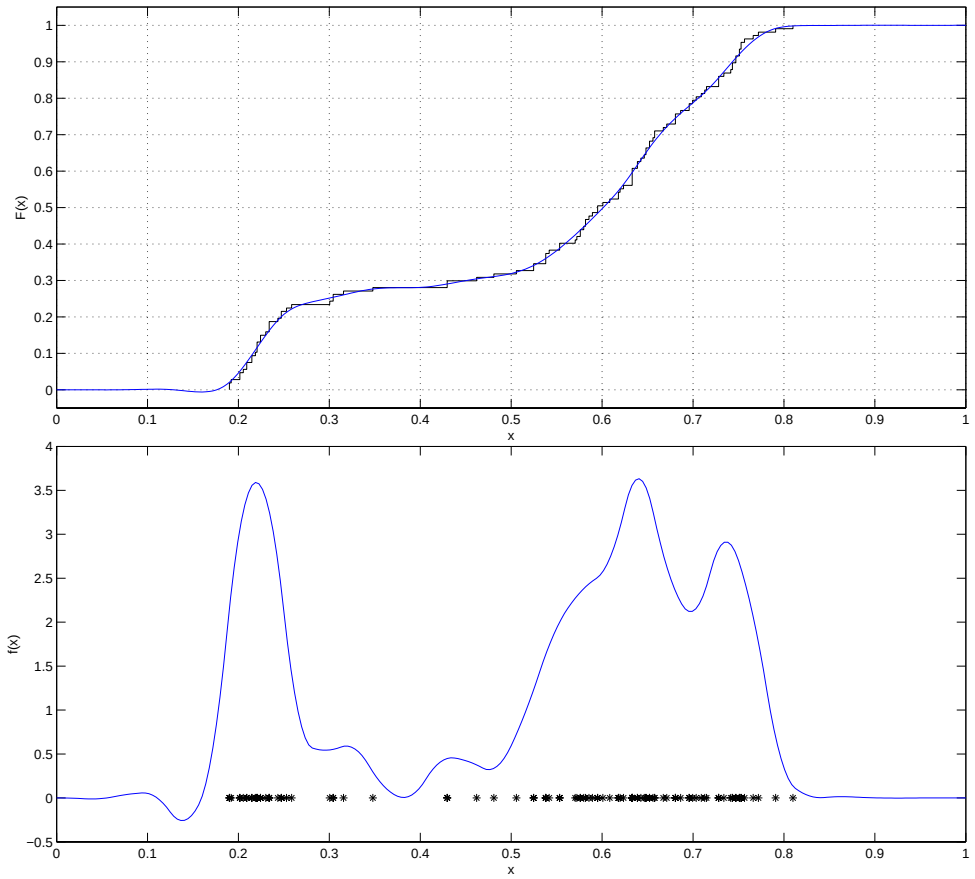

Abbildung 4.4: Dichteschätzung der Eruptionszeiten des Old Faithful Geysirs.

4.4.4 Diskussion

Einige Vor- und Nachteile des Verfahrens werden bereits in den soeben angeführten Beispielen deutlich. So lässt sich das Verfahren mit Hilfe der gegebenen Parameter gut steuern, wobei in dem Kontext über eine geeignete *a-priori* Parameterwahl gesprochen werden muss. Hat man die Originaldichte vorliegen, so lässt sich diese bei festem Diskretisierungslevel *a-posteriori*, abhängig von der Datenmenge, durch Wahl eines geeigneten Glättungsparameters angemessen gut rekonstruieren. Um eine *a-priori* Bestimmung der Parameter durchführen zu können, sind Verfahren wie die Bootstrap-Methode oder verallgemeinerte Kreuzvalidierung durchzuführen (siehe [Wah90], [Sco92]). Auf diese werden wir in dieser Arbeit aber nicht genauer eingehen.

Ein weiterer Vorteil des Verfahrens ist sicherlich die saubere theoretische Grundlage, auch wenn für die Anwendung eher die Ergebnisse im Vordergrund stehen. Als Nachteil des Ansatzes stellt man jedoch fest, dass die approximierte Dichtefunktion die Eigenschaft der Nichtnegativität einer Dichtefunktion nicht erfüllt. Dies liegt daran, dass die ermittelte Näherung an die Verteilungsfunktion nicht zwingend monoton steigend ist. Natürlich ist das für reale Anwendungen

weniger relevant.

Für die Komplexität des Verfahrens ist von Nachteil, dass durch das Differenzieren die Basisfunktionen, im Vergleich zu direkt die Dichtefunktion bestimmende Verfahren, eine Ordnung höher gewählt werden müssen. Dadurch wird die Matrix des aufzustellenden Gleichungssystems dichter besetzt, was sich auf Speicher und Rechenzeit zum Lösen des Systems auswirkt. Der gleiche Effekt tritt bei den Regularitätsanforderungen auf, die ebenfalls eine Ordnung höher anzusetzen sind.

Zum Schluss lässt sich noch als Vorteil nennen, dass das zu lösende Gleichungssystem linear ist und daher direkt gelöst werden kann. Im Gegensatz dazu wird beim Maximum-a-posteriori-Ansatz, der im folgenden Kapitel vorgestellt wird, ein System von nichtlinearen Gleichungen aufgestellt, das nur näherungsweise gelöst werden kann. Trotz des höheren Aufwands wird man jedoch feststellen, dass diese Nichtlinearität einen verbesserten Dichteschätzer liefert und sich daher der Mehraufwand lohnt.

5 Maximum a-posteriori mit Gauß-Prior

In diesem Kapitel präsentieren wir einen Maximum-a-posteriori-Ansatz, der auf Hegland und Griebel zurückgeht. Analog zu dem in 3.1.2 beschriebenen üblichen Maximum-a-posteriori-Verfahren, bestimmen wir zunächst aus den Datenpunkten die Likelihood-Funktion, hier allerdings als Funktional einer Parameterfunktion, die also einen unendlichen Parametervektor darstellt. Mit einer a-priori Annahme zur Verteilung der Parameterfunktionen bestimmen wir anschließend über den Satz von Bayes ein a-posteriori Funktional, dessen Minimum unseren Dichteschätzer charakterisiert. Da Dichtefunktionen im Unendlichen nicht im klassischen Sinne existieren, muss man sich zusätzlich Gedanken über einen verallgemeinerte Begriff der Dichte machen.

Der hier vorgestellte Maximum-a-posteriori-Ansatz kann nicht nur zur Dichteschätzung, sondern auch für Regression und Klassifikation verwendet werden. Daher wird in diesem Kapitel zunächst der allgemeine Ansatz vorgestellt, wovon in den nächsten Kapiteln der Spezialfall der Dichteschätzung betrachtet wird.

5.1 Herleitung des MAP-Funktionals

Wir betrachten ein allgemeines Lernproblem über einer Menge von Datenpunkten $x_i \in X$ mit zugehörigen Antwortvariablen $y_i \in Y$

$$D = \{(x_1, y_1), ..., (x_n, y_n)\}.$$

Gesucht ist eine Funktion $f : X \to Y$, die den Zusammenhang zwischen den Variablen möglichst gut darstellt.

Für unseren Ansatz gehen wir davon aus, dass diesen Daten eine Wahrscheinlichkeitsverteilung mit Dichtefunktion $p(x, y)$ zugrunde liegt. Ziel ist es die bedingte Wahrscheinlichkeit $p(y|x)$ mit Hilfe der Daten zu rekonstruieren. Je nachdem welche Gestalt X und Y haben, entstehen verschiedene Probleme:

- Falls X unendlich und Y endlich ist, resultiert ein *Klassifikationsproblem.*
- Sind X und Y unendlich, bekommt man ein *Regressionsproblem.*
- Ist X unendlich und besteht Y nur aus einem Element, erhält man eine *Dichteschätzung.*

Zunächst können die Daten durch ihre Likelihood-Funktion

$$L(D) = \prod_{i=1}^{n} p(y_i|x_i; \theta) \tag{5.1.1}$$

beschrieben werden. Wenn die Menge der variablen Parameter θ der Punktdichten $p(y_i|x_i;\theta)$ endlich ist, wendet man ein Standard- Maximum-Likelihood-Verfahren an, wie es in Kapitel 3.1.1 vorgestellt wurde. Dieses bestimmt $\hat{f}$ als Maximum der Likelihood-Funktion $L(D)$ nach den einzelnen Parametern. Das Gesetz der großen Zahlen garantiert einem dann für $n \to \infty$ die Konvergenz gegen die wirkliche Wahrscheinlichkeitsdichte. Es treten an dieser Stelle die in 3.1 angeführten Probleme parametrischer Dichteschätzungsverfahren auf.

Wir betrachten daher einen etwas anderen Ansatz, bei dem statt eines endlichen Parametervektors θ eine Parameterfunktion u eingeführt wird, die man auch als unendlichen Parametervektor interpretieren kann. Gesucht ist also eine Funktion $u \in \mathbb{R}^{X\times Y}$. Damit schränkt man die Klasse von möglichen Verteilungen nur sehr gering ein.

Analog zu anderen Regressionsaufgaben stellt man fest, dass dieses Problem ohne weitere Voraussetzungen an die Dichte f, beziehungsweise die Parameterfunktion u schlecht gestellt ist. Es ist demnach eine gewisse Regularisierung notwendig. Dies geschieht hier durch eine Einschränkung auf einen Funktionenraum $H \subset \mathbb{R}^{X\times Y}$, aus dem die gesuchte Funktion gewählt wird. Dazu führen wir eine Familie von Verteilungen ein, aus denen wir unsere Lösung bestimmen wollen.

5.1.1 Exponentialfamilien

In der statistischen Lerntheorie werden häufig so genannte *Exponentialfamilien* von Verteilungen betrachtet. Dabei wird die Wahrscheinlichkeitsdichte f beschrieben durch die Exponentialfunktion einer Linearkombination von *suffizienten Statistiken* der Verteilung (siehe[SS01]).

Definition 5.1.1 (Exponentialfamilie)**.** Eine Familie von Verteilungen, deren Wahrscheinlichkeitsdichten sich schreiben lassen als

$$f(x;\theta) = \exp\left(\langle\phi(x),\theta\rangle - g(\theta)\right)$$

für gegebene Parameter θ heißt *Exponentialfamilie*. Dabei ist

- θ der so genannte *kanonische Parameter*
- $\phi(x)$ die *suffiziente Statistik* von x
- $g(\theta)$ die *log-partition-Funktion*, die garantiert, dass die Verteilung zu 1 integriert. Also
$$g(\theta) = log \int_X \exp\left(\langle\phi(x),\theta\rangle\right) dx.$$

Alternativ wird die log-partition Funktion auch als die *Momente erzeugende Funktion* bezeichnet. Wie der Name schon sagt, erhält man nach differenzieren von $g(\theta)$ die Momente der Verteilung. Leitet man $g(\theta)$ einmal nach θ ab, so erhält man den Erwartungswert bezüglich $f(x;\theta)$.

$$\frac{\partial g(\theta)}{\partial\theta} = \frac{\int \phi(x)\exp(\langle\phi(x),\theta\rangle)dx}{\int \exp(\langle\phi(x),\theta\rangle dx} = E[\phi(x)]$$

Die zweite Ableitung liefert analog die Kovarianz, entsprechend ergeben höhere Ableitungen Momente höherer Ordnung.

Ein bekanntes Beispiel einer Exponentialfamilie ist die Normalverteilung mit Dichtefunktion

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

Diese kann man analog zu obiger Definition schreiben als

$$\begin{aligned} f(x) &= \exp\left(-\frac{1}{2\sigma^2} + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \\ &= \exp\Bigg(\langle\underbrace{(x, x^2)}_{\phi(x)}, \theta\rangle - (\underbrace{\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)}_{g(\theta)})\Bigg) \end{aligned}$$

Nach einigen Rechnungen resultiert $\theta_1 = \mu\sigma^{-2}$ und $\theta_2 = -\frac{1}{2}\sigma^{-2}$ als Parametervektor und die zugehörige log-partition-Funktion als

$$g(\theta) = -\frac{1}{4}\theta_1^2\theta_2^{-1} + \frac{1}{2}\log 2\pi - \frac{1}{2}\log(-2\theta_2).$$

Im Folgenden suchen wir eine Parameterfunktion u, die eine Dichtefunktion der Form

$$p(x|y) = \exp(u(x,y) - g(u,x))$$

charakterisiert. Damit hat man sehr viel Spielraum für die Wahl von u und schränkt sich nicht wie bei einem Standard MAP-Ansatz auf eine feste Verteilung ein.

Mit Hilfe der Exponentialfamilien haben wir die Aufgabenstellung umformuliert in eine Suche nach einer geeigneten Parameterfunktion u, an die bisher wenige Einschränkungen gemacht werden. Später wird u dann als Linearkombination von Basisfunktionen aus einem Funktionenraum V gewählt.

Die Wahrscheinlichkeit für das Eintreten einer Datenmenge D bei gegebenem u kann man mit Hilfe der Likelihood-Funktion schreiben als

$$L(u, D) = \prod_{i=1}^{n} p(y_i|x_i) = \prod_{i=1}^{n} \exp(u(x_i, y_i) - g(u, x_i)) = P(D|u). \tag{5.1.2}$$

Diese wird oft der einfacheren Handhabung wegen durch die logarithmierte Likelihood Funktion

$$l(u, D) = \sum_{i=1}^{n} (u(x_i, y_i) - g(u, x_i)) \tag{5.1.3}$$

ersetzt.

Betrachtet wird nun ein *Maximum-a-posteriori-Ansatz*, bei dem die Wahrscheinlichkeit für ein u bei gegeben Daten D maximiert werden soll. Voraussetzung ist eine *a-priori* Annahme an die Wahrscheinlichkeit $P(u)$ mit der eine Vorabinformation an die Verteilung der Daten eingeht. Hier wird später angenommen, dass die Verteilung der möglichen Funktionen u normalverteilt ist. Man spricht dann auch von einem *Gauß-Prior*.

Nach dem Satz von Bayes ist die bedingte Wahrscheinlichkeit

$$P(u|D) = \frac{P(u)P(D|u)}{P(D)},$$

wobei

$$P(D) = \int P(u)P(D|u)\mu_0(du)$$

mit geeignetem Maß μ_0 ist. Als Maß interpretiert lautet die Gleichung:

$$\mu_1(A) = \frac{\int_A L(u,D)\mu_0(du)}{\int_{\mathbb{R}^{X\times Y}} L(u,D)\mu_0(du)}.$$

Wenn in späteren Betrachtungen vom *a-priori-Maß* die Rede ist, ist das Maß μ_0 gemeint. Das *a-posteriori-Maß*, bezüglich dem ein Maximum gesucht wird, ist hier mit μ_1 bezeichnet.

Wir bestimmen nun analog zu dem Standard MAP-Verfahren aus Kapitel 3.1.2 ein Maximum dieser Wahrscheinlichkeit $P(u|D)$, beziehungsweise eine Maximalstelle des Maßes μ_1. Dazu sucht man so genannte kritische Punkte beziehungsweise Maximalstellen der Verteilung von u. Die Wahrscheinlichkeit, dass das gesuchte u sehr nahe bei solchen Punkten liegt, ist dann sehr groß. Die Frage ist, wie man solche Punkte findet. Offensichtlich hängt ein Maximum stark von der Wahl des Maßes μ_0 ab. Es kann vorkommen, dass $P(u|D)$ Maximalstellen an Punkten hat, die bzgl. μ_0 sehr geringes Maß haben und daher nicht gewählt werden. Um dies zu verhindern, könnte man translationsinvariante Maße wie das Lebesguemaß betrachten. Diese sind aber nur für endlich dimensionale Räume definiert. Demnach besitzen weder die a-priori Wahrscheinlichkeit $P(u)$, als auch die a-posterior Wahrscheinlichkeit $P(u|D)$ eine Dichte im gewöhnlichen Sinne.

Um nun Punkte mit maximaler Dichte zu finden, braucht man einen allgemeineren Begriff der Dichte. Daher wurden in [GH06] und [Heg06] verallgemeinerte Dichten für spezielle unendlichdimensionale Funktionenräume definiert, die es dann ermöglichen eine numerische Lösung zu bestimmen (siehe auch [Bog98]). Dabei stimmen die im nächsten Unterkapitel vorgestellten *verallgemeinerten Dichten* im endlich dimensionalen Fall bis auf eine Konstante mit den gewöhnlichen Dichten überein.

In bisherigen Verfahren des maschinellen Lernens wurden im Allgemeinen nur endlich dimensionale Verteilungen betrachtet, die durch eine gemeinsame Verteilung von u an den jeweiligen Datenpunkten x_i $i = 1, ..., n$ charakterisiert sind (siehe [YA04],[See04]). Damit umgeht man zwar das Problem der unendlichdimensionalen Dichten, bekommt aber immer noch keine variationelle Form für das eigentliche maximum a-posteriori Problem.

Der Ansatz wie wir ihn im Folgenden betrachten werden, macht an die a-priori Verteilung der Parameterfunktionen u die Annahmen, dass sie normalverteilt sind. Man wählt also einen Gauß-Prior, welcher als Gaußmaß über dem $R^{X\times Y}$ definiert ist. Alternativ kann man diesen Prior auch als unendlichen Zufallsvektor von normalverteilten Zufallsvariablen $u(x)$ interpretieren.

Über die a-posteriori Verteilung kann man in der Regel keine genaueren Aussagen machen. Diese sind im Allgemeinen auch keine Gaußmaße, können aber, wie man später sehen wird, analog charakterisiert werden.

Im Gegensatz zur a-priori und a-posteriori Verteilung besitzt die Likelihood-Funktion $L(u, D)$ eine Dichte im klassischen Sinne. Sie besteht aus dem endlichen Produkt der Dichtefunktionen an den Datenpunkten (siehe (5.1.2)).

Eine komplette Darstellung der hier aufgeführten Theorien findet sich in dem Artikel [Heg06], auf dem die folgenden Abschnitte basieren.

5.1.2 Gaußscher Prior

Auch wenn im Gegensatz zu endlich dimensionalen Maßen ein Gaußsches Wahrscheinlichkeitsmaß γ über dem $\mathbb{R}^{X\times Y}$ im allgemeinen keine Dichte besitzt, so wird es doch durch Erwartungswert und Varianz charakterisiert. Bevor wir diese angeben können, müssen wir einen Definitionsbereich für das Maß und eine entsprechende σ-Algebra definieren.

Eine natürliche Topologie auf $\mathbb{R}^{X\times Y}$ ist die durch die Topologie auf $\mathbb{R}$ induzierte Produkttopologie. Diese ist die maximale Topologie, für die alle Punktauswertungsfunktionale der Form $u \to u(x, y)$ stetig sind. Der Dualraum des Funktionenraumes $\mathbb{R}^{X\times Y}$ ist dann die Menge aller stetigen linearen Funktionale

$$\phi(u) = \sum_{j=1}^{n} c_j u(x_j, y_j), \tag{5.1.4}$$

für $c_j \in \mathbb{R}$ und $x_j \in X$, $y_j \in Y$. Man benötigt diese Funktionale, um eine σ-Algebra für das Wahrscheinlichkeitsmaß auf $\mathbb{R}^{X\times Y}$ angeben zu können. Diese ergibt sich für stetige lineare Funktionale ϕ dann in der Form

$$\{u \in \mathbb{R}^{X\times Y} : \phi(u) < c\},$$

für eine Konstante $c < \infty$. Als Prior nimmt man nun ein *Gauß-Radon-Maß* γ auf dieser σ-Algebra. Dabei heißt ein Wahrscheinlichkeitsmaß μ *Radon-Maß*, falls $\mu(A)$ das Supremum von $\mu(K)$ über alle kompakten Teilmengen $K \subset A$ ist. Dass heißt bei einem Gauß-Radon-Maß lassen sich alle messbaren Mengen durch kompakte Teilmengen approximieren und die Auswertungen aller Funktionen $\phi(u)$ sind normalverteilte Zufallsvariablen.

Hat man nun ein u gegeben, dass gemäß einem solchen Maß γ verteilt ist, so sind die punktweisen Auswertungen $u(x, y)$ eine Familie von normalverteilten Zufallsvariablen und u somit ein sogenanntes *Gaußsches Zufallsfeld*. Die Werte eines stetigen Funktionals ϕ sind dann normalverteilt mit Erwartungswert

$$\int_{\mathbb{R}^{X\times Y}} \sum_{i=1}^{n} c_i u(x_i, y_i)\gamma(du) = \sum_{i=1}^{n} c_i \bar{u}(x_i, y_i)$$

und Varianz

$$\int_{\mathbb{R}^{X\times Y}} \sum_{i,j=1}^{n} c_i c_j [u(x_i,y_i) - \bar{u}(x_i,y_i)][u(x_j,y_j) - \bar{u}(x_j,y_j)]\gamma(du)$$
$$= \sum_{i,j=1}^{n} c_i c_j K(x_i,y_i,x_j,y_j),$$

wobei $\bar{u}(x,y)$ der Mittelwert des Zufallsvektors und $K(x,x',y,y')$ die Kovarianzmatrix ist, die auch als *Kern* bezeichnet wird. Diese beiden Funktionale definieren also den Prior γ und charakterisieren das „wahrscheinlichste“ u, falls keine Daten gegeben sind, sowie die Kovarianz zwischen Funktionswerten von u an verschiedenen Stellen (x,y). Mit Hilfe dieser Funktionale kann man Gebiete einführen, an denen eine hohe Korrelation der Funktionswerte vorliegt und die damit interresant für mögliche Maxima sind.

Der Prior γ definiert eine Seminorm

$$\|u\|_{H(\gamma)} = \sup_{\phi}\{\phi(u) : R(\phi,\phi) \leq 1\}$$

wobei das Supremum über alle stetigen linearen Funktionale der Form (5.1.4) gebildet wird und

$$R(\phi,\phi) = \sum_{i,j=1}^{n} c_i c_j K(x_i,y_i,x_j,y_j)$$

ist. Das Gebiet dieser Seminorm ist dann ein so genannter *Cameron-Martin-Raum*

$$H(\gamma) = \{u : \|u\|_{H(\gamma)} < \infty\}, \tag{5.1.5}$$

der ein Hilbertraum mit reproduzierendem Kern K ist. Mehr Details dazu finden sich beispielsweise in [Bog98]. Es kann gezeigt werden, dass dieser Raum gerade aus allen Funktionen $u \in \mathbb{R}^{X\times Y}$ besteht, die sich als Grenzwert von Folgen $u_1, u_2, ...$ mit

$$u_n(x,y) = \sum_{i=1}^{n} c_{i,n} K(x_{i,n}, y_{i,n}, x, y)$$

schreiben lassen. Dabei sind die Folgen $c_{i,n}$ und $(x_{i,n}, y_{i,n})$ so zu wählen, dass die entsprechenden Funktionale ϕ_n mit

$$\phi_n(u) = \sum_{i=1}^{n} c_{i,n}[u(x_{i,n},y_{i,n}) - \bar{u}(x_{i,n},y_{i,n})]$$

in $L_2(\gamma)$ konvergieren.

Anzumerken ist dabei, dass der Grenzwert $\tilde{\phi}$ im allgemeinen kein stetiges lineares Funktional auf $X \times Y$ ist. Allerdings kann er als stetiges lineares Funktional auf dem Hilbertraum $H(\gamma)$ angenommen werden, wodurch man

$$\tilde{\phi}(u) = \|u\|^2_{H(\gamma)}$$

erhält. Die Menge der auf diese Weise definierten $\tilde{\phi}$ beinhalten alle stetigen linearen Funktionale ϕ auf $\mathbb{R}^{X\times Y}$.

Damit haben wir nun sowohl das Maß γ, als auch den Raum aus dem wir unsere Lösung bestimmen wollen. Jetzt müssen wir noch eine notwendige Bedingung an Maximalpunkte bezüglich γ stellen. Dazu werden im nächsten Abschnitt *stationäre Punkte* und die so genannte *logarithmische Ableitung* definiert.

5.1.3 Logarithmische Ableitung und stationäre Punkte

Wir betrachten nun allgemeine, eventuell nichtlineare Funktionale der Form

$$\psi(u) := \theta(\phi_1(u), ..., \phi_m(u))$$

wobei θ aus dem Raum der unendlich oft differenzierbaren Funktionen mit kompaktem Träger über dem $\mathbb{R}^m$ ($C_b^\infty(\mathbb{R}^m)$) ist, und die Funktionale ϕ_i aus dem Dualraum X^* von X sind. Dies sind so genannte *glatte zylindrische Funktionale*, deren Fréchet Ableitung von der Form

$$\partial_v \psi(u) = \sum_{i=1}^{m} \frac{\partial \theta}{\partial \phi_i}(u)\phi_i(v)$$

ist. An die Stelle von ψ wird später die Likelihood-Funktion aus Gleichung (5.1.2) treten.

Definition 5.1.2 (Logarithmische Ableitung)**.** Man sagt ein Radonmaß μ ist differenzierbar in Richtung $v \in \mathbb{R}^X$, falls ein Funktional $\beta_v^\mu \in L^1(\mu)$ existiert, so dass für alle glatten zylindrischen Funktionale ψ

$$\int_X \partial_v \psi(u)\mu(du) = -\int_X \psi(u)\beta_v^\mu(u)\mu(du)$$

gilt. Das Funktional β_v^μ wird *logarithmische Ableitung* von μ nach v genannt.

Man kann zeigen, dass für ein Gauß-Radon-Maß γ mit Erwartungswert 0

$$\beta_v^\gamma(u) = -(v, u)_{H(\gamma)}$$

und γ differenzierbar für alle $u \in H(\gamma)$ nach $v \in H(\gamma)$ ist. Mit $(.,.)_{H(\gamma)}$ sei dabei das zur Norm $\|.\|_{H(\gamma)}$ gehörige Skalarprodukt bezeichnet.

Für ein beliebiges $L \in L^1(\gamma)$ sei $\mu = L \cdot \gamma$ ein Maß, so dass $L = d\mu/d\gamma$ die Radon-Nikodym-Ableitung ist. Also ergibt sich

$$\mu(A) = \int_A L(u)\gamma(du).$$

Falls L differenzierbar ist, folgt dann auch, dass $\mu = L \cdot \gamma$ differenzierbar ist und als logarithmische Ableitung resultiert

$$\beta_v^\mu(u) = -(v, u)_{H(\gamma)} + \frac{\partial_v L(u)}{L(u)}. \tag{5.1.6}$$

Definition 5.1.3 (Stationärer Punkt)**.** Ein Punkt $u \in \mathbb{R}^X$ heißt *stationärer Punkt* des Maßes μ, falls die logarithmische Ableitung

$$\beta_v^\mu(u) = 0$$

für alle Richtungen v ist, entlang derer μ differenzierbar ist.

Man sieht, dass dies offensichtlich auch die stationären Punkte für ein endlich dimensionales Maß μ charakterisiert. Mit Hilfe der logarithmischen Ableitung hat man nun also das Prinzip der stationären Punkte von Wahrscheinlichkeitsdichten auf allgemeine Maße μ der Form $\mu = L \cdot \gamma$ übertragen. Dabei hängt der Begriff der logarithmischen Ableitung nicht von der Existenz einer Dichte ab. Aus Definition 5.1.3 und Gleichung (5.1.6) ergibt sich für ein Maß $\mu = L \cdot \gamma$ die folgende Charakterisierung eines stationären Punktes u:

$$-(v,u)_{H(\gamma)} + \frac{\partial_v L(u)}{L(u)} = 0, \quad \text{für} \quad v \in H(\gamma).$$

Offensichtlich ist der Erwartungswert eines Gaußmaßes ein stationären Punkt. Um Maximalstellen von anderen stationären Punkten unterscheiden zu können, benötigt man zusätzliche Bedingungen, die im nächsten Abschnitt genauer untersucht werden. Dies führt zum Begriff der *verallgemeinerten Dichte*, mit deren Hilfe man Maximalstellen von unendlich dimensionalen Maßen unter gewissen Vorraussetzungen bestimmen kann.

5.1.4 Verallgemeinerte Dichte

Man betrachtet nun ein allgemeines Radonmaß μ, welches auf $\mathbb{R}^{X \times Y}$ definiert ist. Dann ist das translatierte Maß μ_h definiert als

$$\mu_h(A) = \mu(A + h).$$

Dabei ist $A + h$ die Menge aller Funktionen $u + h$ mit $u \in A$. Vorausgesetzt A ist messbar, ist auch $A + h$ messbar. Falls μ_h absolut stetig in Bezug zu μ ist, existiert nach Satz 2.3.2 eine messbare Funktion $r_h(u)$, so dass

$$\mu_h(A) = \int_A r_h(u)\mu(du).$$

Also ist die Radon-Nikodym-Ableitung

$$r_h = \frac{d\mu_h}{d\mu}.$$

Man bezeichnet ein Maß μ als *quasi-invariant* unter Translation mit h, falls die Maße μ_h und μ gegenseitig absolutstetig, also äquivalent sind ($\mu_h \equiv \mu$). Sei H die Menge aller Funktionen $u \in \mathbb{R}^{X \times Y}$, für die μ quasi-invariant unter Translation mit u ist, also:

$$H := \{u : \mu \equiv \mu_u\}.$$

Offensichtlich ist $0 \in H$, und für jedes $h \in H$ auch $-h \in H$. Weiterhin gilt falls $\mu_{h_1} \equiv \mu$, $\mu_{h_1+h_2} \equiv \mu_{h_2}$ mit Radon-Nikodym-Ableitung

$$\frac{d\mu_{h_1+h_2}}{d\mu_{h_2}}(u) = r_{h_1}(u + h_2).$$

Mit der Kettenregel für Radon-Nikodym-Ableitungen erhält man dann

$$\frac{d\mu_{h_1+h_2}}{d\mu}(u) = \frac{d\mu_{h_1+h_2}}{d\mu_{h_2}}(u)\frac{d\mu_{h_2}}{d\mu}(u) = r_{h_1}(u + h_2)r_{h_2}(u).$$

Ersetzt man nun u durch 0, h_1 durch u und h_2 durch h, so erhält man:

$$r_{u+h}(0) = \frac{d\mu_{u+h}}{d\mu}(0) = \frac{d\mu_{u+h}}{d\mu_h}(0)\frac{d\mu_h}{d\mu}(0) = r_u(h)r_h(0).$$

Daraus resultiert:

$$r_h(u) = \frac{r_{u+h}(0)}{r_u(0)}.$$

Damit kann man die Funktionen $r_h(u)$ durch ihren Wert an der Stelle 0 charakterisieren. Wir bezeichnen $r_u(0)$ als die *verallgemeinerte Dichte* in u.

Falls $H = \mathbb{R}^{X\times Y}$ ist, muss die Menge $X \times Y$ endlich sein (siehe [Bog98]). Im Folgenden nehmen wir an, wir suchen ein $u \in H$, an dem μ ein Maximum hat. Wir sagen dazu $u + h$ ist *wahrscheinlicher* als u, falls

$$r_{u+h}(0) \geq r_u(0).$$

Eine Maximalstelle $r_u(0)$ nennen wir *Modus* von μ in Anlehnung an den häufigsten Wert einer Häufigkeitsverteilung (auch *Modalwert* genannt).

Für endliche Mengen $X \times Y$, die eine Wahrscheinlichkeitsdichte $p(u) > 0$ besitzen, kann gezeigt werden, dass dann $r_u(0) = p(u)/p(0)$. Also ist ein Modus im obigen Sinne im endlichen Fall gerade bis auf eine Normierung ein Maximalpunkt der Dichte und damit ein Modalwert im gewöhnlichen Sinne. Die verallgemeinerte Dichte eines Gaußmaßes mit Erwartungswert 0 kann als

$$r_u(0) = \exp(-\|u\|_H^2)$$

mit der Norm des Cameron-Martin-Raumes H geschrieben werden.

Nun können wir die Diskussion auf Wahrscheinlichkeitsmaße μ ausdehnen, die äquivalent zu einem Gaußmaß γ sind. Sei H der Cameron-Martin-Raum des Gaußmaßes γ aus (5.1.5). Die verallgemeinerte Dichte ergibt sich nun als

$$r_u(0) = \frac{d\mu}{d\gamma}\exp(-\|u\|_H^2),$$

wobei $d\mu/d\gamma$ wieder die Radon-Nikodym-Ableitung ist.

Wir haben jetzt also ein Kriterium für mögliche Extremstellen eines Gaußschen Wahrscheinlichkeitsmaßes hergeleitet. Diese beruht auf verallgemeinerten Dichten, die auf einer Teilmenge $H \subset \mathbb{R}^{X\times Y}$ definiert sind. Falls wir Maße haben, die äquivalent zu einem Gaußmaß sind, lässt sich die verallgemeinerte Dichte explizit in Abhängigkeit von der Radon-Nikodym-Ableitung und der Norm im entsprechenden Cameron-Martin-Raum darstellen. Damit haben wir die Möglichkeit eine variationelle Form für den „wahrscheinlichsten“ Punkt einer Verteilung angeben zu können.

Nun muss diese Methodik entsprechend auf den maximum a-posteriori Ansatz mit Exponentialfamilien übertragen werden, bei welchem das a-posteriori Maß äquivalent zum Gauß-Prior ist.

5.1.5 Anwendung auf Exponentialfamilien

Um die Theorie aus den vorherigen Abschnitten auf die maximum a-posteriori Methode anwenden zu können, braucht man eine genügend glatte Likelihood-Funktion. Diese Anforderung an die Regularität überträgt sich dann auf die log-partition-Funktion $g(u)$ aus Gleichung (5.1.1).

$$g(u,x) = \log \int_Y \exp(u(x,y))dy.$$

Ein wesentlicher Bestandteil ist der Integraloperator

$$u \in H \to \int_Y u(x,y)dy.$$

Wir bezeichnen im Folgenden mit $(\mathbb{R}^{X\times Y})^*_\gamma$ den Dualraum von $\mathbb{R}^{X\times Y}$ bezüglich $L^2(\gamma)$. Bedingung ist nun, dass dieser Integraloperator wohl definiert ist, und ein Element von $(\mathbb{R}^{X\times Y})^*_\gamma$ ist. Daraus folgt, dass der Operator ein stetiges lineares Funktional auf $H(\gamma)$ ist. Nach dem *Darstellungssatz von Riesz* (siehe z.B. [Alt02]) existiert dann ein $k_x \in H(\gamma)$, so dass

$$\int_Y u(x,y)dy = (k_x, u)_{H(\gamma)}$$

mit

$$k_x = \int_Y k_{x,y}dy,$$

wobei $k_{x,y}$ der reproduzierende Kern ist. Damit schreiben wir die log-partition-Funktion als

$$g(u,x) = \log((k_x, \exp(u))_{H(\gamma)}).$$

Als Komposition von messbaren Funktionen ist g somit messbar und für beliebige $u \in \mathbb{R}^{X\times Y}$ γ-fast-überall definiert. Sei $H(\gamma)$ stetig in $C(X\times Y)$, dem Raum der stetigen Funktionen auf $X\times Y$, eingebettet. Für ein $h \in H$ und $t \in \mathbb{R}$ gilt dann:

$$\left|\frac{dg(u+th,x)}{dt}\right| = \left|\frac{\int_Y exp(u(x,y)+th(x,y))h(x,y)dy}{\int_X exp(u(x,y)+th(x,y))dy}\right| \leq \|h\|_\infty \leq C\ \|h\|_{H(\gamma)}$$

Also ist g eine absolut stetige Funktion in t. Betrachten wir nun die Likelihood-Funktion

$$L(u,D) = \prod_{i=1}^{n} \exp(u(x_i,y_i) - g(u,x_i))$$

mit den gegebenen Datenpunkten $D = (x_1,y_1),...,(x_n,y_n)$. Mit den Eigenschaften von g ist L bezüglich des Produktmaßes $\gamma(du)\times dy_1 \times ... \times dy_n$ messbar und

$$\int_{\mathbb{R}^n} L(u,D)dy_1...dy_n = 1.$$

Nach dem *Satz von Tonelli* ist $L(\cdot,D)$ in $L^1(\gamma)$ für fast alle D und man kann die Reihenfolge der Integration vertauschen

$$\int_y \int_{\mathbb{R}^n} L(u,D)\gamma(du)dy_1...dy_n = \int_{\mathbb{R}^n}\int_Y L(u,D)\gamma(du)dy_1...dy_n = 1.$$

Damit können wir folgende Proposition angeben

Proposition 5.1.4. *Sei γ eine Gauß-Radon-Maß auf $\mathbb{R}^{X\times Y}$, so dass $H(\gamma) \subset C(X\times Y)$ eine stetige Einbettung ist und $h \in H(\gamma)$. Weiterhin sei das Funktional $u \to \int_y u(x,y)dy$ aus $L^2(\gamma)$ für alle $x \in X$. Dann erfüllt die logarithmische Ableitung des Maßes $\mu = L(\cdot, D)\cdot\gamma$*

$$\beta_h^\mu = \beta_h^\gamma + \frac{\partial_h L(\cdot, D)}{L(\cdot, D)}.$$

Das a-posteriori Maß ist das bedingte Maß μ, definiert durch

$$\mu(A|D) = \frac{\int_A L(u,D)\gamma(du)}{\int_{\mathbb{R}^{X\times Y}} L(u,D)\gamma(du)}$$

für alle messbaren Mengen $A \in \mathbb{R}^{X\times Y}$. Dies ist also nur eine skalierte Version des im vorletzten Abschnittes betrachteten Maßes μ, voraus folgt, dass die beiden Verteilungen die gleichen Maxima und stationären Punkte haben.

Setzen wir nun die negative logarithmierte Likelihood-Funktion

$$l(u,D) = -\sum_{i=1}^{n}(u(x_i,y_i) - g(u,x_i))$$

ein, so erhalten wir die Bedingung an einen stationären Punkt

$$(h,u)_{H(\gamma)} + \partial_h l(u,D) = 0, \quad \text{für alle} \quad h \in H(\gamma).$$

Daraus folgt, dass ein Maximalpunkt von $\mu(A|D)$ das Funktional

$$J(u) = \|u\|_H^2 + l(u) \tag{5.1.7}$$

minimiert.

5.1.6 Konvexität und Eindeutigkeit

Nachdem wir eine variationelle Form für das maximum a-posteriori Verfahren angegeben haben, wollen wir uns jetzt die Lösbarkeit dessen anzusehen. Dazu setzen wir die logarithmierte Likelihood-Funktion $l(u,D)$ und die log-partition-Funktion $g(u,x)$ in (5.1.7) ein und erhalten

$$J(u) = \|u\|_H^2 - \sum_{i=1}^{n}\left(u(x_i,y_i) - \log\int_Y \exp(u(x,y))dy)\right).$$

Der Beweis der folgenden Proposition zeigt für $J(u)$ die Konvexität und eindeutige Lösbarkeit. Die Aussage und deren Beweis stammen aus [GH06].

Proposition 5.1.5. *Sei das nichtlineare Funktional $\phi : H \to \mathbb{R}$ definiert auf einem reproduzierendem Kern-Hilbert-Raum mit Norm $\|.\|_H$ durch*

$$\phi(u) = \log\int_Y \exp(u(x,y))dy.$$

Dann ist das Funktional

$$J(u) = \|u\|_H^2 + \sum_{i=1}^{n}\phi(u)(x_i) - \sum_{i=1}^{n}u(x_i,y_i)$$

strikt konvex und besitzt ein eindeutiges Minimum.

Beweis. Für beliebige Funktionen u_1, u_2 und $\beta \in (0,1)$ gilt offensichtlich

$$\phi(\beta u_1 + (1-\beta)u_2) = \log \int_Y \exp(u_1(\cdot,y))^\beta (\exp(u_2(\cdot,y)))^{1-\beta} dy.$$

Mit der Ungleichung von HÖLDER ergibt sich

$$\begin{aligned} &\int_Y \exp(u_1(\cdot,y))^\beta (\exp(u_2(\cdot,y)))^{1-\beta} dy \\ &\leq \left(\int_Y \exp(u_1(\cdot,y)) dy \right)^\beta \left(\int_Y \exp(u_2(\cdot,y)) dy \right)^{1-\beta}. \end{aligned}$$

Da der Logarithmus monoton steigend ist, folgt die Konvexität von ϕ

$$\phi(\beta u_1 + (1-\beta)u_2) \leq \beta\phi(u_1) + (1-\beta)\phi(u_2).$$

Nun ist $J(u)$ strikt konvex, wenn $\|.\|_H^2$ positiv definit ist (Der Term $\sum_{i=1}^n u(x_i)$ ist nur eine konstante Verschiebung).

Sei c eine Konstante, sodass gilt

$$|u(x,y)| \leq c\|u\|_H.$$

Weiterhin sei $\int_Y 1 dy = C$. Daraus resultieren obere und untere Schranken für das Funktional ϕ

$$\phi(u)(x) \geq \log \int \exp(-\|u\|_\infty) dy \geq -C\|u\|_\infty \geq C\|u\|_H$$

$$\phi(u)(x) \leq \log \int \exp(\|u\|_\infty) dx \leq C\|u\|_\infty \leq C\|u\|_H.$$

Es gilt somit $\phi(u)(x) \in [-C\|u\|_H, C\|u\|_H]$. Damit kann man Schranken für J angeben

$$-(Cn)^2 \leq \|u\|_H^2 - 2Cn\|u\|_H \leq J(u) \leq \|u\|_H^2 + 2Cn\|u\|_H.$$

und hat somit gezeigt, dass J ein eindeutiges endliches Minimum besitzt. □

5.2 Minimierung des Zielfunktionals

Das zu minimierende Zielfunktional hat die Gestalt

$$J(u) = \|u\|_H^2 - \sum_{i=1}^n \left(u(x_i,y_i) - \log \int_Y \exp(u(x_i,y)) dy \right) \to \min.$$

Ähnlich wie bei Regularisierungsnetzwerken existieren zwei grundsätzliche Ansätze um dieses zu lösen: Zum einen die Darstellung als Linearkombination von auf den Datenpunkten zentrierten Kernfunktionen, und zum anderen die Formulierung der Galerkin-Gleichungen für eine Minimierung in einem Funktionenraum.

Zunächst gehen wir auf den *datenbasierten* Ansatz mittels Kerndarstellung ein. Anschließend werden die *gitterbasierten* Galerkin-Gleichungen für das Problem hergeleitet.

5.2.1 Kerndarstellung

Sei $H(\gamma)$ der Hilbertraum aus dem die Lösung u gesucht wird. Weiterhin sei $K : H(\gamma)^* \to H(\gamma)$ eine Bijektion des Dualraums von $H(\gamma)$ auf sich selbst, die nach dem Darstellungssatz von Riesz existiert, mit

$$\phi(v) = (K\phi, v)_{H(\gamma)}, \quad \text{für alle} \quad \phi \in H(\gamma)^* \quad \text{und} \quad v \in H(\gamma).$$

Der Operator K ist offensichtlich der reproduzierende Kern des Hilbertraumes und man kann daher jede Funktion $v \in H(\gamma)$ mit Dirac'scher Deltafunktion δ schreiben als:

$$v(x, y) = \delta_{x,y}(v) = (K\delta_{x,y}, v)_{H(\gamma)}, \quad \text{für} \quad (x, y) \in X \times Y.$$

Man bezeichnet K als den *reproduzierenden Kernoperator*. Als Folgerung daraus bekommt man die Aussage:

Proposition 5.2.1. *Sei u^* ein Minimum des nichtlinearen Funktionals $J(u) = \|u\|^2_{H(\gamma)} + l(u, D)$. Dann gilt*

$$u^* + \frac{1}{2} K \nabla l(u^*, D) = 0, \tag{5.2.1}$$

wobei K der reproduzierende Kernoperator von $H(\gamma)$ ist.

Beweis. Das Minimum von $J(u)$ muß ein stationärer Punkt sein, also

$$\langle \nabla J(u^*), v \rangle = 0 \quad \text{für alle} \quad v \in H(\gamma).$$

Nach Einsetzen der Ableitung von $\|u\|^2_{H(\gamma)}$ und der Definition von K ergibt sich

$$\langle \nabla J(u^*), v \rangle = 2(u^*, v)_{H(\gamma)} + (K \nabla l(u^*, D), v)_{H(\gamma)}$$

und die Aussage ist bewiesen, da dies für alle $v \in H$ Null ist. □

Hier ist wieder das bekannte Prinzip angewendet worden, den Regularisierungsterm durch einen Kern auszudrücken. Das heißt die Glattheitsvorraussetzungen werden jetzt durch Wahl geeigneter Kernfunktionen und nicht mehr explizit durch einen Regularisierungsterm vorgegeben.

5.2.2 Funktionenraumdarstellung

Unser Ziel ist die Minimierung des konvexen Funktionals $J(u)$.

$$J(u) \xrightarrow{u \in H} min$$

$$J(u) = \|u\|^2_H - \sum_{i=1}^{n} \left(u(x_i, y_i) - \log \int_Y \exp(u(x_i, y)) dy \right). \tag{5.2.2}$$

Wobei $\|u\|^2_H = n\lambda/2\langle u, Tu \rangle, \quad u \in H$ die Norm des reproduzierenden Kern-Hilbertraumes H mit einem Differentialoperator T ist. Mit $\langle \cdot, \cdot \rangle$ sei dabei das

L_2-Skalarprodukt bezeichnet. Mit dem Glättungsparameter λ lässt sich die Regularität der Lösung steuern.

Stellen wir nun u bezüglich einer (unendlichen) Funktionenraumbasis $\{\varphi_j\}_{j=1}^{\infty}$ des Hilbertraumes H dar

$$u(x,y) = \sum_{j=1}^{\infty} \alpha_j \varphi_j(x,y).$$

Einsetzen in $J(u)$ ergibt

$$J(u) = (\sum_{j=1}^{\infty} \alpha_j \varphi_j, \sum_{l=1}^{\infty} \alpha_l \varphi_l)_H - \sum_{i=1}^{n} \left(\sum_{j=1}^{\infty} \alpha_j \varphi_j(x_i,y_i) - \log \int \exp\left(\sum_{j=1}^{\infty} \alpha_j \varphi_j \right) dy \right).$$

Nach Differentiation in den Koeffizienten α_k, $k = 1, ..., \infty$ erhält man

$$\begin{aligned} \frac{\partial J(u)}{\partial \alpha_k} =& 2 \sum_{j=1}^{\infty} \alpha_j (\varphi_k, \varphi_j)_H \\ & - \sum_{i=1}^{n} \left(\sum_{j=1}^{\infty} \alpha_j \varphi_j(x_i,y_i) - \int \varphi_k(x,y) \frac{\exp(\sum_{j=1}^{\infty} \alpha_j \varphi_j(x,y))}{\int \exp(\sum_{j=1}^{\infty} \alpha_j \varphi_j(x,y)) dy} dy \right), \end{aligned}$$

bzw. die notwendige Bedingung ($k = 1, ..., \infty$)

$$2(\varphi_k, u)_H - \sum_{i=1}^{n} \left(\varphi_k(x_i,y_i) - \int \varphi_k(x_i,y) \frac{\exp(u(x_i,y))}{\int \exp(u(x_i,z)) dz} dy \right) \stackrel{!}{=} 0.$$

Wählt man nun geeignete Testfunktionen v, so resultieren daraus die Galerkin-Gleichungen in der schwachen Form

$$a(v,u) + \frac{1}{n} \sum_{i=1}^{n} \int v(x_i,y) \frac{\exp(u(x_i,y))}{\int \exp(u(x_i,z)) dz} dy = \frac{1}{n} \sum_{i=1}^{n} v(x_i,y_i) \quad \forall v \in H. \tag{5.2.3}$$

Dabei ist die Bilinearform a definiert als

$$a(v,u) = \frac{2}{n}(v,u)_H = \lambda \langle T'v, T'u \rangle, \tag{5.2.4}$$

mit entsprechendem Differentialoperator T'. Der erste lineare Term entspricht einer Steifigkeitsmatrix, wie er auch bei der Behandlung von partiellen Differentialgleichungen auftritt. Man bezeichnet diese auch als *Regularisierungsmatrix*, da mit ihr die Regularität der Lösung gesteuert wird. Auf der rechten Seite haben wir in dieser Terminologie dann den Quellterm. Die Schwierigkeit beim Lösen kommt erst durch den zusätzlichen nichtlinearen (Massen-)Term hinein. In den meisten bisherigen Arbeiten wie [HHR99] [MV99] tauchen an dieser Stelle lineare Massenterme auf. Alle resultierenden Gleichungssyteme sind linear, da sie aus der Minimierung der quadratischen Fehlerfunktionale hervorgehen. Die

Hoffnung ist nun, dass beim hier vorgestellten MAP-Ansatz durch die Nichtlinearität zusätzliche Information einfließt, so dass ein besseres Lernverfahren entsteht.

Um mehr Übersichtlichkeit zu schaffen, führen wir noch ein paar Hilfnotationen ein. Sei

$$F(u)(x,y) := \frac{\exp(u(x,y))}{\int \exp(u(y,z))dz} = p(y|x)$$

die gesuchte Wahrscheinlichkeitsdichte und

$$[v,u] := \frac{1}{n}\sum_{i=1}^{n}\int_Y v(x_i,y)u(x_i,y)dy$$

ein weiteres Skalarprodukt auf V. Den Quellterm schreiben wir als

$$q(v) := \frac{1}{n}\sum_{i=1}^{n} v(x_i,y_i).$$

Damit ergeben sich die Galerkin-Gleichungen zu

$$a(u,v) + [v,F(u)] = q(v), \quad \forall v \in H. \tag{5.2.5}$$

Wie man bereits hier sieht, entsteht die Hauptschwierigkeit beim Lösen der Galerkin-Gleichungen durch den nichtlinearen Term $[v,F(u)]$. Dieser enthält unter anderem ein Integral über die Exponentialfunktion der Parameterfunktion u, welches in den meisten Fällen nicht analytisch lösbar ist. Daher ist es notwendig diese Gleichungen nach geeigneter Diskretisierung numerisch zu lösen.

Regularisierungsnetzwerke und der MAP-Ansatz

Interressant ist an dieser Stelle der Zusammenhang zwischen der hier vorgestellten Maximum a-posteriori Methode und anderen klassischen Verfahren der Lerntheorie. Wählt man als Exponentialfamilie eine Normalverteilung mit

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(u(x)-y)^2}{2\sigma^2}\right),$$

so hat die logarithmierte Likelihood-Funktion die Gestalt

$$\begin{aligned} -\log(L(u)) =: l(u) &= \sum_{i=1}^{n}\log(p(y_i|x_i)) \\ &= \sum_{i=1}^{n}\left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(u(x_i)-y_i)^2\right) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(u(x_i)-y_i)^2. \end{aligned}$$

Als Zielfunktional resultiert

$$J(u) = \|u\|_H^2 + \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(u(x_i)-y_i)^2.$$

Vernachlässigt man an dieser Stelle den für die Minimierung irrelevanten zweiten, konstanten Term, so hat man wieder die klassische Form einer Regularisierung nach TIKHONOV, wie sie auch bei *smoothing Splines* oder *Regularisierungsnetzwerken* auftritt:

$$J'(u) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (u(x_i) - y_i)^2 + \|u\|_H^2.$$

Dies zeigt wie mächtig der Maximum a-posteriori Ansatz ist. Ist die gesuchte Dichte normal, so entsteht ein Minimierungsproblem mit quadratischem Fehlerfunktional und Glättungsterm $\|u\|_H^2$, wobei die Varianz σ^2 die Rolle des Regularisierungsparameters übernimmt.

5.2.3 Dichteschätzung mit der MAP-Methode

Bisher wurden Galerkin-Gleichungen und Kerndarstellung für ein allgemeines Lernproblem aufgestellt. Nun werden diese für den Spezialfall der Dichteschätzung betrachtet.

Kerndarstellung

Das Zielfunktional J hat für ein Dichteschätzungsproblem die Gestalt

$$J(u) = \|u\|_H^2 + n \log \int \exp(u(x))dx - \sum_{i=1}^{n} u(x_i) \tag{5.2.6}$$

und man erhält somit den Gradienten der logarithmierten Likelihood-Funktion als

$$\nabla l(u, D) = -\sum_{i=1}^{n} \delta_{x_i} + \frac{n \int_X \exp(u(x))\delta_x dx}{\int_X \exp(u(x))dx}.$$

Die zu lösenden Gleichungen sind demnach

$$u(x) + \frac{n}{2} \frac{\int_X \exp(u(y))k(x,y)dy}{\int_X \exp(u(y))dy} = \frac{1}{2} \sum_{i=1}^{n} k(x_i, x)$$

für einen Kern k. Diese nichtlinearen Fredholmschen Integralgleichungen der zweiten Art, sind nun mit einem geeigneten numerischen Verfahren zu lösen. Einen Überblick möglicher Lösungsverfahren gibt das Buch [Hac89].

Alternativ gibt es wiederum die Darstellung der Lösung in einem Funktionenraum.

Funktionenraumdarstellung

Nach Minimierung von (5.2.6) resultieren in der schwachen Form die Galerkin-Gleichungen

$$a(\varphi_j, u) + \frac{\int_X \varphi_j(x) \exp(u(x))dx}{\int_X \exp(u(z))dz} = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(x_i) \tag{5.2.7}$$

mit bekannter Bilinearform $a(v,u) = \lambda\langle Tv, Tu\rangle$ für eine Funktionenraumbasis $\{\varphi_j\}_{j=1}^{\infty}$.

Die Hilfnotationen lauten nun

$$F(u)(x) := \frac{\exp(u(x))dx}{\int \exp(u(z))dz}$$

und

$$q(v) := \frac{1}{n}\sum_{i=1}^{n} v(x_i).$$

Also können die Gleichungen in verkürzter Form geschrieben werden als

$$a(u,v) + \langle v, F(u)\rangle = q(v), \quad \forall v \in H. \tag{5.2.8}$$

Dabei ist wieder mit $\langle\cdot,\cdot\rangle$ das L^2-Skalarprodukt bezeichnet.

Diskussion

Insgesamt haben wurden zwei verschiedene Darstellungen für das Maximum-a-posteriori-Verfahren mit Gauß-Prior zur Dichteschätzung gezeigt.

Im ersten Fall war

$$u(x) + \frac{n}{2}\frac{\int_X k(x,y)\exp(u(y))dy}{\int_X \exp(u(y))dy} = \frac{1}{2}\sum_{i=1}^{n} k(x_i,x) \tag{5.2.9}$$

für Kernfunktionen k zu lösen. Beim anderen Ansatz entstehen in der schwachen Form die Galerkin-Gleichungen für eine Funktionenraumbasis $\{\varphi_i\}_{i=1}^{N}$:

$$a(\varphi_j,u) + \frac{\int_X \varphi_j(x)\exp(u(x))dx}{\int_X \exp(u(z))dz} = \frac{1}{n}\sum_{i=1}^{n}\varphi_j(x_i) \tag{5.2.10}$$

mit einer Bilinearform $a(\cdot,\cdot)$.

Die Frage ist nun, für welchen der beiden Ansätze man sich entscheidet. Bei beiden Ansätzen sind Integrale zu lösen, für die es keine explizite analytische Lösung gibt. Insofern muss in beiden Fällen zur Integration ein numerisches Verfahren verwendet werden. Des weiteren sind beides nichtlineare Gleichungen in u, so dass eine direkte Lösung nicht möglich ist und ein iteratives Verfahren verwendet werden muss. Danach kann man jedoch die üblichen Unterschiede zwischen Kernverfahren und Funktionenraumverfahren feststellen.

Ein klarer Vorteil des datenbasierten Verfahrens ist, dass dieses nach Wahl eines Kernes die exakte Lösung bestimmt und nicht erst noch diskretisiert werden muss. Dafür ist die entstehende Matrix allerdings in der Regel voll besetzt und der Aufwand zur Lösung ist von der Ordnung $O(n^3)$. Es ist jedoch möglich diese etwas zu verbessern, indem man die Kerne nur approximativ aufstellt, wodurch jedoch das Argument der exakten Lösung wieder abgeschwächt wird. Trotzdem ist das Verfahren somit für große Datenmengen nicht geeignet, da die Lösung des Gleichungssystems mit der $n \times n$ Kernmatrix zu viel Rechenzeit benötigt.

Das gitterbasierte Verfahren hingegen skaliert nur linear mit der Anzahl an Datenpunkten. Entscheidend für die Komplexität ist hier die Auflösung des

verwendeten Gitters bzw. das Diskretisierungslevel in der jeweiligen Dimension. Für einen d-dimensionalen Datenraum ergibt sich demnach eine Komplexität von $O(N^d)$, falls in jeder Dimension ein äquidistantes volles Gitter verwendet wird. Somit hängt es vom Verhältnis von n^3 zu N^d ab, für welche Variante man sich entscheiden sollte. Hat man eine mäßig große Anzahl von Datenpunkten in hohen Dimensionen vorliegen, ist die Kernvariante die bessere Wahl. Ist die Dimension moderat und hat man sehr viele Datenpunkte, sollte man sich für die Lösung im diskretisierten Funktionenraum entscheiden.

Mit Hilfe dünner Gitter, wie sie in Kapitel 6.3 vorgestellt werden, lässt sich dieses Verhältnis jedoch zu Gunsten der Gitterlösung verschieben. Diese hat dann nur noch eine Komplexität von $O(N \cdot \log(N)^{(d-1)})$. Falls die Lösung genügend glatt ist, erreicht die Dünngitter-Lösung dabei eine Genauigkeit von $O(N^{-2} \cdot \log(N)^{d-1})$ im Vergleich zu $O(N^{-2})$ für ein volles Gitter.

Da man in den meisten Anwendungen nach geeigneter Vorverarbeitung die Dimension des Datenraumes auf eine moderate Anzahl an Dimensionen reduzieren kann (siehe [Car97]), konzentrieren wir uns hier im Folgenden auf eine Diskretisierung mit geeigneten Basen eines Funktionenraums. Genauer werden wir die so genannte *Kombinationstechnik* für dünne Gitter verwenden. Man ist damit in der Lage mit sehr großen, moderat hochdimensionalen Datensätzen zu arbeiten.

Im folgenden Kapitel stellen wir mögliche iterative Lösungswege vor mit denen man die Nichtlinearität behandeln kann.

6 Diskretisierung und Lösung

Um den im vorherigen Kapitel hergeleiteten Ansatz lösen zu können, muss man sich zunächst Gedanken über ein geeignetes iteratives Verfahren zur näherungsweisen Lösung machen. Dazu werden hier zwei verschiedene Ansätze vorgestellt. Erst ein recht naiver, voll expliziter Ansatz, danach ein etwas aufwendigeres Verfahren, welches den nichtlinearen Massenterm quadratisch approximiert.

Im letzten Teilkapitel werden dünne Gitter als Diskretisierungstechnik vorgestellt, die in unserem Verfahren mittels der Kombinationstechnik implementiert werden.

6.1 Voll expliziter Lösungsansatz

Als bestimmende Gleichungen für den gesuchten Dichteschätzer ergaben sich im vorherigen Kapitel:

$$a(u,v) + \langle v, F(u)\rangle = q(v), \quad \forall v \in H.$$

Dabei liegt offensichtlich die Schwierigkeit beim Lösen dieser Gleichungen in dem nichtlinearen Term

$$\langle v, F(u)\rangle = \frac{\int v(x)\exp(u(x))dx}{\int \exp(u(z))dz}. \tag{6.1.1}$$

Um dieses Problem zu behandeln, ist ein erster Ansatz ein explizites Iterationsverfahren.

6.1.1 Der Algorithmus

Ausgehend von einem Startwert u^0 wird jeweils die alte Iterierte in den nichtlinearen Term eingesetzt und auf die rechte Seite gebracht. Das entstehende Gleichungssystem ist somit nur noch linear in der neuen Iterierten und kann daher mit geeigneten Verfahren gelöst werden. Die Gleichungen lauten demnach in jeder Iteration $(k = 1, ..., N)$

$$\sum_{j=1}^{N} u_j^{i+1} a(\varphi_k, \varphi_j) = \frac{1}{n}\sum_{i=1}^{n} \varphi_k(x_i) - \int \varphi_k(x) \frac{\exp(\sum_{j=1}^{N} u_j^i \varphi_j(x))}{\int \exp(\sum_{j=1}^{N} u_j^i \varphi_j(z))dz} dx,$$

wobei u_j^i der j-te Koeffizient der Basisdarstellung der i-ten Iterierten u^i ist. In Matrixform geschrieben lautet das in jeder Iteration zu lösende Gleichunssystem:

$$\mathcal{A}\boldsymbol{u}^{i+1} = \boldsymbol{q} - \boldsymbol{\Phi}(\boldsymbol{u}^i)$$

mit so genannter Regularisierungsmatrix $\mathcal{A}_{i,j} = a(\varphi_i, \varphi_j)$ und Massenvektor $\boldsymbol{\Phi}(\boldsymbol{u})_k = \langle \varphi_k, F(u) \rangle$. Dabei bezeichnet $\boldsymbol{u}$ nun den Koeffizientenvektor der Funktion u bezüglich einer geeigneten Funktionenraumbasis $\{\varphi_j\}_{j=1}^N$. Nach Wahl eines Startvektors $\boldsymbol{u}^0$ ergibt sich die jeweils nächste Iterierte als

$$\boldsymbol{u}^{i+1} = \boldsymbol{u}^i + \omega \mathcal{A}^{-1}(\boldsymbol{q} - \boldsymbol{\Phi}(\boldsymbol{u}^i)) \tag{6.1.2}$$

mit Dämpfungsparameter $\omega > 0$. Die Iteration wird abgebrochen, falls das Residuum $r^i = \mathcal{A}\boldsymbol{u}^i - \boldsymbol{q} + \boldsymbol{\Phi}(\boldsymbol{u}^i)$ kleiner als eine vorgegebene Toleranz ist oder eine maximale Anzahl von Iterationen erreicht wird. Hier löst man also in jeder Iteration ein lineares Gleichungssytem, wobei die rechte Seite jeweils durch die alte Iterierte den nichtlinearen Term beinhaltet. Mit dem Dämpfungsparameter ω lässt sich eine Konvergenz des Verfahrens erzwingen. In Algorithmus 1 ist das hier vorgeschlagene Verfahren noch mal dargestellt.

Eingabe : Datenpunkte $\{x_i\}_{i=1}^n$ Basisfunktionen $\{\varphi_j\}_{j=1}^N$
Resultat : Dichtefunktion $\hat{f}$
Regularisierungsmatrix $\mathcal{A}_{i,j} = a(\varphi_i, \varphi_j)$ aufstellen ;
Festen Anteil der rechten Seite $\boldsymbol{q}$ aufstellen ;
Wähle Startvektor $\boldsymbol{u}^0 = 0$;
Setze Dämpfungsparameter $\omega > 0$, Toleranz ϵ und i_{max};
$i = 0$; $r = \|\mathcal{A}\boldsymbol{u}^0 - \boldsymbol{q} + \boldsymbol{\Phi}(\boldsymbol{u}^0)\|$;
while $r \geq \epsilon$ & $i \leq i_{max}$ **do**
  $i = i + 1$;
  $\boldsymbol{b} = \boldsymbol{q} - \boldsymbol{\Phi}(\boldsymbol{u}^{i+1})$;
  Löse $\mathcal{A}\boldsymbol{u}^i = b \Rightarrow \boldsymbol{u}^i$;
  $\boldsymbol{u}^i = \boldsymbol{u}^{i-1} + \omega \boldsymbol{u}^i$;
  $r = \|\mathcal{A}\boldsymbol{u}^i - \boldsymbol{q} + \boldsymbol{\Phi}(\boldsymbol{u}^i)\|$;
end
$\hat{f} = F(\boldsymbol{u}^i)$;

Algorithmus 1 : Explizites Lösungsverfahren

6.1.2 Eigenschaften des Verfahrens

Da eine theoretische Analyse der Konvergenzeigenschaft aufgrund des komplexen Zusammenspiels der Parameter und der Gestalt des nichtlinearen Terms nicht direkt möglich ist, beschränken wir uns auf experimentelle Untersuchungen. Dabei hat sich herausgestellt, dass das Verfahren sehr stark gedämpft werden muss ($\omega \leq 0.1$) um überhaupt eine Konvergenz zu erreichen. Mit einem geringen ω erzwingt man diese zwar, aber die Iterierte nähert sich nur sehr langsam der Lösung.

Für die folgenden Ergebnisse wurde ein Datensatz aus 100 Punkten nach einer aus zwei Normalverteilungen bestehenden Dichtefunktion erzeugt. Gewählt wurde hier $\mu_1 = 4, \mu_2 = 7$ mit $\sigma_1 = 1, \sigma_2 = 1/4$. Anschließend wurden die erzeugten Daten auf das Einheitsintervall $[0, 1]$ transformiert. Genauer beschrieben ist diese Datenerzeugung in Kapitel 4.4.1. Die Schwierigkeit besteht hierbei

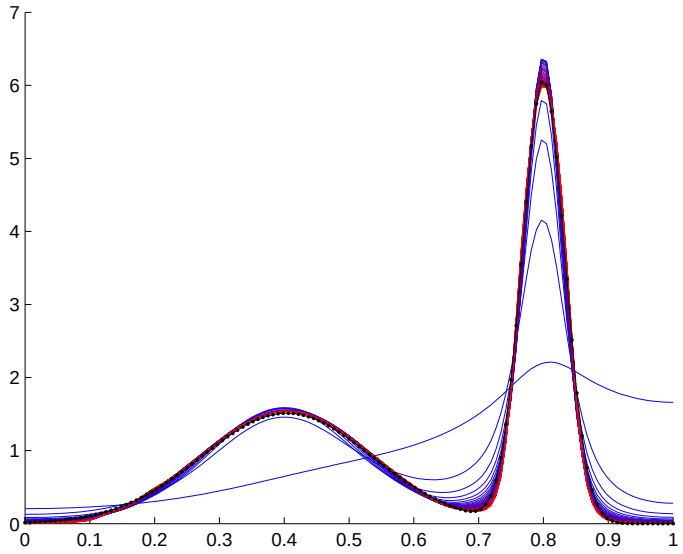

Abbildung 6.1: Lösung der expliziten Fixpunktiteration mit $\omega = 0.1$. Abgebildet ist jede zehnte der dreihundert ersten Iterationen. Die Originaldichte ist als schwarze gepunktete Linie dargestellt.

in den deutlich verschiedenen Varianzen der beiden Gaußglocken. Ein gutes Verfahren sollte beide Spitzen der Verteilung gut approximieren und trotzdem in der Breite nicht zu oszillieren anfangen. Weiterhin muss eine saubere Trennung zwischen den beiden Gaußglocken erfolgen.

In Abbildung 6.1 ist für diesen eindimensionalen synthetischen Datensatz jede zehnte Zwischenlösung der Iteration eingezeichnet. Als Parameter wurden dabei $\lambda = 5$ und $\omega = 0.1$ gesetzt. Die verwendeten Basisfunktionen sind kubische B-Splines mit dem Gradienten als Regularisierungsoperator. Als Gitter für die Ansatzfunktonen ist das Einheitsintervall in 2^5 Teilintervalle unterteilt worden, auf dem jeweils die kubischen B-Splines angesetzt werden. Die in den Gleichungen auftretenden Integrale werden auf einem verfeinertem Gitter mit Hilfe der Simpsonregel numerisch bestimmt.

Man sieht, dass die Originaldichte von Iteration zu Iteration zwar immer besser approximiert wird, sich der Schätzer aber nur sehr langsam der Lösung der nichtlinearen Gleichungen nähert. In der linken Grafik von Abbildung 6.2 ist das Residuum in der $l2$-Norm abgebildet, während die rechte Grafik den mittleren quadratischen Fehler (MSE) auf dem verfeinertem Diskretisierungsgitter darstellt. Das Residuum fällt zu Beginn zwar relativ schnell, flacht mit höheren Iterationen aber immer stärker ab. Ähnlich verhält sich der gemessene Fehler, welcher ungefähr bei der 50. Iteration schon relativ gering ist, später aber wieder leicht ansteigt.

Aufgrund der hohen benötigten Anzahl an Iterationen und der damit verbundenen Rechenzeit, ist das hier beschriebene Verfahren für höhere Dimensionen

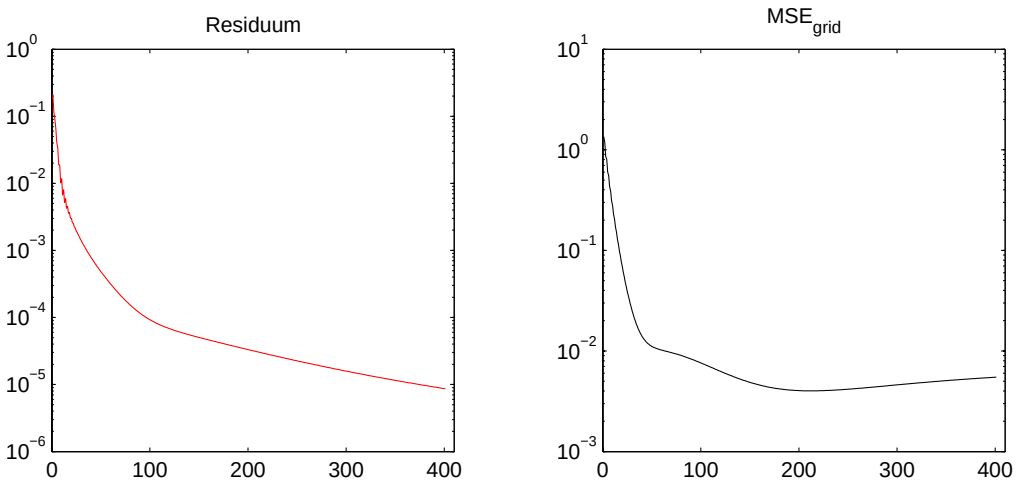

Abbildung 6.2: Links Residuum, rechts mittlerer $l2$-Fehler auf dem Diskretisierungsgitter jeweils gegen Anzahl Iterationen geplottet.

nicht geeignet. Trotzdem stellt man fest, dass der Grenzwert - also der gesuchte Maximum-a-posteriori-Schätzer - die Originaldichte sehr gut rekonstruiert. Man benötigt nur ein schnelleres Verfahren um diesen zu bestimmen.

Hauptproblem scheint dabei zu sein, dass der nichtlineare Term $\boldsymbol{\Phi}(\boldsymbol{u})$ den gewichtigeren Teil der Gleichungen ausmacht. Das explizite Verfahren approximiert diesen durch eine Konstante, welche dann in die rechte Seite mit einfließt. Scheinbar ist eine konstante Approximation aber noch zu ungenau, um in angemessener Zeit eine Lösung zu finden. Daher muss man sich Gedanken über Approximationen höherer Ordnung machen.

6.2 Quadratische Approximation der Nichtlinearität

Ein natürliches Vorgehen zur Verbesserung der Approximationsgüte ist eine lineare Näherung an den nichtlinearen Massenterm. Dazu wird eine quadratische Approximation an das ursprüngliche Zielfunktional $J(u)$ bestimmt, die nach Minimierung in ein lineares Gleichungssystem resultiert. Im Folgenden wird diese Vorgehensweise näher erläutert.

6.2.1 Die Approximation

Gesucht ist nach wie vor ein Minimum über alle $u \in H$ des Funktionals

$$J(u) = \|u\|_H^2 + n \log \int_Y \exp(u(x))dx - \sum_{i=1}^{n} u(x_i).$$

Wir bilden nun analog zu einem Newton-Verfahren eine lokal quadratische Näherung an J, sodass nach Minimierung ein lineares Gleichungssystem entsteht.

Der zu approximierende, nichtlineare Term des Funktionals $J(u)$ hat die Gestalt

$$\log \int \exp(u(x))dx.$$

Um diesen besser handhaben zu können, nutzen wir die Reihenentwicklungen des Logarithmus und der Exponentialfunktion und schneiden nach der zweiten Ordnung ab:

$$\exp(a+b) = \exp(a)(1+b+\frac{1}{2}b^2+O(b^3))$$
$$\log(a+b) = \log(a)+\frac{b}{a}-\frac{1}{2}\frac{b^2}{a^2}+O(\frac{b^3}{a^3}).$$

Damit ergibt sich für jedes $u \in H$ und kleine $h \in H$

$$\log \int \exp(u+h)(x)dx = \log \int \exp(u(x))(1+h(x)+\frac{1}{2}h^2(x))dx + O(h^3)$$

und im nächsten Schritt, wobei das Argument x der Übersichtlichkeit halber nicht angegeben ist

$$\log \int \exp(u+h)dx = \log \int \exp(u)dx + \frac{\int h\exp(u)dx}{\int \exp(u)dx} + \frac{1}{2}\frac{\int h^2\exp(u)dx}{\int \exp(u)dx} - \frac{1}{2}\left(\frac{\int h\exp(u)dx}{\int \exp(u)dx}\right)^2 + O(h^3).$$

Lassen wir an dieser Stelle alle Terme ab Ordnung h^3 weg, so erhalten wir ein quadratisches Funktional in h, welches das ursprüngliche Funktional lokal approximiert. Der Fehler hierbei ist dementsprechend von der Ordnung $O(h^3)$.

6.2.2 Das lokal quadratische Zielfunktional

Wir suchen nun ein Minimum von

$$M_u(h) := \frac{1}{n}(J(u+h)-J(u)-O(h^3)).$$

Die in h konstante Verschiebung $J(u)$ verändert dabei die Lage des Minimums nicht, führt aber zu einer übersichtlicheren Darstellung.

Mit obigen Approximationen schreiben wir unser neues Zielfunktional als

$$\begin{aligned}
M_u(h) =& \frac{1}{n}(\|u+h\|_H^2 - \|u\|_H^2) + \frac{\int h\exp(u)dx}{\int \exp(u)dx} + \frac{1}{2}\frac{\int h^2\exp(u)dx}{\int \exp(u)dx} \\
& - \frac{1}{2}\left(\frac{\int h\exp(u)dx}{\int \exp(u)dx}\right)^2 - \frac{1}{n}(\sum_{i=1}^n (u+h)(x_i) - \sum_{i=1}^n u(x_i)) \\
=& \frac{1}{n}(\|u\|_H^2 + 2(u,h)_H) + \langle h, F(u)\rangle + \frac{1}{2}[h^2, F(u)] \\
& - \frac{1}{2}\langle h, F(u)\rangle^2 - \frac{1}{n}\sum_{i=1}^n h(x_i) \\
=& \frac{1}{2}a(h,h) + a(u,h) + \langle h, F(u)\rangle + \frac{1}{2}\langle h^2, F(u)\rangle - \frac{1}{2}\langle h, F(u)\rangle^2 - q(h).
\end{aligned}$$

Notwendige Bedingung für ein Minimum ist das Verschwinden der ersten Ableitung. Dazu stellen wir h bezüglich einer Funktionenraumbasis $\{\varphi_j\}_{j=1}^N$ eines endlichen Teilraumes $H_N \subset H$ dar:

$$\begin{aligned} M_u(h) =& \frac{1}{2}a(\sum_j \alpha_j\varphi_j, \sum_j \alpha_j\varphi_j) + a(u, \sum_j \alpha_j\varphi_j) + \langle \sum_j \alpha_j\varphi_j, F(u)\rangle \\ &+ \frac{1}{2}\langle (\sum_j \alpha_j\varphi_j)^2, F(u)\rangle - \frac{1}{2}\langle \sum_j \alpha_j\varphi_j, F(u)\rangle^2 - q(\sum_j \alpha_j\varphi_j). \end{aligned}$$

Differentiation in den Koeffizienten α_k, $k = 1, ..., N$ ergibt:

$$\begin{aligned} \frac{\partial M_u(h)}{\partial \alpha_k} =& a(\varphi_k, \sum_j \alpha_j\varphi_j) + a(u, \varphi_k) + \langle \varphi_k, F(u)\rangle + \langle \sum_j \alpha_j\varphi_j\varphi_k, F(u)\rangle \\ &- \langle \sum_j \alpha_j\varphi_j, F(u)\rangle\langle \varphi_k, F(u)\rangle - q(\varphi_k) \stackrel{!}{=} 0. \end{aligned}$$

Dies ist äquivalent zu:

$$\begin{aligned} \sum_j \alpha_j \Big(a(\varphi_k, \varphi_j) + \langle \varphi_j\varphi_k, F(u)\rangle - \langle \varphi_j, F(u)\rangle\langle \varphi_k, F(u)\rangle \Big) \\ = q(\varphi_k) - \langle \varphi_k, F(u)\rangle - a(u, \varphi_k). \end{aligned}$$

In analoger Notation zum vorherigen Kapitel erhalten wir für alle Testfunktionen $v \in H$ die Galerkin-Gleichungen:

$$a(v, h) + \langle vh, F(u)\rangle - \langle v, F(u)\rangle\langle h, F(u)\rangle = q(v) - \langle v, F(u)\rangle - a(u, v). \quad (6.2.1)$$

6.2.3 Der Algorithmus

Die betrachteten Gleichungen lassen sich in Matrixschreibweise darstellen als

$$(\mathcal{A} + \mathcal{B}_{\boldsymbol{u}} - \boldsymbol{\Phi}_{\boldsymbol{u}}\boldsymbol{\Phi}_{\boldsymbol{u}}^t) \cdot \boldsymbol{\alpha} = \boldsymbol{q} - \boldsymbol{\Phi}_{\boldsymbol{u}} - \mathcal{A}\boldsymbol{u}. \quad (6.2.2)$$

Dabei ist $\mathcal{A}$ die quadratische, positiv definite *Regularisierungsmatrix*

$$(\mathcal{A})_{j,k} = a(\varphi_j, \varphi_k) = \lambda\langle T\varphi_j, T\varphi_k\rangle, \quad j, k = 1, ..., N$$

mit Differentialoperator T. $\mathcal{B}_{\boldsymbol{u}}$ ist ebenfalls quadratisch mit Einträgen

$$\{(\mathcal{B})_{\boldsymbol{u}}\}_{j,k} = \langle \varphi_j\varphi_k, F(u)\rangle, \quad j, k = 1, ..., N.$$

Massenvektor und *Datenvektor* sind analog zum vorherigen Abschnitt von der Gestalt

$$(\boldsymbol{\Phi}_{\boldsymbol{u}})_j = \langle \varphi_j, F(u)\rangle, \quad q_j = \sum_{i=1}^M \varphi_j(x_i), \quad j = 1, ..., N.$$

Mit $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)^t$ und $\boldsymbol{u} = (u_1, ..., u_N)^t$ seien die Koeffizientenvektoren von h beziehungsweise u bezüglich der gewählten Funktionenraumbasis bezeichnet.

An die Stelle von $\boldsymbol{u}$ wird nun ausgehend von einem Startvektor (üblicherweise wählt man $\boldsymbol{u}^0 = \mathbf{0}$) jeweils die gegebene Iterierte eingesetzt und die neue Iterierte als $\boldsymbol{u}^{i+1} = \boldsymbol{u}^i + \omega\boldsymbol{\alpha}$ bestimmt. Dabei ist ω wiederum ein in Abhängigkeit von $\mathcal{A}$ zu wählender Dämpfungsparameter. Es ist demnach in jeder Iteration ein Gleichungssystem zu lösen, welches linear in den Koeffizienten von h ist. Als rechte Seite steht hierbei gerade das Residuum der ursprünglichen Galerkin-Gleichungen (5.2.5), welches Null wird für die exakte Lösung u.

In Algorithmus 2 ist das komplette Verfahren noch einmal übersichtlich dargestellt.

Eingabe : Datenpunkte $\{x_i\}_{i=1}^n$ Basisfunktionen $\{\varphi_j\}_{j=1}^N$
Resultat : Dichtefunktion $\hat{f}$
Regularisierungsmatrix $\mathcal{A}_{i,j} = a(\varphi_i, \varphi_j)$ mit Differentialoperator T und Glättungsparameter λ aufstellen ;
Festen Anteil der rechten Seite $\boldsymbol{q}$ aufstellen ;
Wähle Startvektor $\boldsymbol{u}^0 = \mathbf{0}$;
Setze Toleranz ϵ, Dämpfungsparameter ω und i_{max};
$i = 0$, $r = 2\epsilon$;
while $r \geq \epsilon$ & $i \leq i_{max}$ **do**
  Vektor $\boldsymbol{\Phi}_{\boldsymbol{u}^i}$ aufstellen;
  Matrix $\mathcal{B}_{\boldsymbol{u}^i}$ berechnen;
  $\boldsymbol{b} = \boldsymbol{q} - \boldsymbol{\Phi}_{\boldsymbol{u}^i} - \mathcal{A}\boldsymbol{u}^i$;
  $\mathcal{C} = \mathcal{A} + \mathcal{B}_{\boldsymbol{u}^i} - \boldsymbol{\Phi}^t_{\boldsymbol{u}^i}\boldsymbol{\Phi}_{\boldsymbol{u}^i}$;
  Löse $\mathcal{C}\boldsymbol{\alpha} = \boldsymbol{b} \Rightarrow \boldsymbol{\alpha}$;
  $\boldsymbol{u}^{i+1} = \boldsymbol{u}^i + \omega \cdot \boldsymbol{\alpha}$;
  $r = \|\boldsymbol{q} - \boldsymbol{\Phi}_{\boldsymbol{u}^{i+1}} - \mathcal{A}\boldsymbol{u}^{i+1}\|$;
  $i = i + 1$;
end
$\hat{f} = F(\boldsymbol{u}^i)$;

Algorithmus 2 : Quadratische Approximation

6.2.4 Konvergenz der Iteration

Analog zum vorherigen Abschnitt über das explizite Verfahren, wollen wir die Konvergenz des Verfahrens analysieren. Dazu untersuchen wir den iterativen Löser anhand des oben vorgestellten Datensatzes.

Wie in Abbildung 6.3 zu sehen sieht, konvergiert das Newton-ähnliche Verfahren wesentlich schneller als die einfache explizite Iteration. Bereits nach wenigen Iterationsschritten verändert sich der Schätzer nur noch minimal. Allerdings hat hierbei der Dämpfungsparameter ω großen Einfluss auf die Konvergenz. In Abbildung 6.4 sind das jeweilige Residuum und der mittlere quadratische Fehler auf den Gitterpunkten dargestellt. Der Fehler ist bereits nach fünf Iterationen minimal und ändert sich anschließend nicht mehr merklich. Das Residuum fällt nach kurzem Anstieg sehr schnell ab.

Analog zum Newton-Verfahren hängt die Konvergenz stark von der Wahl des

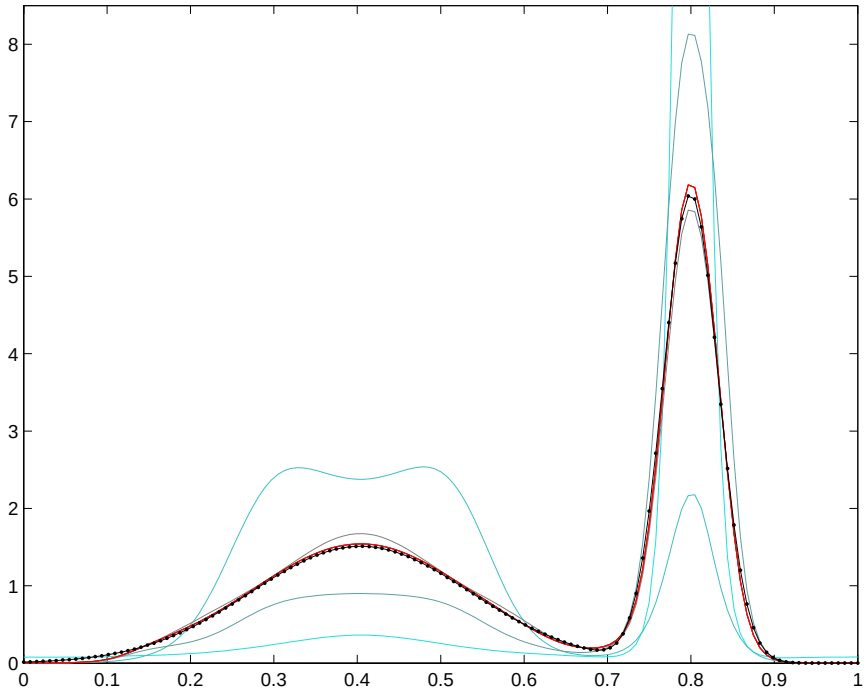

Abbildung 6.3: Die Iterierten 1 (cyan) bis 8 (rot) des Approximationsverfahrens auf Level 5 mit Parametern $\lambda = 0.01$ und $\omega = 1$. Die schwarze, gepunktete Linie stellt die zu rekonstruierende Originaldichte dar.

Startwertes ab. Falls der Startwert genügend nahe an der gesuchten Lösung liegt, konvergiert die Iteration quadratisch gegen die Lösung. Wie günstig der Startwert gewählt ist, hängt dabei auch von der vorgegebenen Regularität ab. Falls λ sehr klein gewählt wird, oszilliert die Lösung entsprechend stark, so dass ein standardmäßig gewählter Startwert von $\boldsymbol{u}^0 = 0$ nicht optimal ist. Unter Umständen liegt dieser zu weit von der Lösung entfernt, sodass nur mit einem Dämpfungsparameter $\omega < 1$ eine Konvergenz erreicht wird.

Im Gegensatz zum nicht gedämpften Verfahren konvergiert die Iterierte für $\omega < 1$ jedoch wesentlich langsamer gegen die Lösung. In dem Fall erreicht das Verfahren nur noch eine Konvergenzornung von 1 (siehe Abbildung 6.5 rechts). Dafür kann mit geeigneter Dämpfung für beliebige Startwerte, Parameter λ und Level immer eine Konvergenz erzwungen werden. Ohne Dämpfung ist eine Konvergenz nicht garantiert. Liegt der Startwert zu weit von der Lösung weg, divergiert die Iteration unter Umständen.

In Abbildung 6.5 ist dieser Sachverhalt veranschaulicht. In der linken Grafik ist für verschiedene Level der Regularisierungsparameter variiert und der Fehler als Summe aus Daten- und Gitterfehler gemessen worden. Auffällig ist dabei, dass das ungedämpfte Verfahren (die gefüllten Kreise) bei höherem Level erst ab einem gewissen Schwellwert konvergiert, während die gedämpfte Variante für alle λ eine Lösung liefert. Dieser Schwellwert liegt interessanterweise meist direkt am Minimum. Eine Möglichkeit könnte also sein, immer mit der ungedämpften Variante zu arbeiten und jeweils das minimale λ zu wählen, für welches diese

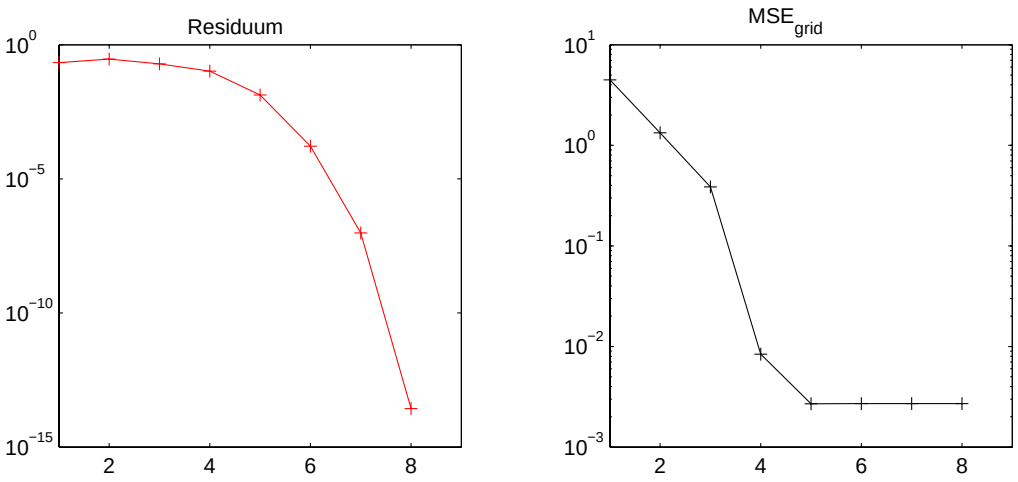

Abbildung 6.4: Residuum der MAP-Methode mittels quadratischer Approximation auf Level 5 ($\lambda = 0.01$), sowie der entsprechende mittlere quadratische Fehler auf dem Gitter gegen die jeweilige Iterierte

konvergiert.

In der rechten Grafik von Abbildung 6.5 ist für das jeweilig optimale λ die Entwicklung des Residuums über die Iterationen dargestellt. Das Verfahren benötigt hierbei für $\omega = 0.8$ zwischen 22 und 24 Iterationen bis der Fehler unter 10^{-10} fällt, während das ungedämpfte Verfahren schon nach circa 8 bis 9 Iterationen diese Grenze deutlich unterschreitet.

Allgemein kann man festhalten, dass ab einem gewissen Schwellwert für den Regularisierungsparameter das nichtgedämpfte Verfahren ($\omega = 1$) divergiert und um eine Konvergenz zu erhalten entweder λ erhöht oder ω verringert werden muss. Bei kleinem ω verschlechtert sich die Konvergenz ähnlich wie bei einem gedämpften Newton-Verfahren deutlich. Zusätzlich ist bei höherem Diskretisierungslevel der Parameter λ entsprechend höher zu wählen, da ansonsten ähnlich wie bei einem Histogramm eine Überanpassung stattfindet.

Weitere Beispiele zu den hier beschriebenen Eigenschaften finden sich in Kapitel 7. Dabei stellt man fest, dass sich das Verfahren mit den gegebenen Parametern sehr gut steuern lässt und gute Ergebnisse liefert.

6.3 Diskretisierung mit dünnen Gittern

Für die numerische Behandlung der aufgestellten Gleichungen ist eine diskrete Darstellung des Funktionenraumes notwendig. Da sich mit geeigneter Transformation jedes beliebige Rechteckgebiet auf den Einheitswürfel $[0,1]^d$ abbilden lässt, werden wir uns bei den weiteren Ausführungen auf diesen beschränken.

Ein üblicher Ansatz zur Diskretisierung wählt ein äquidistantes Gitter der Maschenweite $h_l = 2^{-l}$ in jede Koordinatenrichtung und stellt darauf die Basisfunktionen auf, die den Raum V_N aufspannen. Mit l ist hier die Verfeinerungstiefe der Diskretisierung bezeichnet, oft auch Diskretisierungslevel genannt. Für

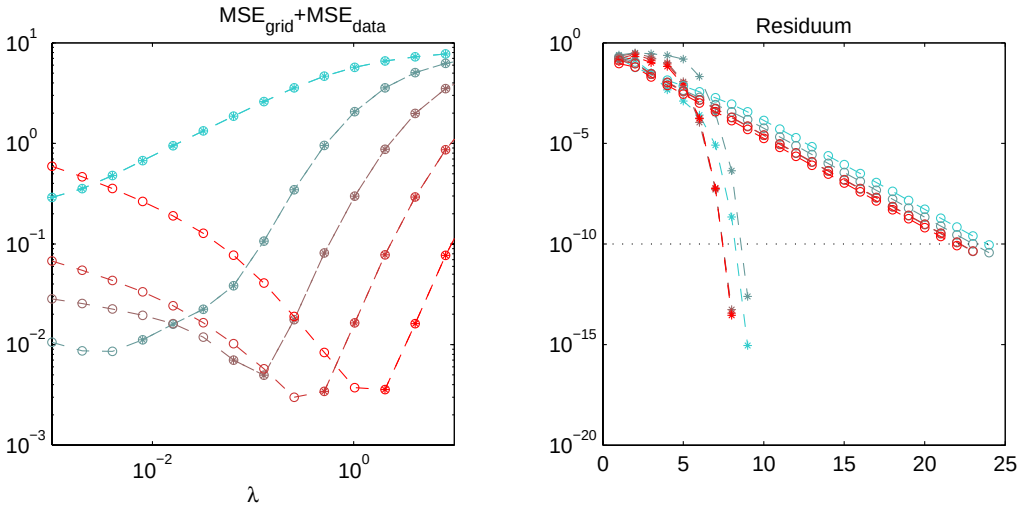

Abbildung 6.5: Links: Ergebnisse des mit $\omega = 0.8$ gedämpften Verfahrens (mit Kreisen markiert) im Vergleich zur nicht gedämpften Variante (gefüllte Kreise) für Level 3 (cyan) bis Level 7 (rot). Rechts: Für den jeweils besten Regularisierungsparameter λ die Entwicklung des Residuums für das gedämpfte (Kreise) bzw. das ungedämpfte Verfahren (mit Sternen markiert).

eine Genauigkeit von $O(h_l)$ entsteht somit eine Komplexität von $O(h_l^{-d})$ bei der Lösung der resultierenden Gleichungssysteme. Dieser exponentiell mit der Dimension wachsende Aufwand wird häufig als *Fluch der Dimension* bezeichnet. Die in diesem Abschnitt vorgestellten dünnen Gitter sind in der Lage diesen zu einem gewissen Grad zu umgehen, und ermöglichen es damit höherdimensionale Probleme zu behandeln.

Die Idee basiert auf einer hierarchischen Basis, einer zur üblichen nodalen Basis äquivalenten Darstellung, bei der gewisse den Fehler nur wenig beeinflussende Basen weggelassen werden. Es kann gezeigt werden, dass damit für eine genügend glatte Lösung bei einer Genauigkeit von $O(h_l^2 \cdot \log(h_l^{-1})^{d-1})$ lediglich ein Aufwand der Ordnung $O(h_l^{-1} \cdot \log(h_l^{(} - 1))^{(d-1)})$ nötig ist.

Realisiert werden dünne Gitter in dieser Arbeit mittels der so genannten *Kombinationstechnik*. Dabei werden auf einer Sequenz von anisotropen, in jeder Koordinatenrichtung äquidistanten Gittern Teillösungen bestimmt, aus denen mittels der Kombinationsformel die Gesamtlösung auf dem dünnen Gitter aufgestellt wird.

Erstmals in der Literatur erschienen ist ein ähnliches Verfahren bei dem russischen Mathematiker Smolyak, der in [Smo63] effiziente Methoden zur numerischen Integration untersucht. Explizit vorgestellt wurden dünne Gitter erstmals in [Zen91],[Bun92],[GSZ92] zur Lösung elliptischer, partieller Differentialgleichungen. Mittlerweile finden dünne Gitter in vielen Bereichen des wissenschaftlichen Rechnens Anwendung. In [Gar04] werden sie bereits für Regressions- bzw. Klassifikationsaufgaben eingesetzt. Des weiteren werden sie für Integration, Approximation und Interpolation, parabolische Differentialgleichungen, Eigenwert-

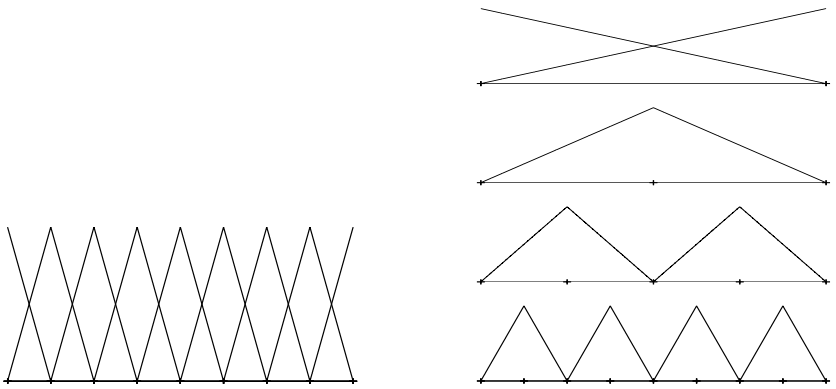

Abbildung 6.6: Vergleich einer nodalen (links) zu einer hierarchischen Basis (rechts) auf Level 3.

probleme, stochastische Differentialgleichungen in der Finanzmathematik und andere Problemstellungen genutzt. Einen guten Überblick über Dünne Gitter und deren Anwendung liefert der Artikel [BG04].

Die folgende Darstellung der Kombinationstechnik für Dünne Gitter ist eng angelehnt an [Gar04].

6.3.1 Hierarchische versus nodale Basis

Ein Standard-Finite-Elemente-Ansatz auf dem betrachteten Gebiet $\Omega = [0,1]^d$ verwendet ein Gitter $\Omega_{\boldsymbol{l}}$ mit Multi-Index $\boldsymbol{l} = (l_1, ..., l_d) \in \mathbb{N}^d$ und zugehörigen Maschenweiten $h_{\boldsymbol{l}} = (h_{l_1}, ..., h_{l_d})$ in jeder Dimension. Dabei haben die Maschenweiten die Gestalt $h_{l_i} = 2^{-l_i}$. Die Gitter sind demnach in jeder Koordinatenrichtung äquidistant, können jedoch verschieden feine Maschenweiten in den einzelnen Richtungen haben. Als Gitterpunkte der i-ten Koordinatenrichtung ergeben sich folglich $x_{l_i} = j_i \cdot h_{l_i} = j_i \cdot 2^{-l_i}\ i = 0, ..., 2^{l_i}$.

Auf jedem Gitter $\Omega_{\boldsymbol{l}}$ wird der Raum

$$V_{\boldsymbol{l}} = \text{span}\{\phi_{\boldsymbol{lj}} | j_i = 0, ..., 2^{l_i}, i = 1, ..., d\}$$

definiert, der durch die Funktionen $\phi_{\boldsymbol{lj}}$, welche sich als Tensorprodukt der eindimensionalen Basisfunktionen

$$\phi_{\boldsymbol{lj}}(x) = \prod_{i=1}^{d} \phi_{l_i, j_i}(x_i) \tag{6.3.1}$$

ergeben, aufgespannt wird. Für unsere Anwendung werden hier die in Definition 3.2.1 eingeführten eindimensionalen B-Splines verwendet.

Ein Beispiel einer solchen *nodalen* Basis mit linearen B-Splines ist in Abbildung 6.6 angegeben. Im Gegensatz dazu arbeiten dünne Gitter auf einer *hierarchisch* angeordneten Menge von Basisfunktionen.

Um diese einführen zu können, brauchen wir die folgenden Differenzenräume

$$W_{\boldsymbol{l}} := V_{\boldsymbol{l}} \setminus \bigcup_{i=1}^{d} V_{\boldsymbol{l}-\boldsymbol{e}_i}, \tag{6.3.2}$$

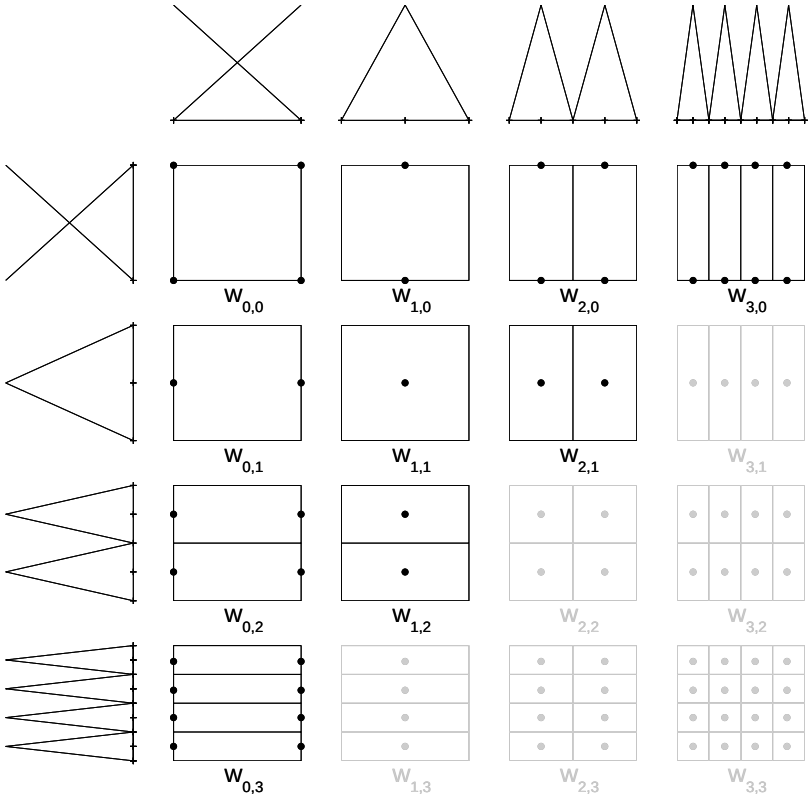

Abbildung 6.7: Die Träger der hierarchischen Basisfunktionen des Raumes V_3 mit Differenzenräumen $W_{0,0}$ (oben links) bis $W_{3,3}$ (unten rechts)

wobei $\boldsymbol{e}_i$ der i-te Einheitsvektor ist. Der Vollständigkeit halber setzen wir $V_{\boldsymbol{l}_i} = \emptyset$, falls $\boldsymbol{l}_i = -1$ für ein i. Mit diesen Differenzenräumen lässt sich der diskrete Raum V_N als direkte Summe

$$V_N = \bigoplus_{l_1=0}^{N} \cdots \bigoplus_{l_d=0}^{N} W_{\boldsymbol{l}} = \bigoplus_{|\boldsymbol{l}|_\infty \leq N} W_{\boldsymbol{l}}$$

schreiben, wobei $|\boldsymbol{l}|_\infty := \max_{1 \leq i \leq d} l_i$ die diskrete L^∞-Norm ist. Die Relation „$\leq$“ ist komponentenweise zu verstehen.

Analog lassen sich die Räume $V_{\boldsymbol{l}}$ als direkte Summe der entsprechenden Differenzenräume schreiben:

$$V_{\boldsymbol{l}} = \bigoplus_{t_1=0}^{l_1} \cdots \bigoplus_{t_d=0}^{l_d} W_{\boldsymbol{l}} = \bigoplus_{\boldsymbol{t} \leq \boldsymbol{l}} W_{\boldsymbol{t}}.$$

Mit Hilfe der Indexmenge

$$B_{\boldsymbol{l}} := \left\{ \boldsymbol{j} \in \mathbb{N}^d \middle| \begin{array}{ll} j_i = 1, ..., 2^{l_i} - 1, \ j_i \text{ ungerade}, & i = 1, ..., d, \text{ falls } l_i > 0 \\ j_i = 0, 2^{l_i}, & i = 1, ..., d, \text{ falls } l_i = 0 \end{array} \right\}$$

können wir nun $W_{\boldsymbol{l}}$ schreiben als:

$$W_{\boldsymbol{l}} = \text{span}\{\phi_{\boldsymbol{lj}}, j \in B_{\boldsymbol{l}}\}.$$

Somit ist die Familie von Basisfunktionen

$$\{\phi_{\boldsymbol{lj}}, j \in B_{\boldsymbol{l}}\}_{\boldsymbol{l}=0}^{\boldsymbol{N}}$$

gerade eine hierarchische Basis von V_N. Hierbei wird mit Hilfe eines Tensorproduktes der eindimensionale Fall auf den d-dimensionalen Fall verallgemeinert. Eine wichtige Eigenschaft der hierarchischen Basisdarstellung ist, dass die Träger der Basisfunktionen disjunkt sind. Für den linearen, eindimensionalen Fall ist in Abbildung 6.6 ein Beispiel gegeben. Zur Veranschaulichung des interessanteren 2-dimensionalen Falles sind in Abbildung 6.7 die Träger der einzelnen Basisfunktionen der Differenzenräume $W_{\boldsymbol{l}}$ angegeben. Stellen an denen die jeweiligen anisotropen Basisfunktionen den Wert 1 haben, sind als Punkte markiert. Nimmt man alle diese Punkte zusammen, so erhält man wieder ein äquidistantes, volles Gitter.

Mit dieser hierarchischen Basis lässt sich jede Funktion $f \in V_N$ schreiben als

$$f(x) = \sum_{|\boldsymbol{l}|_\infty \leq N} \sum_{\boldsymbol{l} \in B_{\boldsymbol{l}}} \alpha_{\boldsymbol{lj}} \phi_{\boldsymbol{lj}}(x)$$

mit Koeffizienten $\alpha_{\boldsymbol{lj}} \in \mathbb{R}$. Da sich der Raum V_N als direkte Summe der Differenzenräume schreiben lässt, kann jedes $f \in V_N$ auch als Summe der Teillösungen $f_{\boldsymbol{l}} \in W_{\boldsymbol{l}}$ geschrieben werden:

$$f = \sum_{|\boldsymbol{l}|_\infty \leq N} f_{\boldsymbol{l}}.$$

Bisher ist die Anzahl benötigter Basisfunktionen im Vergleich zur nodalen Basisdarstellung unverändert. Sowohl in nodaler, als auch in hierarchischer Darstellung sind $(2^N + 1)$ Basisfunktionen nötig, um eine Funktion f aus V_N darzustellen. Der Aufwand steigt exponentiell mit dem Diskretisierungslevel und der Dimension an.

6.3.2 Dünne Gitter und die Kombinationstechnik

Die Idee ist nun Basisfunktionen der hierarchischen Darstellung, welche nur geringen Einfluss auf die Gesamtlösung haben, zur Reduzierung des Aufwandes wegzulassen. Für den zweidimensionalen Fall zeigt Zenger in [Zen91] den Zusammenhang zwischen der Größe des Trägers einer hierarchischen Basisfunktion und dessen Anteil an der Interpolation einer Funktion.

Es kann gezeigt werden, dass unter gewissen Voraussetzungen an die Glattheit der zu interpolierenden Funktion, der Anteil einer hierarchischen Basisfunktion

an der Approximation durch die Größe des Trägers dieser Basisfunktion nach oben beschränkt ist. Für eine detailliertere Analyse sei auf [Bun92] verwiesen. Dort wird auch gezeigt, dass die zu approximierenden Funktionen aus dem so genannten *Sobolevraum mit dominierender gemischter Glattheit* H^2_{mix} stammen müssen. Die zugehörige Norm ist definiert als

$$\|f\|_{H^2_{mix}} = \sum_{0 \leq \boldsymbol{k} \leq \mathbf{2}} \left| \frac{\partial^{|\boldsymbol{k}|_1}}{\partial \boldsymbol{x}_{\boldsymbol{k}}} f \right|_2^2 .$$

Diese Räume weisen analog zum diskreten Raum V_N eine Tensorproduktstruktur auf. Das heißt sie lassen sich als Tensorprodukt der eindimensionalen Sobolevräume H^2 schreiben.

Als Schlussfolgerung aus diesen Feststellungen werden in [Zen91] die so genannten *dünnen Gitter* definiert, in denen die hierarchischen Basisfunktionen mit kleinem Träger und somit kleinem Beitrag zur Funktionsdarstellung nicht mehr zum diskreten Ansatzraum gehören. In hierarchischer Darstellung erreicht man dies, indem die diskrete Maximumsnorm $|\cdot|_\infty$ durch die diskrete L^1-Norm $|\cdot|_1$ ersetzt wird. Der Dünngitterraum wird damit definiert als

$$V_N^{dg} = \bigoplus_{|\boldsymbol{l}|_1 \leq N} W_{\boldsymbol{l}}, \tag{6.3.3}$$

was ein diskreter Teilraum von V_N ist.

In Abbildung 6.7 sind die wegfallenden Basisfunktionen grau eingezeichnet. Das Schema wird bildlich gesprochen an der Diagonalen abgeschnitten. Für den Aufwand bei der numerischen Behandlung bedeutet dies, dass statt $O(2^{Nd})$ Basisfunktionen nur noch $O(N^{d-1}2^N)$ verwendet werden. Der Aufwand wird damit substanziell reduziert. Für Interpolationsprobleme und Approximationsprobleme, die von elliptischen partiellen Differentialgleichungen zweiter Ordnung herrühren, wird beispielsweise in [GSZ92] gezeigt, dass die Dünngitterlösung f_N^{dg} einer Verfeinerungstiefe N nahezu die gleiche Genauigkeit liefert wie die Lösung f_N auf dem vollen Gitter. Es gilt

$$\|f - f_N^{dg}\|_2 = O(h_N^2 \log(h_N^{-1})^{(d-1)})$$

auf einem dünnen Gitter mit $h_N = 2^{-N}$ im Gegensatz zu

$$\|f - f_N\|_2 = O(h_N^2)$$

auf dem vollen Gitter. Dabei hat die Dünngitterlösung eine höhere Glattheitsvoraussetzung zu erfüllen. Es muss $f_N^{dg} \in H^2_{mix}$ im Gegensatz zu $f_N \in H^2$ erfüllt sein. Entsprechende Aussagen existieren ebenfalls für die Maximumsnorm.

Die so genannte *Kombinationstechnik* ist ein Verfahren, welches auf einer Sequenz von anisotropen, in jeder Koordinatenrichtung äquidistanten Gittern gemäß der *Kombinationsformel* die Lösung auf dem dünnen Gitter berechnet. Genauer betrachtet man die Sequenz von $\Omega_{\boldsymbol{l}}$ mit

$$\sum_{i=1}^{d} l_i = N - q, \quad , q = 0, ..., d-1, \quad l_i \geq 0.$$

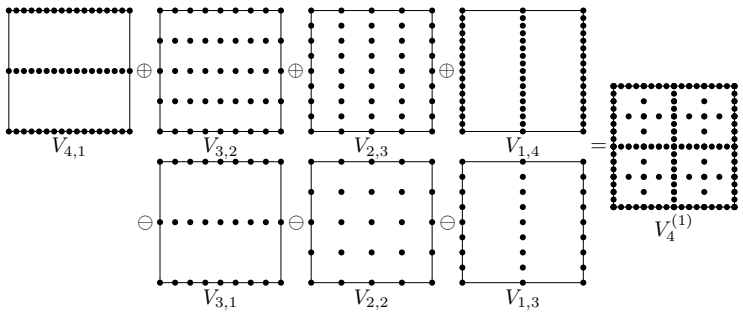

Abbildung 6.8: Die bei der Kombintationstechnik verwendeten Gitter und das resultierende dünne Gitter auf Level 4.

Die Teillösungen $f_{\boldsymbol{l}}$ auf diesen Teilräumen werden mit Hilfe der Kombinationsformel wie folgt zur Dünngitterkombinationslösung f_N^{KT} kombiniert:

$$f_N^{KT}(x) = \sum_{q=0}^{d-1} (-1)^q \binom{d-1}{q} \sum_{|\boldsymbol{l}|_1 = N-q} f_{\boldsymbol{l}}(x). \tag{6.3.4}$$

Die resultierende Funktion f_N^{KT} lebt im oben definierten Dünngitter-Raum V_N^{DG}. Für den Fall der Interpolation wird in [GSZ92] gezeigt, dass der so erhaltene Interpolant mit dem Interpolanten f_N^{DG} übereinstimmt. Allerdings ist dies nicht bei allen Anwendungen garantiert. Ebenfalls in [GSZ92] wird gezeigt, dass bei der numerischen Behandlung von Differentialgleichungen die mit der Kombinationstechnik erzielte Lösung zwar nicht mit der Galerkin-Lösung im Dünngitterraum übereinstimmt, der Fehler aber in der Regel von der gleichen Ordnung ist. Detailliertere Ausführungen zu dem Thema finden sich in [BGRZ94] und den dort angegebenen Referenzen.

In dieser Arbeit wird mit einer leicht abgewandelten Form der dünnen Gitter gearbeitet, bei denen die starke Randdominanz des Gitters zu einem gewissen Teil reduziert wird. Ähnlich wie bei partiellen Differentialgleichungen mit Dirichlet-Randwerten, wo sich auf dem Rand keine Freiheitsgrade befinden, macht es auch für die Dichteschätzung Sinn die Randdominanz zu reduzieren. Bei späteren Anwendungen werden die Daten so auf den Einheitswürfel transformiert, dass keine Datenpunkte in unmittelbarer Nähe des Randes liegen. Dadurch wird erreicht, dass die Dichtefunktion auf dem Rand gegen Null geht.

In Abbildung 6.8 sind für den zweidimensionalen Fall die verwendeten Gitter dargestellt. Die entsprechende Kombinationsformel lautet für diese Gitter in einer leicht abgewandelten Form:

$$f_N^{KT}(x) = \sum_{q=0}^{d-1} (-1)^q \binom{d-1}{q} \sum_{|\boldsymbol{l}|_1 = N+(d-1)-q} f_{\boldsymbol{l}}(x), \tag{6.3.5}$$

wobei folgende Sequenz von Gittern verwendet wird:

$$\sum_{i=1}^{d} l_i = N + (d-1) - q, \quad , q = 0, ..., d-1, \quad l_i > 0.$$

Ein Hauptargument für die Nutzung der Kombinationstechnik ist die relativ einfache Implementierung, da nach wie vor mit äquidistanten Gittern gearbeitet wird. Weiterhin sind die einzelnen Teilprobleme in ihrer Komplexität stark reduziert. Anstatt ein Problem der Ordnung $O(2^{Nd})$, sind $O(dN^{d-1})$ Teilprobleme der Größe $O(2^N)$ zu behandeln. Aufgrund der Unabhängigkeit der Teillösungen kann ein solches Verfahren in natürlicher Weise parallelisiert werden. Das heißt die Lösungen auf den Teilgittern können auf verschiedenen Prozessoren unabhängig voneinander gelöst werden. Sobald alle Prozesse beendet sind, werden die Teillösungen an einen Prozessor gesendet, der nach der Kombinationsformel die Gesamtlösung aufstellt. Dies bedeutet eine große Beschleunigung des Verfahrens.

7 Numerische Ergebnisse

In diesem Kapitel werden erste Ergebnisse mit dem Maximum a-posteriori Verfahren auf exponentiellen Familien mit Gauß-Prior auf dünnen Gittern vorgestellt (im Folgenden kurz *MAP-Verfahren* genannt). Dazu werden im ersten Abschnitt einige Details zur Implementierung angegeben, bevor das Verfahren zunächst für den univariaten Fall untersucht wird. Für einen simulierten Datensatz wird die Konvergenz der Lösung bei zunehmender Anzahl an Datenpunkten gegen die Originaldichte gezeigt. Weiterhin wird das Verhalten des Schätzers in Abhängigkeit von Level, Regularisierungsparameter und Größe der vorliegenden Datenmenge analysiert.

Zum Vergleich des MAP-Verfahrens mit anderen Schätzverfahren werden die in Kapitel 3.2.1 definierten verallgemeinerten Histogramme, das in Kapitel 4.3 vorgestellte ecdf-basierte Verfahren, sowie ein Standard-Kernschätzer mit Gaußkern verwendet.

Im dritten Teilkapitel kommen dünne Gitter für zweidimensionale Dichteschätzungen zum Einsatz. Dabei vergleichen wir die auf dem dünnen Gitter berechnete Lösung mit dem Vollgitteransatz und untersuchen das Konvergenzverhalten für verschieden große Datensätze. Anschließend werden für simulierte Datensätze die jeweiligen Dichteschätzer des MAP-Verfahrens mit denen eines kubischen Histogramms verglichen. Zum Abschluss wird das implementierte Verfahren auf einige reale Datensätze angewendet.

Sämtliche in diesem Kapitel abgebildete Grafiken wurden dabei mit MATLAB erzeugt [The06].

7.1 Implementierungsdetails

Nachdem die grundsätzliche Diskretisierung im vorherigen Kapitel vorgestellt wurde, werden hier Aspekte der Umsetzung angesprochen. Grundsätzlich wird mit der in Kapitel 6.2 vorgestellten Variante, bei der das nichtlineare Zielfunktional quadratisch approximiert wird, gearbeitet.

Zunächst ist ein geeigneter Regularisierungsoperator zu wählen, mit dem das Problem erst eindeutig lösbar wird und der die Gestalt der Lösung mit vorgibt. Falls nicht explizit anders erwähnt, wird für die Berechnung der Regularisierungsmatrix $\mathcal{A}_{i,j} = a(\varphi_i, \varphi_j)$ der Gradient als Regularisierungsoperator verwendet. Die Bilinearform $a(\cdot, \cdot)$ hat damit die Gestalt

$$a(\varphi_i, \varphi_j) = \lambda \cdot \int \nabla\varphi_i(x)\nabla\varphi_j(x)dx.$$

Die Gradienten der Basisfunktionen φ_j werden dabei vorab analytisch berechnet. Als Basisfunktionen verwenden wir für die folgenden Experimente ku-

bische B-Splines, die auf den Knoten des Gitters aufgesetzt werden. Als Regularisierungsmatrix entsteht eine dünn besetzte Matrix, die je nach Ordnung der B-Splines unterschiedlich viele Einträge ungleich Null enthält. Je höher die Ordnung der Basisfunktionen ist, desto dichter ist die Regularisierungsmatrix besetzt und desto aufwändiger ist folglich das Lösen des resultierenden Gleichungssystems.

Eine Schwierigkeit bei der Aufstellung der Gleichungen besteht in den auftretenden Integralen. Da sich

$$\langle \varphi, F(u) \rangle = \int \varphi \frac{\exp(u(y))}{\int \exp(u(z))dz} dy$$

analytisch nicht berechnen lässt, wird eine numerische Integration mittels der *Simpson-* oder *Trapezregel* durchgeführt [HB02]. Um die Genauigkeit der Integration zu verbessern, wird dazu auf einem zusätzlich verfeinerten Gitter die jeweilige Iterierte als Kollokationslösung aufgestellt. Die diskreten Lösungen werden somit sowohl als Koeffizientenvektor der Basisdarstellung, als auch als Vektor von Funktionswerten auf dem verfeinertem Gitter gespeichert.

Das entstehende Gleichungssystem ist mit dem Gradienten als Regularisierungsoperator symmetrisch und positiv definit und wird mit einem diagonal vorkonditionierten CG-Verfahren gelöst [HB02]. Als Startwert wird bei der ersten Iteration $u_0 = 0$ gewählt; bei allen weiteren Iterationen wird die vorherige Iterierte als Startwert verwendet. Dadurch spart man sich einige Iterations-Schritte im Vergleich zum Startwert Null.

Um in allen Fällen eine Konvergenz gegen die Lösung des MAP-Verfahrens zu garantieren, wird in den folgenden Experimenten stets mit der gedämpften Variante gearbeitet. Die hier vorgestellten Algorithmen wurden in C++ bzw. MATLAB implementiert.

7.2 Das Verfahren in 1D

7.2.1 Konvergenz der diskreten Lösung

Um die Konvergenz der diskreten Lösung gegen die Originaldichte zu untersuchen, erzeugen wir einen Datensatz von zwei überlagerten Normalverteilungen nach der bereits vorgestellten Inversionsmethode (siehe Kapitel 4.4.1).

Da die Originaldichte f explizit gegeben ist, werden die in Kapitel 2.5 vorgestellten Normen

$$MSE_{grid}\{f - \hat{f}\} = \frac{1}{N} \sum_{i=1}^{N} \Big(f(t_i) - \hat{f}(t_i) \Big)^2$$

auf den Gitterpunkten t_i und der analoge Datenfehler

$$MSE_{data}\{f - \hat{f}\} = \frac{1}{n} \sum_{i=1}^{n} \Big(f(x_i) - \hat{f}(x_i) \Big)^2$$

verwendet.

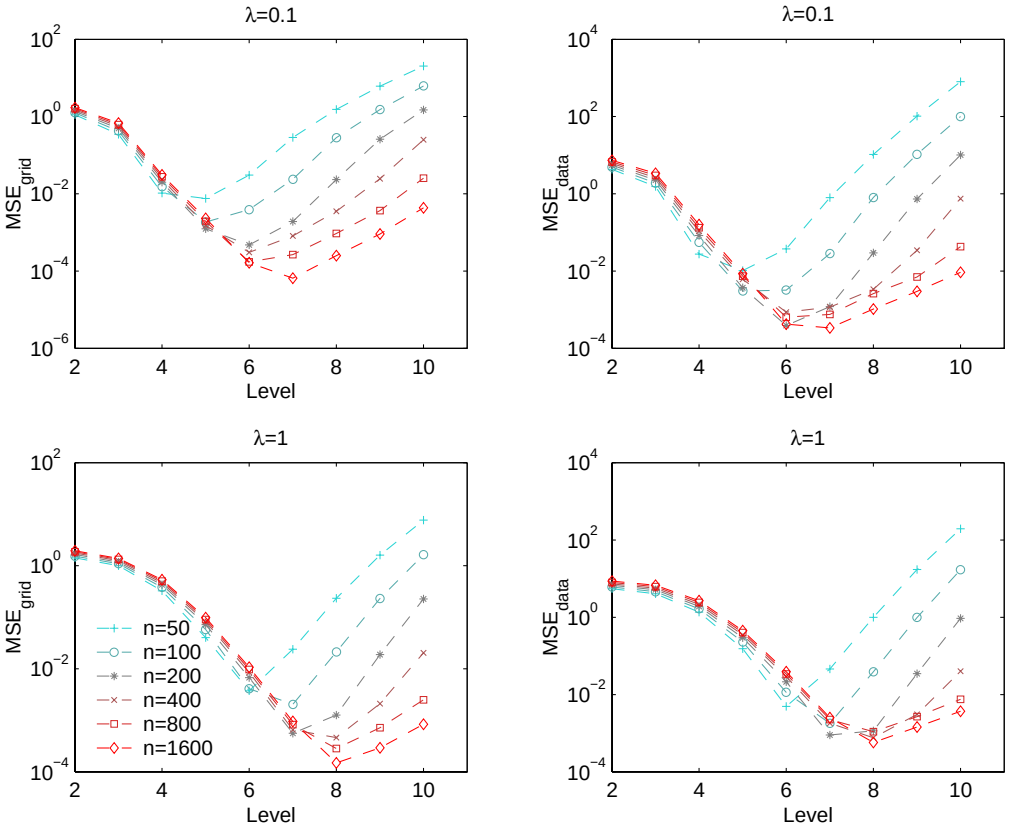

Abbildung 7.1: MSE auf Gitterpunkten (links) und Datenpunkten (rechts) für $\lambda = 0.1$ (oben) und $\lambda = 1$ (unten) bei verschieden großen Datensätzen.

	$\lambda = 1$, $n = 1600$		$\lambda = 1$, $n = 3200$	
Level l	$\sqrt{\frac{MSE^{l-1}_{grid}}{MSE^{l}_{grid}}}$	$\sqrt{\frac{MSE^{l-1}_{data}}{MSE^{l}_{data}}}$	$\sqrt{\frac{MSE^{l-1}_{grid}}{MSE^{l}_{grid}}}$	$\sqrt{\frac{MSE^{l-1}_{data}}{MSE^{l}_{data}}}$
3	1.1868	1.1317	1.1874	1.1280
4	1.6070	1.5813	1.5944	1.5679
5	2.3430	2.4453	2.3277	2.4247
6	3.0182	3.4054	2.9962	3.3676
7	3.3594	3.9221	3.3296	3.9100
8	2.5291	2.0943	3.1984	3.1621
9	0.7130	0.6300	0.9764	0.7161
10	0.5909	0.6258	0.5912	0.6099

Tabelle 7.1: Konvergenzverhalten bei 1600 (links) bzw. 3200 Datenpunkten (rechts) mit Regularisierungsparameter $\lambda = 1$.

Wir betrachten zunächst das Verhalten des Schätzers in Abhängigkeit von der Datenmenge n. Dazu ist in Abbildung 7.1 für zwei verschiedene Regularisierungsparameter λ bei größer werdenden Datensätzen der Fehler auf dem für die Integration verwendeten Gitter sowie die analoge Fehlernorm auf den Datenpunkten in Abhängigkeit vom Diskretisierungslevel dargestellt.

In Tabelle 7.1 ist die relative Abnahme des Fehlers bei 512 Datenpunkten und $\lambda = 1$ (linke Hälfte) und 1024 Datenpunkten und $\lambda = 10$ (rechte Hälfte) dargestellt. Als Fehlernorm wird hier die diskrete L^2-Norm verwendet, die sich als Wurzel des MSE ergibt.

Wie bereits in der Grafik zu sehen ist, hängt das optimale Diskretisierungslevel von der Datenmenge und dem Regularisierungsparameter ab. Bei beiden Experimenten ist zu sehen, dass der minimale Fehler bei größer werdender Datenmenge kleiner wird. Bei geringem n steigt der Fehler bei steigendem Level recht schnell wieder an, falls der Glättungsparameter nicht entsprechend erhöht wird. Man gelangt dann in den Bereich der Überanpassung. Hier tritt analog zu Kernschätzern oder Histogrammen der Effekt auf, dass die Träger der Basisfunktionen bei zunehmendem Level immer schmaler werden, so dass nur wenige oder gar keine Datenpunkte darin liegen. Dadurch oszilliert der resultierende Schätzer relativ stark. Wie jedoch in Abbidung 7.1 unten rechts zu sehen ist, kann dieser Effekt durch eine stärkere Glättung merklich abgeschwächt werden.

Insgesamt sieht man den zu erwartenden Zusammenhang, dass bei steigender Datenmenge der Fehler immer geringer wird, sofern das Diskretisierungslevel und der Regularisierungsparameter erhöht werden. Falls n groß und die Informationsdichte demnach hoch genug ist, konvergiert der Schätzer bei steigendem Level offensichtlich gegen die Lösung. Von einer festen Konvergenzordnung (siehe Tabelle 7.1) kann hier jedoch nicht gsprochen werden. Vielmehr müssen sowohl das Level als auch der Glättungsparameter λ an die Datenmenge angepasst werden. Ab einem gewissen Level wächst der Fehler sogar wieder, da der Bereich der Überanpassung erreicht wird.

Um das Wechselspiel von Diskretisierungslevel und Regularisierung noch einmal besser zu verdeutlichen, ist in Abbildung 7.2 für die gleiche Originaldichte mit 100 simulierten Daten auf verschiedenen Diskretisierungsleveln λ variiert und die Summe aus Gitter- und Datenfehler gemessen worden. Für dieses Beispiel ist offensichtlich Level 6 bei $\lambda \approx 0.3$ optimal. Auf höherem Diskretisierungslevel steigt der Fehler wieder leicht an. Außerdem fällt auf, dass der Regularisierungsparameter λ offensichtlich exponentiell mit dem Level ansteigend zu wählen ist, um optimale Ergebnisse zu erreichen.

Weiterhin sieht man, dass am rechten Rand der Grafik - also für großes λ - sich der Fehler auf allen Levels einem Grenzwert nähert. Hier konvergiert die approximierte Dichte bei zu starker Glättung gegen die Dichtefunktion einer Gleichverteilung auf dem Intervall $[0, 1]$.

Als nächstes Experiment wird die optimale Wahl des Regularisierungsparameters bei unterschiedlicher Datenmenge n untersucht. Abbildung 7.3 stellt in der linken Grafik für verschiedene Level (5-9) den optimalen Wert von λ in Abhängigkeit der Datenmenge dar. Ausgehend von einem Datensatz mit $n = 50$

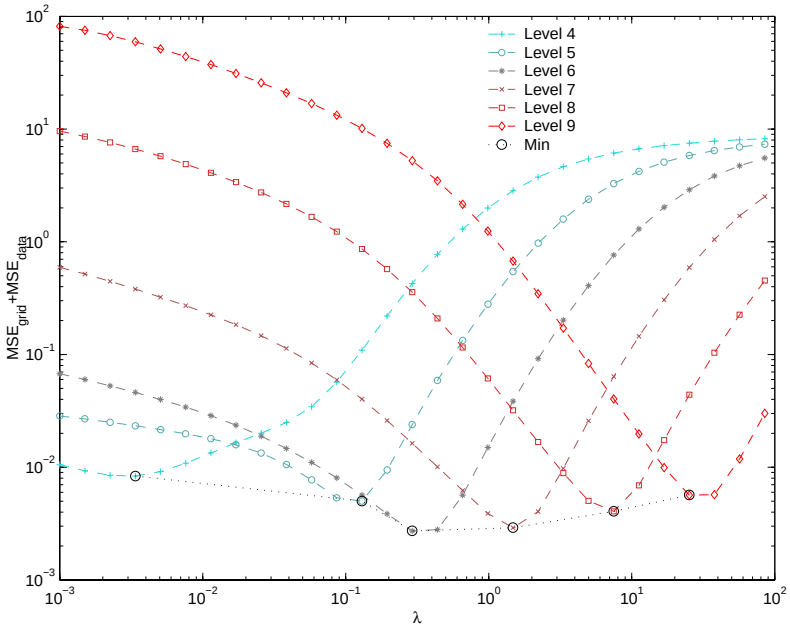

Abbildung 7.2: MSE auf Gitter und Daten für verschiedene Diskretisierungslevel (4-9) in Abhängigkeit vom Regularisierungsparameter λ.

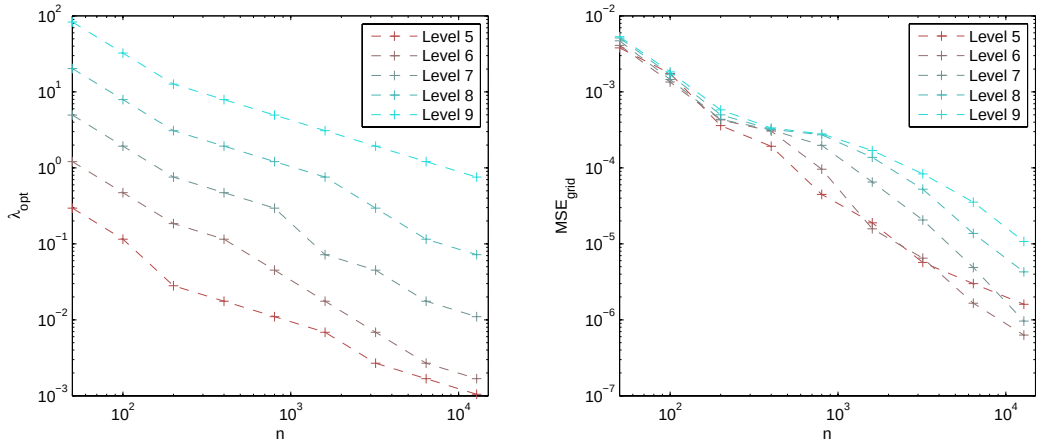

Abbildung 7.3: Links: Die optimale Größe des Regularisierungsparameters λ in Abhängigkeit der Datenmenge n für die Level 5 bis 9. Rechts: Der zugehörige mittlere quadratische Fehler.

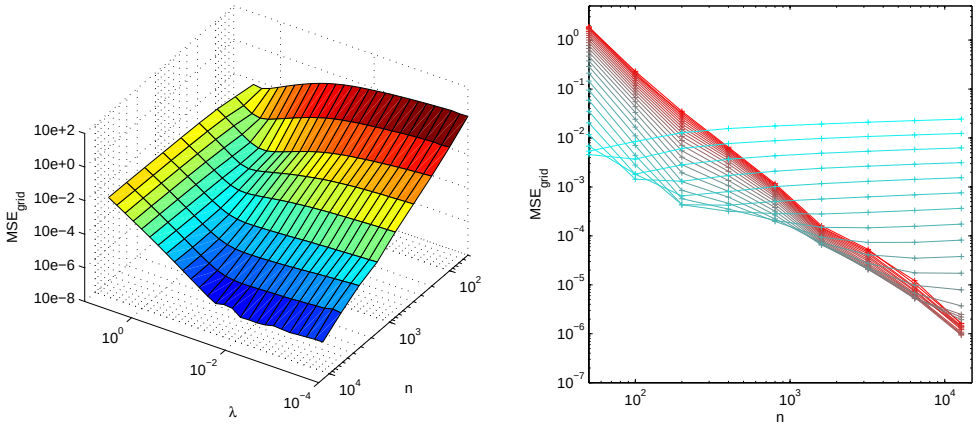

Abbildung 7.4: Der Fehler auf dem Gitter in Abhängigkeit von Glättungsparameter λ und Anzahl an Datenpunkten n (links) sowie die Entwicklung des Fehlers bei größer werdender Datenmenge für verschiedene feste λ (rechts) jeweils auf einem Diskretisierungslevel von 7.

Punkten ist diese sukzessiv bis zu einer Größe von $n = 12\,800$ verdoppelt worden. Die drei entstehenden Kurven verlaufen in doppelt logarithmischer Darstellung parallel versetzt. Der Glättungsparameter ist demnach bei größerem Level exponentiell höher zu wählen. Weiterhin wird ein umgekehrt proportionaler Zusammenhang zwischen Datenmenge n und optimalem Parameter λ deutlich: Bei Verdopplung der Datenmenge ist der Glättungsparameter zu halbieren.

Auf der rechten Seite von 7.3 ist der jeweilige mittlere quadratische Fehler auf dem Gitter in Abhängigkeit von der Anzahl an Eingangsdaten dargestellt. Dabei ist jeweils ein optimales λ gewählt worden. Es tritt der zu erwartende, deutliche Abfall des Fehlers bei jeweiliger Verdopplung der Datenmenge auf. Als Konvergenzrate wurde im Mittel circa 1 gemessen.

Abbildung 7.4 stellt für eine Diskretisierung der Tiefe 7 die Auswirkungen von λ und n auf den Fehler dar. In der linken Grafik wird ein „Tal“ erkennbar, welches sich diagonal durch die Kurve zieht. Noch besser zu erkennen ist dies in der rechten Grafik, welche die einzelnen Schnitte der zweidimensionalen Kurve zeigt. Hier wird jeweils für einen festen Regularisierungsparameter der mittlere quadratische Fehler in Abhängigkeit von der Größe des Datensatzes als Kurve dargestellt. Enthält der vorliegende Datensatz wenig Datenpunkte, ist eine hohe Regularisierung - in der rechten Abbildung türkis gezeichnet - optimal, während bei sehr großen Werten von n ein kleines λ - rot gezeichnet - den Fehler minimiert.

Sofern also auf einem festen Diskretisierungslevel bei größer werdendem Datensatz der Regularisierungsparamter angepasst wird, ist eine Konvergenz gegen die Originaldichte festzustellen. Für einen festen Datensatz ist nicht garantiert, dass die Approximation bei Erhöhung der Dikretisierungsfeinheit einen geringeren Fehler aufweist. Die optimale Approximation hängt vom Diskretisierungslevel und dem dafür zu wählenden Glättungsparameter ab.

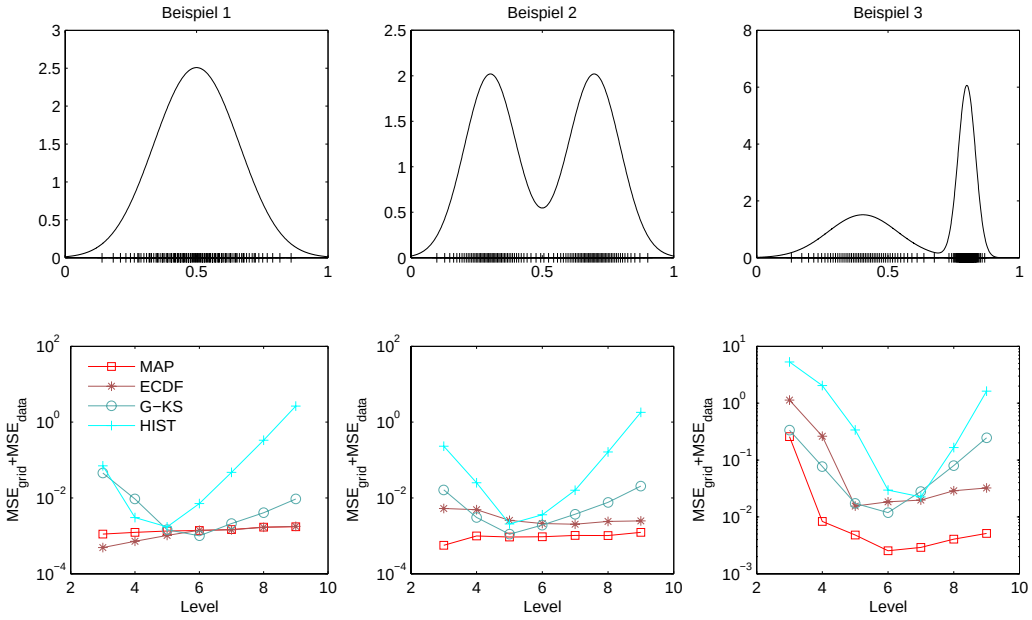

Abbildung 7.5: Drei erste Beispiele zum Vergleich der Verfahren. In der oberen Reihe ist jeweils die Originaldichte mit den verwendeten Datenpunkten abgebildet. Die untere Reihe stellt den Fehler für die verglichenen Verfahren in Abhängigkeit vom Diskretisierungslevel dar.

7.2.2 Verfahrensvergleich

Nachdem beschrieben wurde, wie sich das MAP-Verfahren bei diversen Änderungen der Eingabeparameter verhält, werden in diesem Abschnitt die erzielten Ergebnisse mit denen anderer Verfahren verglichen. Als Originaldichten werden zunächst jeweils aus zwei Gaußglocken bestehende Dichtefunktionen gewählt. Mit Hilfe der bereits beschriebenen Inversionsmethode wurden daraus 100 Datenpunkte erzeugt, die gemäß der jeweiligen Verteilung gleichmäßig verteilt sind. Dadurch ist es möglich eine Fehlermessung ohne stochastische Einflüsse durchzuführen.

Verglichen wird das MAP-Verfahren mit dem in Kapitel 4.3 vorgestellten, linearen ecdf-basierten Verfahren, einem verallgemeinerten Histogramm mit kubischen B-Splines, sowie einem Standard-Kernschätzer mit Gaußkern.

In Abbildung 7.5 sind drei erste Experimente abgebildet. Die obere Reihe stellt die jeweiligen Datenpunkte mit zugehöriger Originaldichte dar, bei der zwei Normalverteilungen in verschiedenem Abstand überlagert worden sind. In der unteren Reihe ist für alle verglichenen Methoden die Summe aus mittlerem quadratischen Fehler auf den Datenpunkten und analogem Fehler auf den Gitterpunkten in Abhängigkeit vom Diskretisierungslevel abgebildet. Für das ecdf-basierte-Verfahren und den MAP-Ansatz wurde jeweils ein optimaler Regularisierungsparameter λ gewählt.

An dieser Stelle mag die Darstellung eines datenbasierten Kernschätzers be-

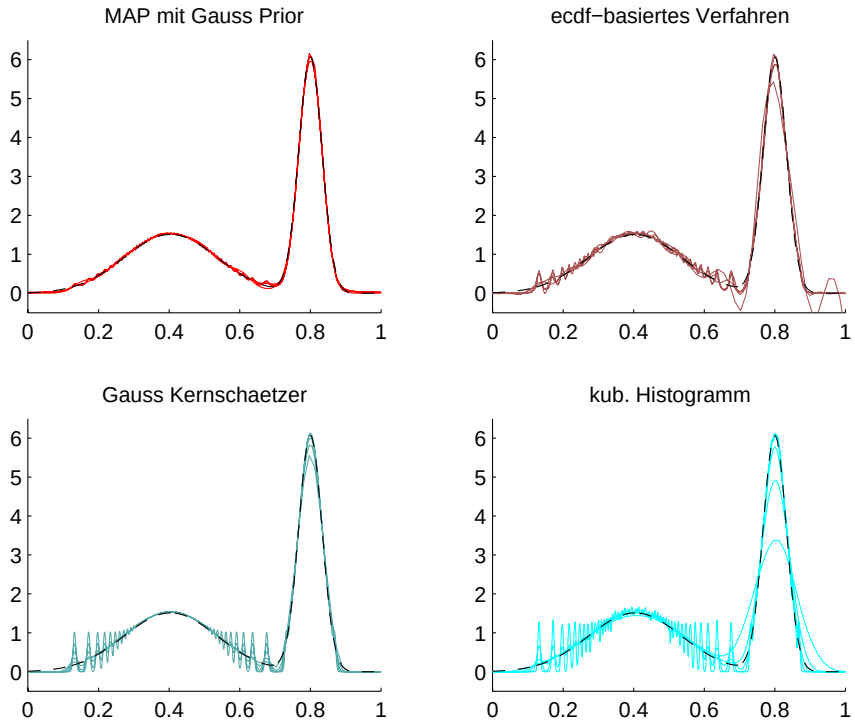

Abbildung 7.6: Die betrachteten Schätzer auf verschiedenem Level (4-8) in jeweils einer Grafik. Die Originaldichte ist als gestrichelte Linie eingezeichnet.

züglich eines Diskretisierungslevels irritieren, da der steuernde Parameter die Bandbreite der Kernfunktionen ist und lediglich für eine diskrete Darstellung der Lösung ein Gitter verwendet wird. Um hier eine Vergleichbarkeit zu den anderen Verfahren zu schaffen, wurde für jedes Beispiel ausgehend von einem Startwert mit steigendem Level jeweils die Bandbreite halbiert. Ein direkter Vergleich des Kernschätzers mit den anderen, gitterbasierten Verfahren macht auf einem festen Level demnach keinen Sinn. Trotzdem kann der Grafik entnommen werden wie die bestmöglichen Dichtefunktions-Approximationen aussehen, da die Ausgangsbandbreiten für den Kernschätzer so gewählt wurden, dass das Minimum der Fehlerkurve im Bereich der betrachteten Level liegt.

In den beiden ersten, sehr einfachen Beispielen liefert unser MAP-Verfahren Resultate, die mindestens ebenso gut sind, wie die der Vergleichsverfahren. Während Kernschätzer und Histogramm jeweils ein deutliches Minimum aufweisen, von dem der Fehler in beide Richtungen stark ansteigt, lassen sich das ecdf-baiserte Verfahren und das MAP-Verfahren durch eine entsprechende Glättung besser steuern und liefern stabilere Ergebnisse. Insbesondere für das sehr einfache erste Beispiel unterscheiden sich die Fehler dieser Ansätze nur gering. Bereits beim zweiten Beispiel ist jedoch das nichtlineare Verfahren dem linearen ecdf-basierten Verfahren überlegen und weist auf allen Diskretisierungsleveln einen geringeren Fehler auf.

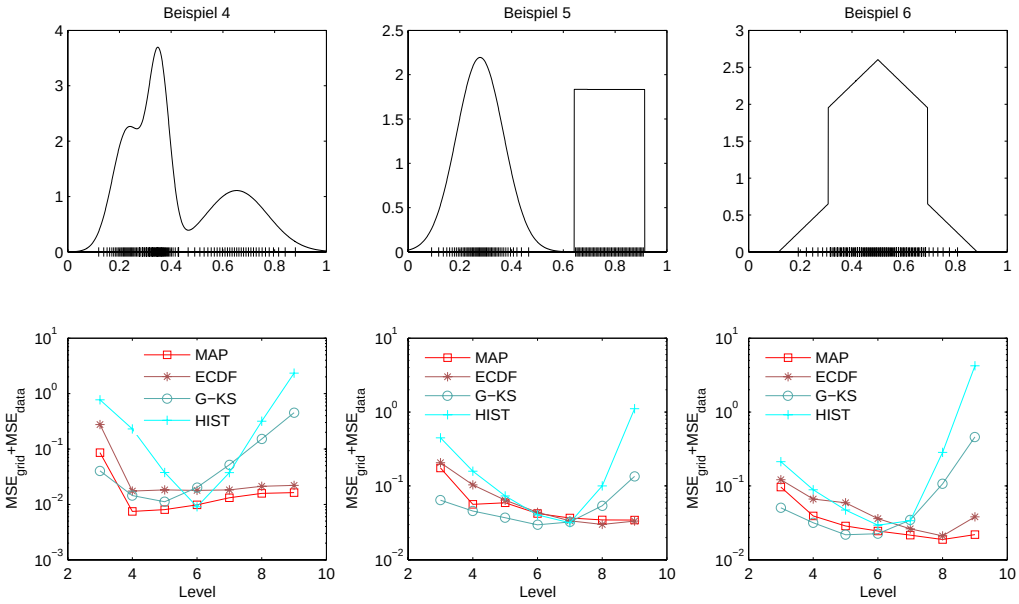

Abbildung 7.7: Drei weitere Beispiele zum Vergleich der Verfahren.

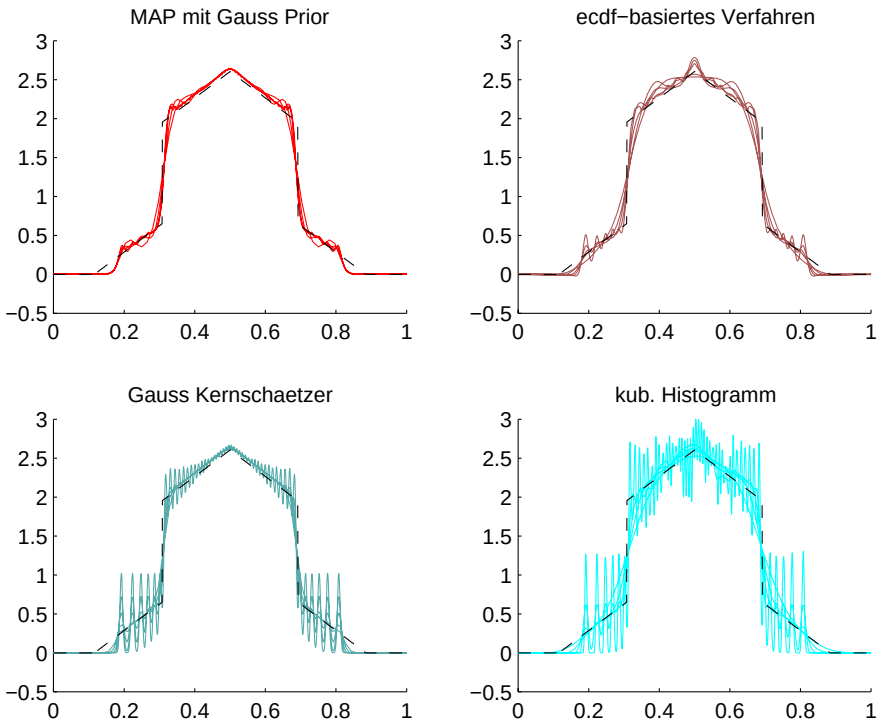

Abbildung 7.8: Die betrachteten Schätzer auf verschiedenem Level (4-8) in jeweils einer Grafik. Die Originaldichte ist als gestrichelte Linie eingezeichnet.

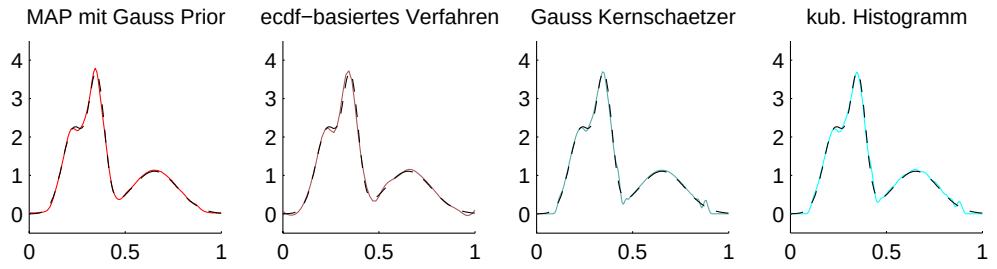

Abbildung 7.9: Die betrachteten Schätzer mit jeweils bester Parameterkonstellation. Die Originaldichte ist als gestrichelte Linie eingezeichnet.

Die mit dem dritten Datensatz erzeugten Dichteschätzer der MAP-Methode unterscheiden sich qualitativ sehr deutlich von denen der Vergleichsverfahren. Dieser Datensatz enthält einige wesentliche Schwierigkeiten, die bei der Dichteschätzung auftreten können. Eine Anforderung an ein Schätzverfahren sollte hier eine saubere Trennung der überlagerten Gaußglocken sein. Weiterhin sollte ein Schätzer in der Breite nicht zu stark oszillieren und trotzdem die volle Spitze der schmaleren Normalverteilung wiedergeben.

Wie der Fehlerkurve entnommen werden kann, bereitet der Datensatz dem MAP-Verfahren deutlich weniger Schwierigkeiten als den Vergleichsverfahren. Der Fehler ist hier ab Level 4 substanziell geringer. Es ist nicht verwunderlich, dass auf Level 3 auch das MAP-Verfahren einen großen Fehler aufweist, da bei diesem Level die Träger der Basisfunktionen entsprechend breit sind. Insofern kann hier keines der Verfahren die schmale Gaußkurve gut approximieren.

Um die Stabilität des Verfahrens zu verdeutlichen, sind in Abbildung 7.6 für jeden Schätzer die Lösungen auf Level 4 bis 8 jeweils in eine Grafik gezeichnet. Offensichtlich lässt sich das MAP-Verfahren sehr gut steuern und neigt auch bei höherem Level nicht zum Oszillieren.

Als viertes Beispiel haben wir nun eine weitere Gaußkurve hinzugenommen und bei allen drei Dichtefunktionen verschiedene Varianzen gewählt. In Abbildung 7.9 ist für jedes Verfahren die jeweils beste Approximation zusammen mit der Originaldichte abgebildet. Wie aus Abbildung 7.7 deutlich wird, weist bei diesem Beispiel auf Level 6 das kubische Histogramm den geringsten Fehler der verglichenen Verfahren auf, wobei der Fehler bei Erhöhung oder Verringerung des Levels stark ansteigt. Im Gegensatz dazu verläuft die Fehlerkurve des MAP-Verfahrens wesentlich glatter.

Die Beispiele 5 und 6 stellen Dichtefunktionen dar, wie sie auch in anderen Arbeiten häufiger als Testfunktionen gewählt werden (z.B. [Hoo99]). Hierbei fällt auf, dass das optimale Diskretisierungslevel des MAP-Verfahrens für diese Beispiele deutlich höher liegt. Offensichtlich wird ein gewisses Level benötigt, um vor allem den Sprung der Gleichverteilung gut genug rekonstruieren zu können. Trotzdem ist auch bei niedrigerem Level der Fehler vergleichbar mit dem der anderen Verfahren.

Abbildung 7.8 enthält wiederum die einzelnen Schätzer auf verschiedenen Levels in einer Grafik. Auch diese in realen Daten wohl nicht zu erwartende Dich-

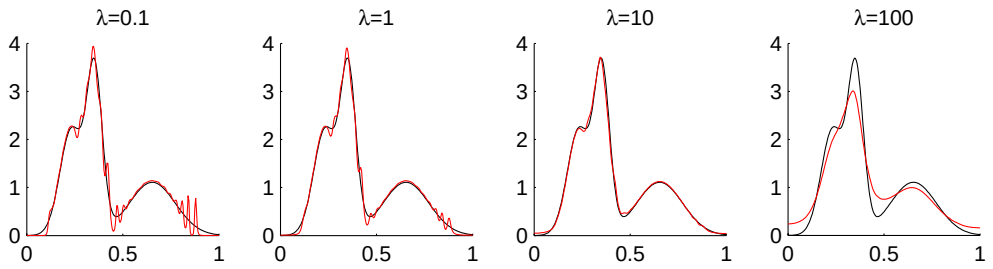

Abbildung 7.10: Approximation mit MAP-Verfahren auf Level 6 für verschiedene Wahl von λ.

tefunktion wird vom MAP-Verfahren gut rekonstruiert.

7.2.3 a-priori Wahl des Regularisierungsparameters

Bei allen bisherigen Experimenten wurde das optimale λ für das ecdf-basierte Verfahren und den MAP-Schätzer a-posteriori bestimmt. Welchen Einfluss der Regularisierungsparameter auf die Lösung hat, ist in Abbildung 7.10 für das 4. Beispiel verdeutlicht. Wie bereits erwähnt, ist λ mit steigendem Diskretisierungslevel exponentiell wachsend zu wählen. Im Folgenden wird eine Möglichkeit vorgestellt aus den *a-posteriori* bestimmten Parametern dieser Beispiele eine Regel für eine *a-priori* Wahl des Regularisierungsparameters zu erstellen.

In Abbildung 7.11 ist für beide Verfahren und die gerade beschriebenen Beispiele der optimale Glättungsparameter λ_{opt} über dem Diskretisierungslevel l in logarithmischer Skala aufgetragen. Nach Ansicht dieser Grafik liegt es nahe hier einen Zusammenhang der Form $\lambda_{opt} = \exp(\alpha + \beta \cdot l)$ zu erwarten. Um die Parameter α und β zu bestimmen, mitteln wir λ_{opt} für die 6 Beispiele über jedem Level und führen eine lineare Regression der logarithmierten Werte durch. Die resultierenden Regressionskurven sind in Abbildung 7.11 als gepunktete Linien eingezeichnet. Ihre genauen Gleichungen lauten für eine Datenmenge der Größe $n = 100$:

$$\lambda_{opt}(l) = \exp(-10.58 + 1.71 \cdot l) \qquad \text{(MAP-Verfahren)}$$
$$\lambda_{opt}(l) = \exp(-8.25 + 1.38 \cdot l) \qquad \text{(ecdf-basiertes Verfahren)}.$$

Geht man zusätzlich von einem umgekehrt proportionalen Verhältnis von n und λ_{opt} aus, wie es anhand diverser Experimente zu vermuten ist, so ergeben sich die Gleichungen:

$$\lambda_{opt}(l, n) = \frac{1}{n} \cdot \exp(-5.98 + 1.71 \cdot l) \qquad \text{(MAP-Verfahren)} \qquad (7.2.1)$$
$$\lambda_{opt}(l, n) = \frac{1}{n} \cdot \exp(-3.65 + 1.38 \cdot l) \qquad \text{(ecdf-basiertes Verfahren)}. \qquad (7.2.2)$$

Dies könnte also eine Möglichkeit sein, den Parameter λ *a-priori* anhand der Datenmenge für ein bestimmtes Diskretisierungslevel zu wählen.

In Abbildung 7.11 ist weiterhin der Zusammenhang zwischen Glattheit der Originaldichte und Größe des optimalen Regularisierungsparameters zu sehen.

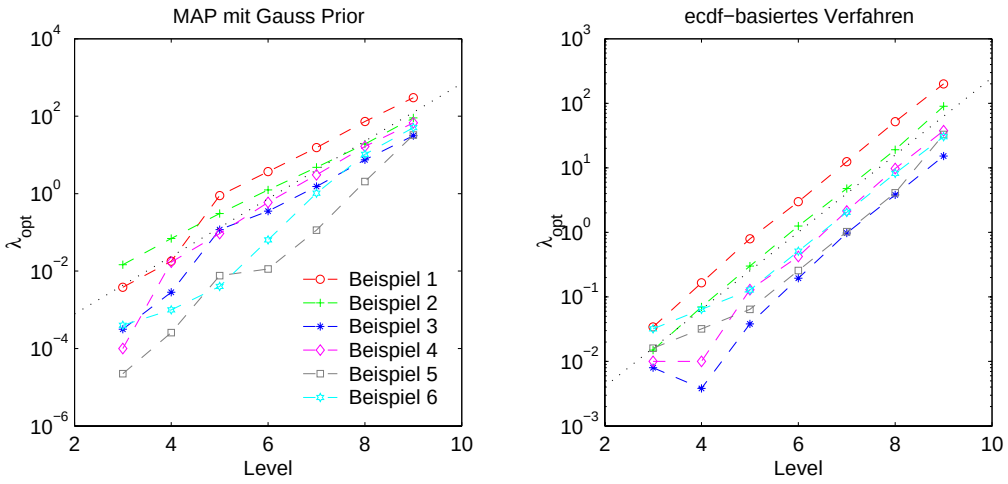

Abbildung 7.11: Optimales λ der 6 Beispiele in Abhängigkeit vom Diskretisierungslevel.

So verlaufen die Kurven von Beispiel 5 und 6 deutlich unter denen der anderen Beispiele. Ein optimaler Regularisierungsparameter hängt somit zusätzlich von der Glattheit der Lösung ab.

Alternativ gibt es die Möglichkeit mit Hilfe so genannter *Kreuzvalidierungsverfahren* oder *Bootstrap-Methoden* ohne Kenntnis der Lösung den Glättungsparameter vorab zu bestimmen (siehe [Sco92]). Auf eine detailliertere Analyse verzichten wir jedoch im Rahmen dieser Arbeit.

Wie wir bereits in den univariaten Beispielen gesehen haben, liefert das entwickelte MAP-Verfahren sehr gute Ergebnisse, die für spezielle Dichtefunktionen sogar eine ganze Ordnung besser als die der verglichenen Verfahren sind. Problematisch bleibt dabei die Wahl der optimalen Eingangsparameter. So lassen sich statt der bisher verwendeten kubischen B-Splines andere Basisfunktionen verwenden. Diese haben wieder Einfluss auf die Lage des optimalen Regularisierungsparameters, da sie je nach Gestalt und Träger die Regularität der Lösung beeinflussen. Des weiteren könnte man statt dem hier verwendeten Gradienten andere Regularisierungs-Operatoren verwenden. Setzt man beispielsweise höhere Ableitungen an, ist ein geringeres λ nötig um eine vergleichbare Glättung zu erreichen.

Insgesamt sind die eindimensionalen Experimente viel versprechend. Insbesondere für Dichtefunktionen die ähnlich wie das dritte Beispiel mehrere, unterschiedlich breite Verteilungen aufweisen, ist das MAP-Verfahren den Vergleichsverfahren deutlich überlegen. Im nächsten Teilkapitel wird das Verfahren in zwei Dimensionen untersucht.

7.3 Zweidimensionale Experimente

In den bisher betrachteten Beispielen ist jeweils mit einem vollen Gitter gearbeitet worden, welches im univariaten Fall mit dem dünnen Gitter übereinstimmt.

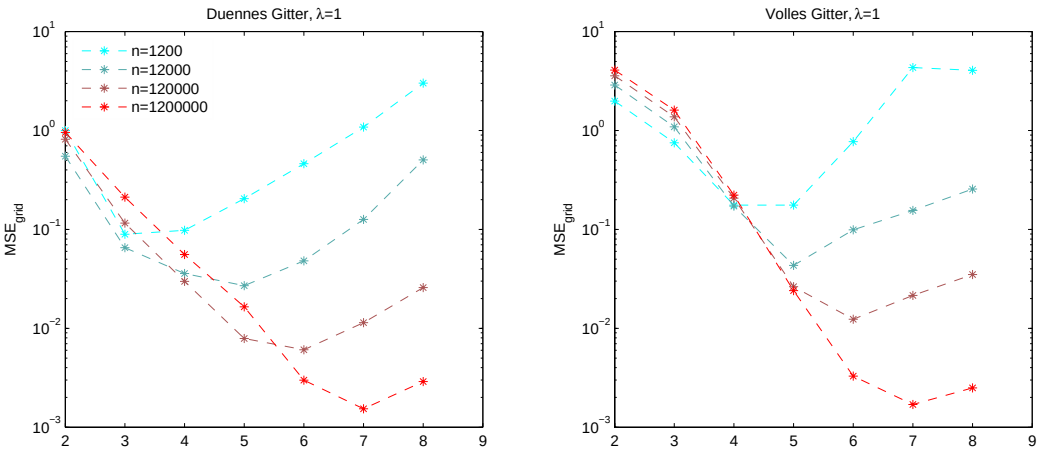

Abbildung 7.12: Konvergenzverhalten der Dünngitterlösung (links) im Vergleich zur Vollgitterlösung (rechts) für verschieden große Datensätze.

Erst für zweidimensionale Datensätze kann zwischen den Diskretisierungstechniken unterschieden werden. Der folgende Absatz untersucht dazu zunächst das Verhalten der diskreten Lösung auf dem dünnen Gitter im Vergleich zum vollen Gitter. Anschließend wird anhand von simulierten Beispieldatensätzen die Güte der mit einem dünnen Gitter erzielten Approximation untersucht. Als Dämpfungsparameter wird in den folgenden Beispielen $\omega = 0.5$ gewählt und die Iteration bei einem Residuum von 10^{-6} bzw. nach maximal 30 Iterationen beendet.

7.3.1 Konvergenz der diskreten Lösung

Zur Konvergenzanalyse werden Dichtefunktionen vorgegeben, aus denen Datensätze simuliert werden. Konvergenz heißt wiederum, dass für großes n die Approximation gegen die Originaldichte konvergieren sollte. Wir simulieren dazu nach einer vorgegebenen Originaldichte Datensätze mit Hilfe der in MATLAB integrierten Funktion *normrnd*. Die gewählten Originaldichten sind dementsprechend Summen von mehreren überlagerten Normalverteilungsdichten.

Für die Konvergenzanalyse verwenden wir nach der Dichtefunktion aus Abbildung 7.14 verteilte Datensätze verschiedener Größe. In Abbildung 7.12 ist der Fehler auf dem Diskretisierungsgitter in Abhängigkeit vom Level für Datensätze bestehend aus 1 200, 12 000, 120 000 und 1.2 Millionen Punkten aufgetragen. Der linke Teil stellt das Konvergenzverhalten auf dem dünnen Gitter dar, der rechte entsprechend auf dem vollen Gitter. Analog zum univariaten Beispiel reduziert sich der Fehler für beide Dikretisierungsvarianten je nach Datenmenge bis zu einem bestimmten Level und steigt dann wieder an. Je größer die Anzahl an Punkten ist, desto höher ist das entsprechende optimale Level, bevor der Bereich der Überanpassung beginnt. Tabelle 7.2 stellt für beide Diskretisierungstechniken die relative Abnahme des Fehlers im Vergleich zum nächstgröberen Level dar.

Level l	$\sqrt{\frac{MSE_{dg}^{l-1}}{MSE_{dg}^{l}}}$	$\sqrt{\frac{MSE_{vg}^{l-1}}{MSE_{vg}^{l}}}$
3	2.1250	1.5945
4	1.9497	2.6888
5	1.8369	3.0426
6	2.3499	2.7043
7	1.3959	1.3907
8	0.7283	0.8246

Tabelle 7.2: Konvergenzverhalten bei 1,2 Millionen Datenpunkten und $\lambda = 0.1$ auf dem dünnen Gitter (links) im Vergleich zum vollen Gitter (rechts).

Ebenfalls analog zum eindimensionalen Fall muss hier mit dem Level auch der Regularisierungsparameter angepasst werden, um die Glattheit der Lösung, die durch Vorgabe der Ansatzfunktionen mit beeinflusst wird, zu steuern.

Abbildung 7.13 zeigt für wachsendes n das Konvergenzvehalten der Dünngitterdiskretisierung verglichen mit der Lösung auf dem vollen Gitter bei jeweils optimaler Wahl des Regularisierungsparameters. In der linken Grafik ist für unterschiedliche Diskretisierungslevel der beiden Diskretisierungen der jeweils optimale Regularisierungsparameter in Abhängigkeit von der Anzahl an Datenpunkten n dargestellt. In der rechten Grafik ist die Entwicklung des zugehörigen mittleren quadratischen Fehlers abgebildet.

Sowohl bei der Vollgitterlösung als auch auf dem dünnen Gitter ist analog zum univariaten Fall ein umgekehrt proportionales Verhältnis zwischen Größe der Datenmenge und optimalem Regularisierungsparameter festzustellen. Bei vier mal so großer Datenmenge ist λ jeweils zu Halbieren.

In der rechten Abbildung ist ein ähnlicher Abfall des Fehlers für die verschiedenen Diskretisierungstechniken zu beobachten. Jedoch flacht der Fehler der Dünngitterlösung, wie aufgrund der $O(h^2 \log(h)^{(d-1)})$-Konvergenz der dünnen Gitter zu erwarten ist, bereits bei kleineren Datenmengen ab. Auf dem vollen Gitter bleibt die Konvergenzrate stabiler (siehe Tabelle 7.3). Ein analoger Effekt tritt hier nur auf Level 4 ein. Sind die Datenmengen also sehr groß, ist auch das Diskretisierungslevel entsprechend größer zu wählen, um eine optimale Approximation zu erzielen.

Weiterhin wird ersichtlich, dass auf dem dünnen Gitter wesentlich weniger stark geglättet werden muss. Dies kann durch die Struktur der anisotropen Teilgitter begründet werden, die bereits eine Glättung implizieren.

Abbildung 7.15 verdeutlicht den unterschiedlichen Einfluss des Regularisierungsparameters für beide Diskretisierungstechniken. Für dieses Beispiel ist der Fehler minimal für eine Vollgitterdiskretisierung vom Level 4. Für Level 5 und 6 ist hingegen der Fehler auf dem dünnen Gitter kleiner als auf dem vollen Gitter. Hier kommt die durch die Teilgitter impilzierte Glättung zum Tragen. Die optimalen Regularisierungsparameter auf dem dünnen Gitter sind daher wesentlich kleiner zu wählen als auf einem vollen Gitter mit analoger minimaler Maschenweite.

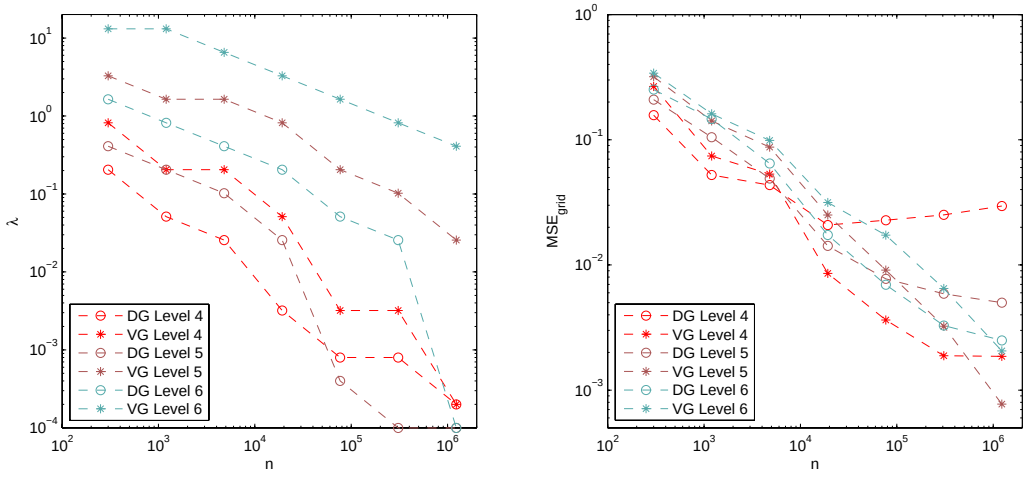

Abbildung 7.13: Das Konvergenzverhalten der Lösung auf dem dünnen Gitter verglichen mit der Lösung auf dem vollen Gitter. Abgebildet ist das jeweils optimale λ in Abhängigkeit von der Anzahl an Datenpunkten n (links), sowie der zugehörige Fehler (rechts) für die Level 4-6.

	Level 5		Level 6	
n	$\sqrt{\frac{MSE_{dg}^{n_{i-1}}}{MSE_{dg}^{n_i}}}$	$\sqrt{\frac{MSE_{vg}^{n_{i-1}}}{MSE_{vg}^{n_i}}}$	$\sqrt{\frac{MSE_{dg}^{n_{i-1}}}{MSE_{dg}^{n_i}}}$	$\sqrt{\frac{MSE_{vg}^{n_{i-1}}}{MSE_{vg}^{n_i}}}$
1200	1.4133	1.5017	1.3094	1.4540
4800	1.4617	1.2722	1.5051	1.2762
19200	1.8586	1.8673	1.9301	1.7699
76800	1.3537	1.6593	1.5818	1.3511
307200	1.1472	1.6757	1.4554	1.6318
1228800	1.0851	2.0466	1.1449	1.7773

Tabelle 7.3: Relative Abnahme des L^2-Fehlers bei größer werdenen Datensätzen auf dem dünnen Gitter (dg) im Vergleich zum vollen Gitter (vg)

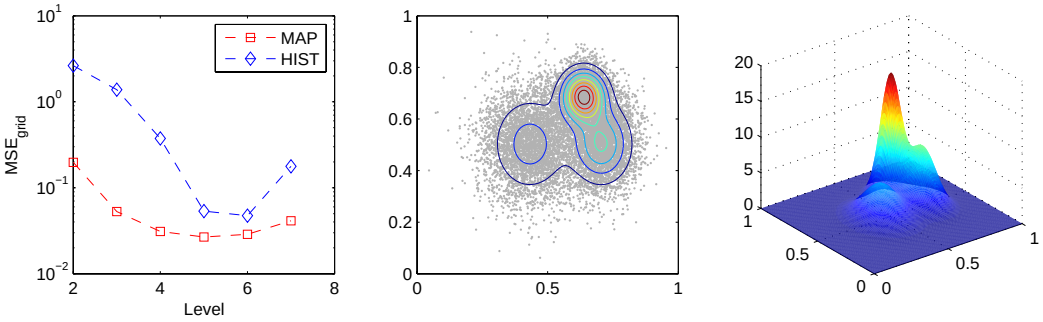

Abbildung 7.14: 1. Beispiel einer Dichteschätzung. In der Mitte sind die 12 000 Datenpunkte mit Höhenlinien der rechts gezeigten Originaldichte abgebildet. Links sind für das MAP-Verfahren und das kubische Histogramm die mittleren quadratischen Fehler auf jeweiligem Diskretisierungslevel abgebildet.

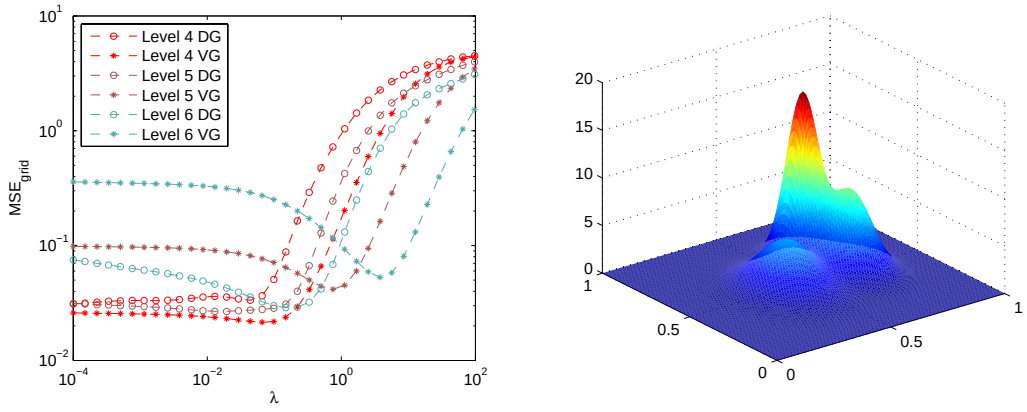

Abbildung 7.15: Experiment mit 12 000 Datenpunkten. Links ist der Fehler für drei verschiedene Level in Abhängigkeit vom Regularisierungsparameter λ für die Lösung auf dem dünnen Gitter im Vergleich zum vollen Gitter dargestellt.

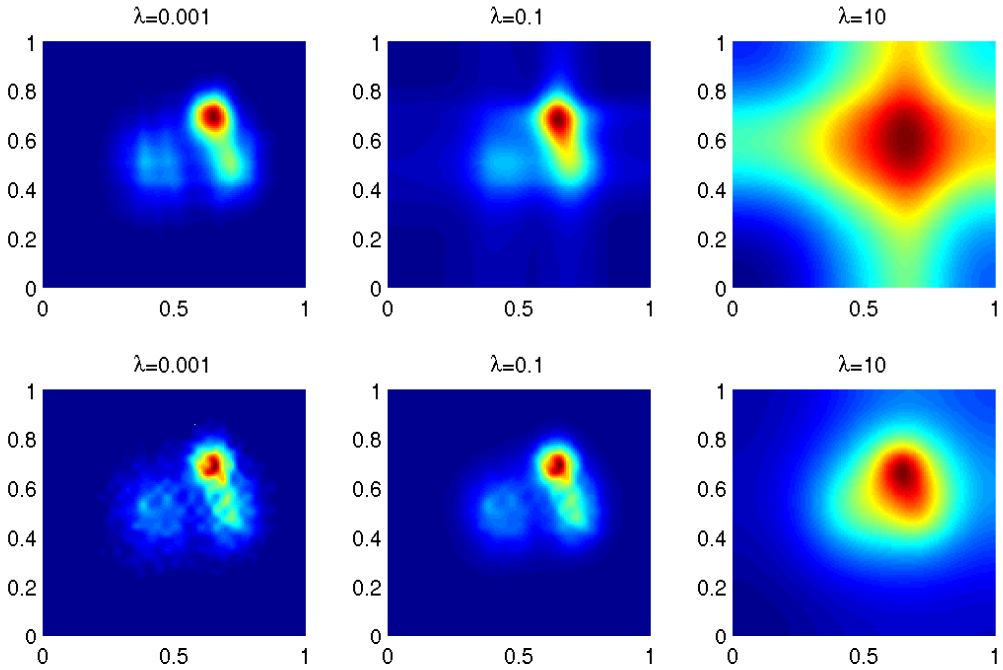

Abbildung 7.16: Experiment mit 12000 Datenpunkten auf Level 5 für verschiedene Glättungsparameter. Die obere Reihe enthält die Dünngitterlösungen und entsprechend die untere die Lösung auf einem vollen Gitter.

Um die unterschiedliche Gestalt der jeweiligen Lösungen zu veranschaulichen, stellt Abbildung 7.16 verschiedene Lösungen auf einem Diskretisierungslevel von 5 dar. Die zugrundeliegende Originaldichte ist die gleiche wie bisher mit einem aus 12 000 Punkten bestehenden Datensatz. Die obere Reihe stellt die Dünngitterlösung für drei verschieden große Werte von λ dar. In der unteren Reihe ist analog die Lösung auf einem vollen Gitter dargestellt. Bei übermäßig starker Glättung ($\lambda = 10$) ist die Struktur der anisotropen Teilgitter zu erkennen, die zu einem „Verschmieren“ in den Koordinatenrichtungen führt.

Daraus folgernd entstand die Idee den Glättungsparameter nicht für alle anisotropen Teilgitter der Kombinationslösung gleich zu wählen, sondern separat auf jedem enthaltenen Gitter ein eigenes $\lambda_{(l_1,l_2)}$ zu bestimmen. Dazu wurden die jeweiligen Lösungen auf den anisotropen Teilgittern für verschiedene λ aufgestellt und der Fehler gegen die Originaldichte gemessen. Um eine Vergleichbarkeit zu schaffen, wurde hier die Teilgitterlösung auf ein volles äquidistantes Gitter interpoliert. Als ein Beispiel ist in Abbildung 7.17 der Fehler der Teilgitter des dünnen Gitters auf Level 6 gegen den Regularisierungsparameter abgetragen. Hierbei ist der Fehler auf fast allen Teilgittern für ein minimales λ am geringsten. Trotzdem ist diese Konstellation in der Summe nicht optimal (siehe Abbildung 7.15).

Offensichtlich ergibt sich die optimale Glättung erst durch Kombination der Teillösungen, so dass dieses Vorgehen keinen Erfolg bringt. Um hier eine weitere

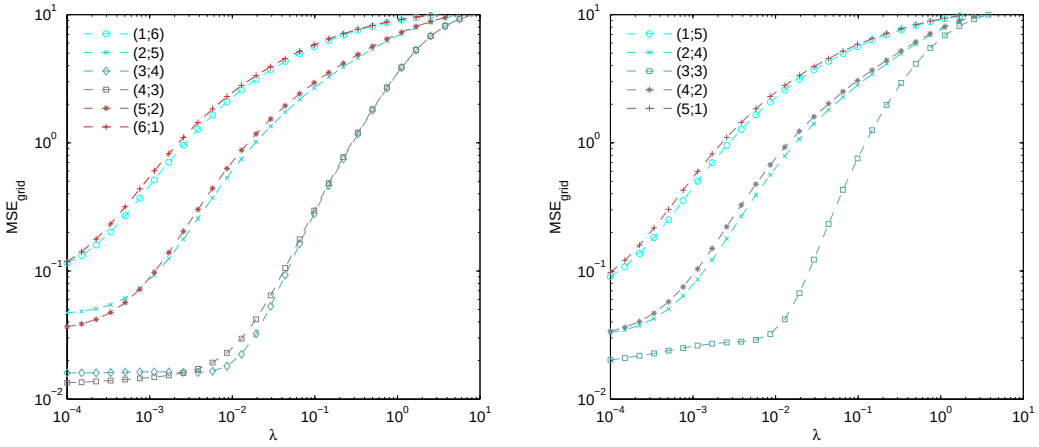

Abbildung 7.17: Optimale Wahl von λ auf den anisotropen Teilgittern des dünnen Gitters vom Level 6. Links die mit $+1$ gewichteten Gitter; rechts analog mit -1 gewichtete Teilgitter.

Verbesserung zu erzielen, müssten die Koeffizienten der Kombinationstechnik, die alle 1 oder -1 sind (siehe Gleichung (6.3.5)), in einer separaten Maximierungsroutine optimiert werden. In [HGC05] und [Gar06] stellen die Autoren einen derartigen Ansatz für Regressionsaufgaben unter dem Namen *Opticom* vor. Dort wird er bereits für Regressions- und Klassifikationsaufgaben eingesetzt.

7.3.2 Zweidimensionale simulierte Datensätze

In Abbildung 7.18 sind für drei verschiedene Beispiele 12 000 Datenpunkte nach einer vorgegebenen Dichte simuliert und der Fehler auf verschiedenem Diskretisierungslevel für ein optimales λ berechnet worden. Abgebildet ist rechts jeweils die Originaldichte und links der mittlere quadratische Fehler auf den Gitterpunkten des MAP-Ansatzes, verglichen mit analoger Fehlernorm für ein kubisches Histogramm. In der Mitte sind die verwendeten Datenpunkte mit den Höhenlinien der Originaldichte abgebildet.

Das kubische Histogramm eignet sich hier als Vergleichsverfahren, da die zugrunde liegenden Ansatzfunktionen identisch zu denen des MAP-Verfahrens sind. Es zeigt sich analog zu den univariaten Beispielen, dass das Verfahren sehr stabile Ergebnisse liefert, die in vielen Fällen eine ganze Ordnung besser sind als die des Vergleichsverfahrens. Weiterhin ist auch hier das optimale Diskretisierungslevel abhängig von der Gestalt der Originaldichte. Während im obersten Beispiel von Grafik 7.18 der Fehler bereits auf Level 2 minimal ist, wird im mittleren Beispiel das Minimum erst bei Level 6 erreicht.

Abbildung 7.19 zeigt für verschiedene Verfeinerungsstufen die unterschiedliche Gestalt der Dichteschätzer. Das MAP-Verfahren lässt sich offenbar auch mit einer Dünngitterdiskretisierung gut steuern.

Um auch bei unbekannter Originaldichte eine Größenordnung des Regularisie-

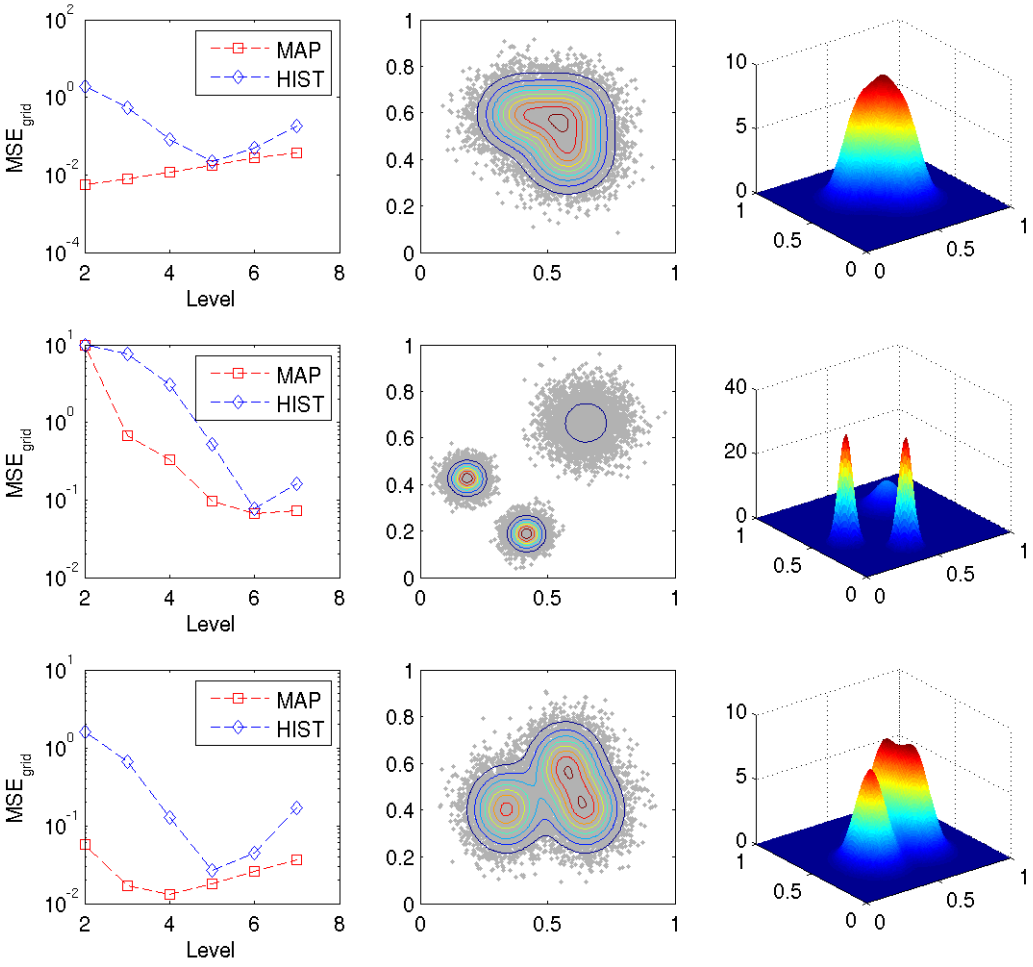

Abbildung 7.18: Verschiedene Beispiele mit jeweils 12 000 Datenpunkten. Abgebildet ist der mittlere quadratische Fehler auf dem Diskretisierungsgitter für die verglichenen Verfahren in Abhängigkeit vom Level (links), die Eingabedaten mit Höhenlinien der Originaldichte (Mitte) sowie die zu rekonstruierende Dichtefunktion (rechts).

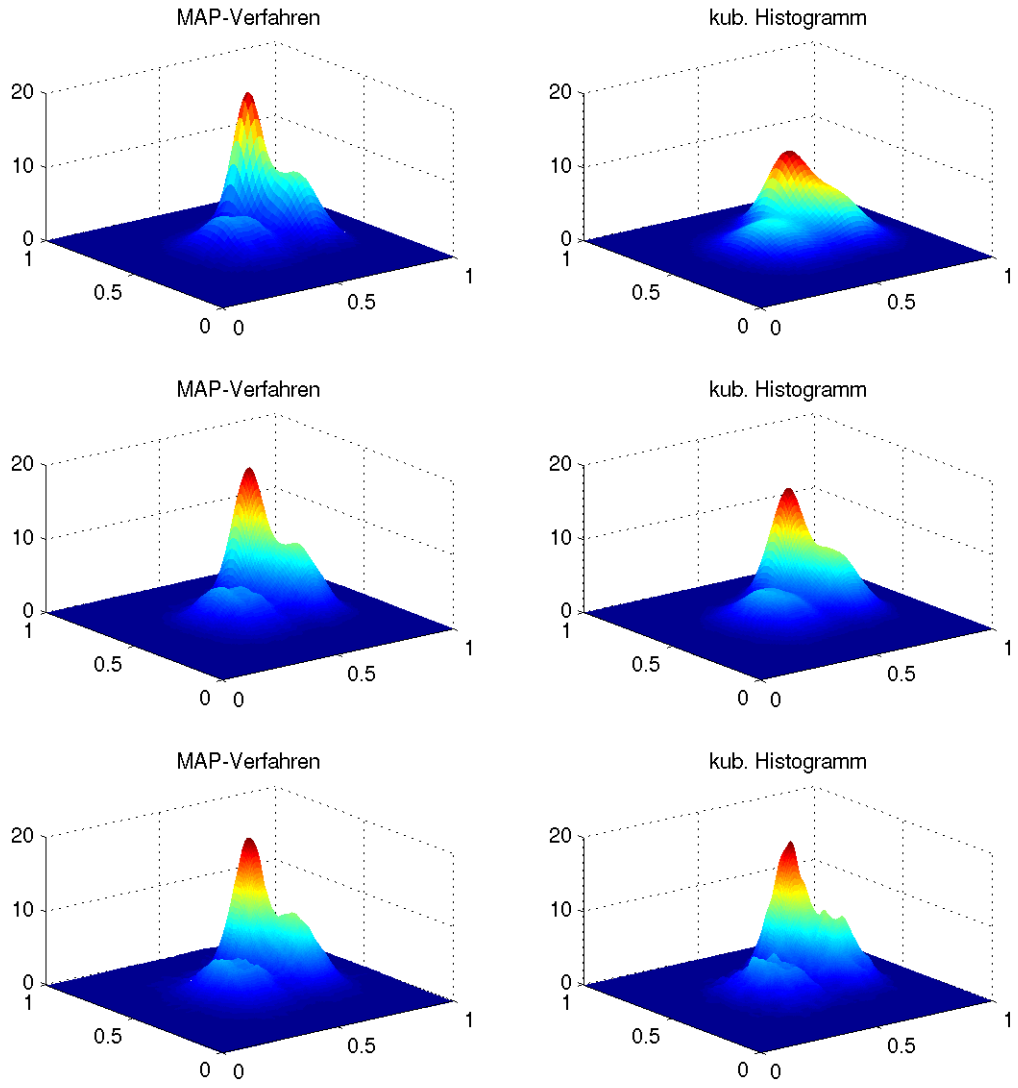

Abbildung 7.19: Vergleich der Lösungen des MAP-Verfahrens mit denen des kubischen Histogramms auf Level 4 (oben) bis Level 6 (unten) für einen aus 12 000 Punkten bestehenden Datensatz

rungsparamters vorgeben zu können, führen wir analog zu den eindimensionalen Beispielen eine Regressionanalyse der a-posteriori bestimmten Regularisierungsparameter in Abhängigkeit vom Level durch. Als Regel für die optimale Wahl von λ erhalten wir damit unter Annahme eines umgekehrt proportionalen Zusammenhangs zwischen Datenmenge und Regularisierungsparameter:

$$\lambda_{opt}(l,n) = \frac{1}{\sqrt{n}} \exp(-5.75 + 1.41 \cdot l).$$

Im nächsten Abschnitt wird das implementierte Verfahren auf einige reale Datensätze angewendet.

7.4 Anwendungsbeispiele auf realen Daten

Die bisher verwendeten Datensätze sind durch geeignete Simulationen erzeugt worden. Solch regelmäßige Strukturen sind in realen Daten jedoch nur sehr selten zu beobachten. In den folgenden Beispielen ist der Glättungsparameter jeweils nach obiger Regel gewählt worden. Ein passendes Diskretisierungslevel wurde a-posteriori nach Ansicht der Dichten gewählt.

Old Faithful Geyser

Der bereits in Kapitel 4 verwendete Datensatz des „Old Faithful Geysers“ ist wohl der am meisten verwendete Datensatz in bestehender Literatur zur Dichteschätzung [Sil86],[Sco92]. Die 107 Werte stellen jeweils die Zeit zwischen zwei Ausbrüchen des Geysirs dar.

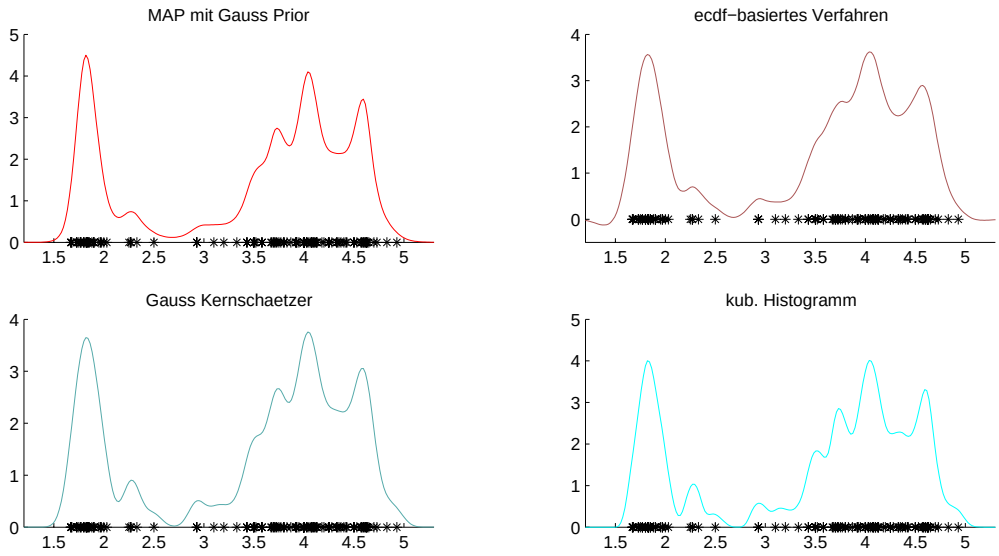

Abbildung 7.20: Dichteschätzung des Old Faithful Geyser-Datensatzes.

In Abbildung 7.20 ist für diesen Datensatz die approximierte Dichtefunktion der verschiedenen Verfahren auf Level 6 angegeben. Der Regularisierungspara-

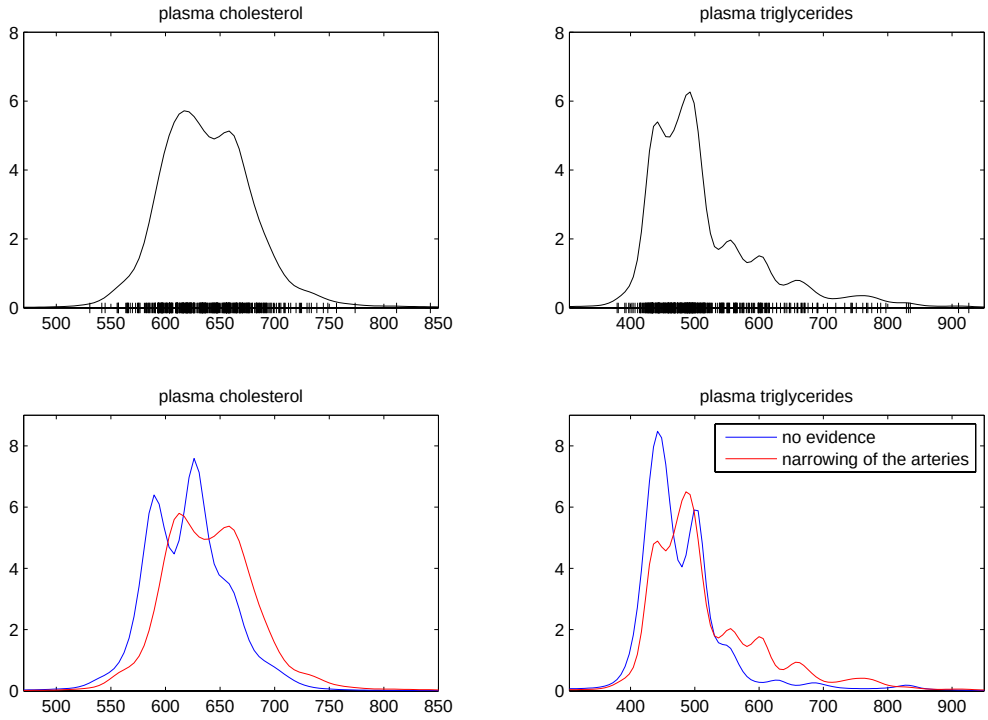

Abbildung 7.21: 1D Dichteschätzungen der Blut-Fett-Konzentrations-Daten auf Level 5. In der oberen Reihe für den kompletten Datensatz; unten getrennt nach den Patientengruppen.

meter λ ist hierbei nach den Regeln aus (7.2.1) und (7.2.2) gewählt worden. Für dieses Beispiel weisen die jeweiligen Schätzverfahren ähnliche Ergebnisse auf.

Blut-Fett-Konzentrations Daten

Der hier untersuchte Datensatz stammt aus dem Buch [Sco92] von SCOTT. Die Daten bestehen aus Cholesterin- und Triglycerid-Werten des Blutplasmas von 371 männlichen Patienten. Davon wiesen 51 Patienten keinerlei Anzeichen eines Herzinfarktes auf. Bei den restlichen 320 Patienten wurden Verengungen der Arterien festgestellt. Um hier eventuell einen Zusammenhang aufzeigen zu können, führen wir eine Dichteschätzung sowohl der gesamten Datenmenge als auch für beide Patientengruppen separat durch. In Abbildung 7.21 ist in der oberen Reihe für beide Messwerte jeweils eine Dichteschätzung auf Level 5 durchgeführt worden. In der unteren Reihe sind die beiden Gruppen separiert untersucht worden. Die Dichtekurve der Patienten mit Anzeichen auf eine Verengung der Arterien ist leicht nach rechts versetzt. Eine mögliche Schlussfolgerung wäre, dass ein Zusammenhang zwischen erhöhten Chlosterin-Werten und Verengung der Arterien besteht.

Abbildung 7.22 stellt links eine zweidimensionale Dichteschätzung des kompletten Datensatzes auf Level 5 dar. In der rechten Grafik sind die einzelnen

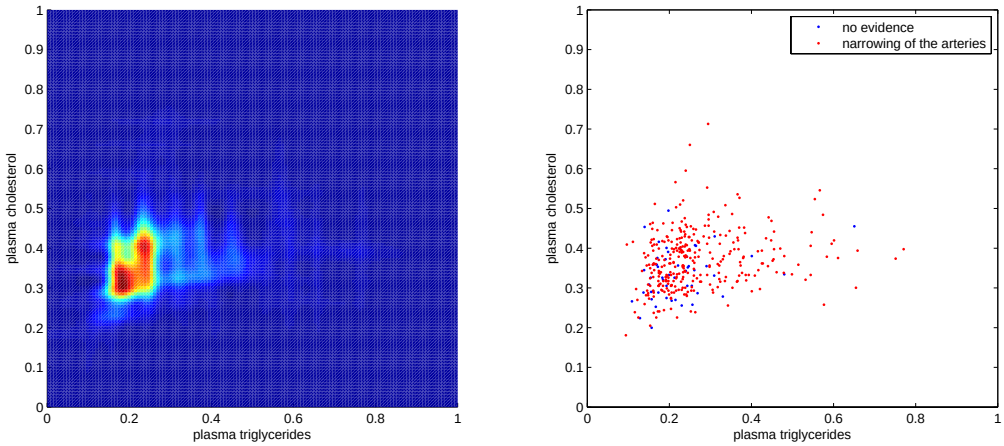

Abbildung 7.22: 2D Dichteschätzung der Blut-Fett-Konzentrations-Daten auf Level 5 mit $\lambda = 0.01$

Messwerte nach Patientengruppe getrennt markiert. Die Dichtefunktion lässt hier zwei Zentren vermuten, die Interpretation soll jedoch einem Fachkundigen überlassen werden.

Iris Daten

Der Iris-Datensatz stammt aus dem Machine Learning Repository [Uni06] der University of California. Er besteht aus Messungen zu 150 Pflanzen, die zur Gattung der Iris (Lilien) gehören. Jeweils 50 Pflanzen gehören einer speziellen Klasse an. Unterschieden wird in „Iris Setosa“, „Iris Versicolor“ sowie „Iris Virginica“. Gemessen wurde für jede Pflanze die Breite des Kelchblattes (*sepal width*), die Länge des Kelchblattes (*sepal length*) und die Breite und Länge des Blütenblattes (*petal length*, bzw. *petal width*). Das Ziel ist eine Zuordnung der Pflanzen zu der jeweiligen Klasse anhand der vier Messungen.

Zunächst untersuchen wir den Datensatz in jeder Dimension einzeln. Bereits hier wird deutlich, dass zumindest eine Klasse leicht linear separiert werden kann (Abbildung 7.23 untere Reihe). So scheint sich bei einer Sorte die Form des Blütenblattes deutlich von den anderen zu unterscheiden, während die Kelchform recht gleichmäßig verteilt ist. Die Dichtefunktion unten rechts legt jedoch nahe, dass die beiden anderen Klassen ebenfalls getrennt werden können. Um dies genauer zu untersuchen, führen wir eine zweidimensionale Dichteschätzung durch. Abbildung 7.24 zeigt eine Dichteschätzung für zwei Merkmalskombinationen der betrachteten Pflanzen. In der oberen Reihe sind die Länge und Breite des Kelchblattes verwendet worden. Unten sind die Breite der Blüte und die Breite des Kelches als Eingangsdaten genommen worden. Ganz rechts sind die Datenpunkte mit verschiedenen Farben für die einzelnen Klassen dargestellt. In der oberen Grafik sieht man nochmal eine saubere Trennung einer der drei Pflanzensorten. Mit Hilfe der unten abgebildeten Dichtefunktion ist zusätzlich eine Trennung der restlichen beiden Klassen möglich. Hier ist durch Kombination zweier Merkmale eine Trennung der Klassen möglich, die anhand der eindimen-

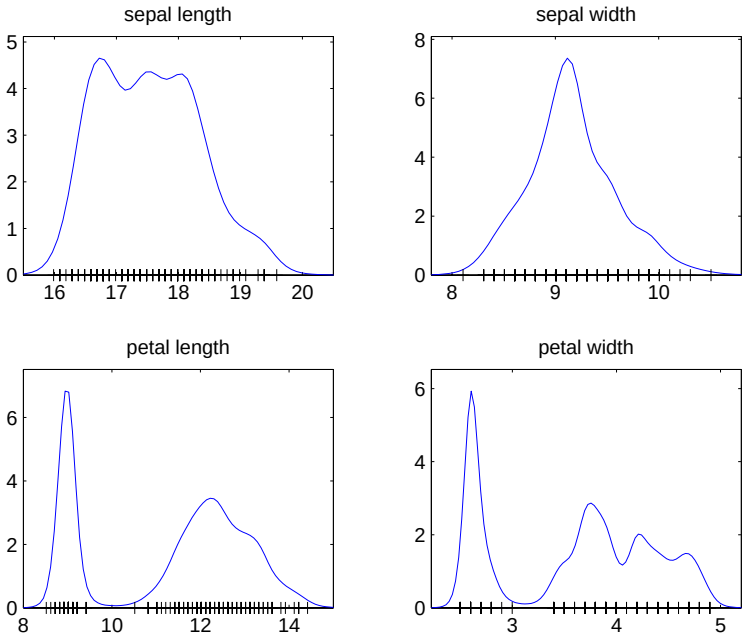

Abbildung 7.23: 1D Dichteschätzungen auf Level 5 der Komponenten des Iris-Datensatzes.

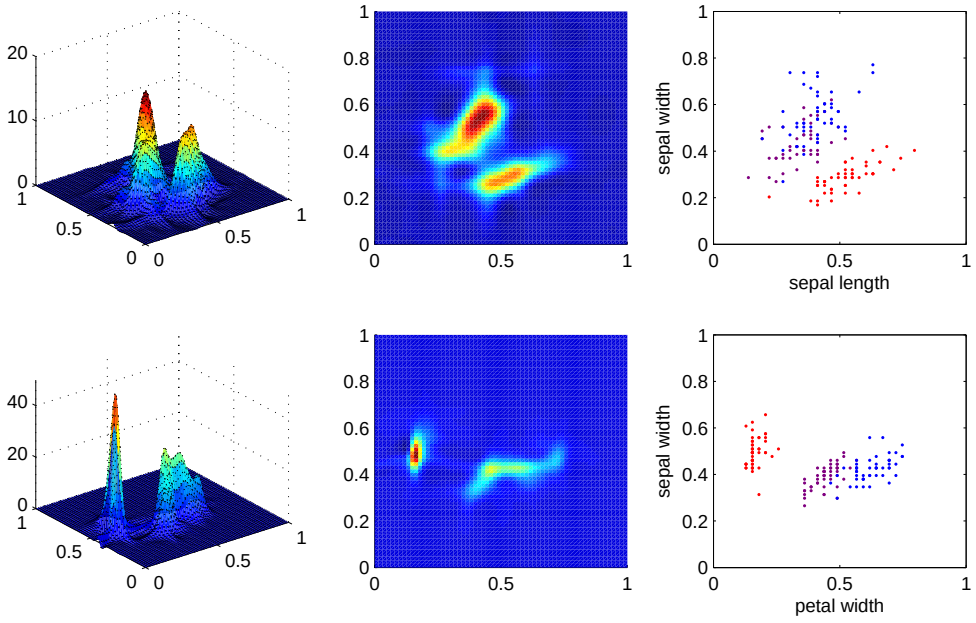

Abbildung 7.24: 2D Dichteschätzung des Iris-Datensatzes auf Level 4. Rechts sind die verschiedenen Sorten der Lilien in unterschiedlichen Farben dargestellt.

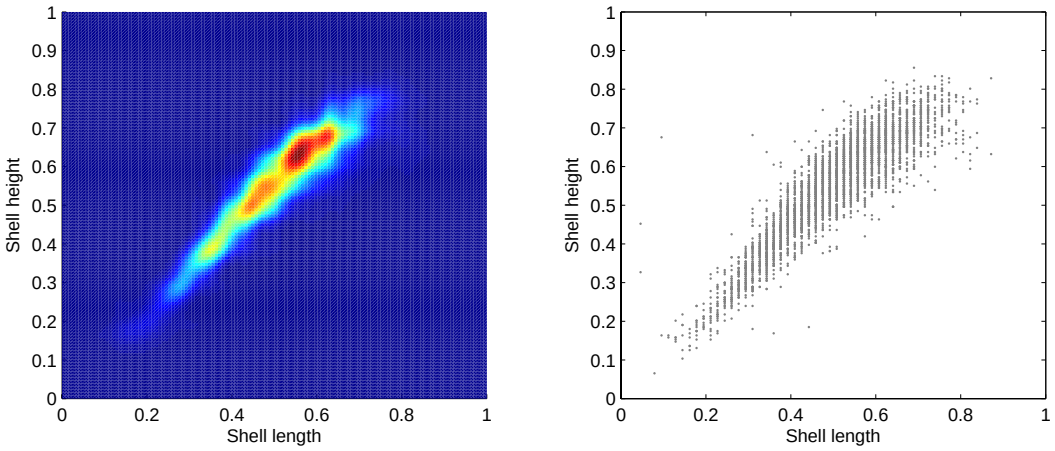

Abbildung 7.25: Links 2D Dichteschätzung des Abalone-Datensatzes auf Level 5. Rechts ein Scatterplot der Messwerte.

sionalen Dichtefunktionen noch nicht realisierbar war.

Abalone Daten

Der hier präsentierte Datensatz stammt ebenfalls vom Machine Learning Repository [Uni06] der University of California und besteht aus Messdaten von 4170 Ohrschnecken, auch Seeohr genannt. Diese Tiere sind in Japan eine begehrte Delikatesse und werden daher systematisch untersucht. Die Zielsetzung ist eine Altersbestimmung anhand der Messdaten. Dieses Beispiel wird angeführt, um den Vorteil einer Dichteschätzung bei der Präsentation und Analyse von Daten gegenüber klassischen Methoden zu verdeutlichen. In Abbildung 7.25 sind auf der rechten Seite Länge und Höhe der Schalen für alle vermessenen Tiere in einem so genannten *Scatterplot* dargestellt. Da die Messwerte ganzzahlige Millimeterwerte sind, liegen an vielen Stellen mehrere Messwerte übereinander. Im Gegensatz dazu liefert die auf der linken Seite von Abbildung 7.25 dargestellte Dichteschätzung wesentlich mehr Informationen. So wird hier deutlich, dass es drei Zentren der Verteilung gibt, an denen besonders viele Punkte liegen. Im einfachen Scatterplot ist diese Struktur nicht zu erkennen.

8 Zusammenfassung und Ausblick

In dieser Arbeit wurde eine neue, nichtlineare Methode der Dichteschätzung vorgestellt, die klassische, statistisch motivierte Ansätze mit effizienten Verfahren der numerischen Mathematik verbindet.

Es wurde in die Theorie der Dichteschätzung eingeführt und bekannte, häufig verwendete Verfahren wurden vorgestellt. Zwei bestehende Ansätze, die mit Regularisierungsnetzwerken [HHR99] beziehungsweise Support-Vektor-Maschinen [MV99] arbeiten, wurden angeführt und ein neues, ebenfalls als Regularisierungsnetzwerk darstellbares Verfahren wurde daraus entwickelt. Sowohl viele bestehende Verfahren der Dichteschätzung, als auch das in Kapitel 4 vorgestellte ecdf-basierte Verfahren, resultieren durch Wahl einer quadratischen Kostenfunktionen in *linearen* Gleichungssystemen. Im Gegensatz dazu ist bei dem in den Kapiteln 5 und 6 vorgestellten *Maximum a-posteriori Ansatz mit exponentiellen Familien und Gauß-Prior* ein *nichtlineares* Gleichunssystem zu lösen. Dabei wird ein üblicherweise in der Statistik zur parametrischen Dichteschätzung verwendetes Maximum-a-posteriori-Verfahren auf unendliche Parametervektoren bzw. Parameterfunktionen erweitert. Durch Verwendung exponentieller Familien von Verteilungen bleibt sehr viel Spielraum für die zu bestimmende Parameterfunktion, so dass ein solches Verfahren für eine große Klasse von Dichtefunktionen anwendbar ist.

Die Nichtlinearität wurde mit einer lokal quadratischen Approximation an das Zielfunktional behandelt, die in der gedämpften Variante mit Konvergenzordnung 1 gegen die Lösung konvergiert. In der ungedämpften Variante fällt die Konvergenz sogar quadratisch aus, kann aber nicht für alle Startwerte und Regularisierungsterme garantiert werden. Es entsteht ein iteratives Verfahren, welches in jeder Iteration ein lineares Gleichungssystem löst. Durch Verwendung dünner Gitter zur Diskretisierung ist es möglich moderat hochdimensionale Probleme zu behandeln und den Fluch der Dimension zu einem gewissen Teil zu umgehen. Da die Datenmenge nur linear in die Komplexität eingeht, ist das vorgestellte Verfahren für sehr große Datensätze geeignet.

Anhand numerischer Experimente wurde das Konvergenzverhalten des Verfahrens für verschieden große Datensätze analysiert und die steuernden Parameter untersucht. Dabei wurden unterschiedliche Auswirkungen des Glättungsparameters auf dem dünnen Gitter im Vergleich zum vollen Gitter beobachtet, die durch die Struktur der anisotropen Teilgitter resultieren. Die gemessenen Fehler der Dünngitterdiskretisierung unterscheiden sich nur gering von denen der Diskretisierung mit einem vollen Gitter. Je nach Diskretisierungslevel und Datenmenge liefert das dünne Gitter sogar bessere Ergebnisse. Weiterhin konnte festgestellt werden, dass sich die resultierenden Approximationen mit Hilfe der Eingabeparameter sehr gut steuern lassen und stabile Ergebnisse liefern.

Die hohe Qualität des resultierenden Schätzers wurde anhand simulierter ein-

und zweidimensionaler Datensätze verifiziert. Beim Vergleich mit anderen klassischen Verfahren wurden insbesondere für Dichtefunktionen, die mehrere unterschiedlich breite Zentren aufweisen, deutlich geringere Fehler gemessen. Der durch die Nichtlinearität verbundene Mehraufwand gegenüber linearen Dichteschätzungsverfahren lässt sich damit rechtfertigen.

Als Erweiterung der bisherigen Arbeit sind Experimente auf höherdimensionalen Datenmengen von Interesse, wobei dimensionsadaptive Ansätze, wie sie bereits in [Gar04] für Regressions- und Klassifikationsaufgaben verwendet werden, zu untersuchen sind. Bei den numerischen Experimenten wurde ein großer Unterschied der Auswirkung des Glättungsparameters festgestellt. Um hier eine Verbesserung zu erzielen, könnten in einer separaten Routine die Koeffizienten der Kombinationstechnik für jedes anisotrope Teilgitter optimiert werden. Ein solches Verfahren wird unter dem Namen *Opticom* bereits zur Regression erfolgreich eingesetzt [Gar06]. In dem Zusammenhang ist zusätzlich eine Parallelisierung des Verfahrens möglich, welche für die Kombinationstechnik in natürlicher Weise umsetzbar ist.

Um sehr hochdimensionale Datensätze mit relativ wenigen Datenpunkten behandeln zu können, kann als Erweiterung der Kern-basierte Ansatz des nichtlinearen Zielfunktionals weiter verfolgt werden, wofür die entstehenden nichtlinearen Integralgleichungen geeignet zu lösen sind. Auf diese Weise erhält man ein zweites nun datenbasiertes Verfahren, so dass je nach Gestalt der Datenmenge ein passendes Verfahren zur Verfügung steht.

Literaturverzeichnis

[Alt02] Alt, Hans Wilhelm: *Lineare Funktionalanalysis*. Springer, Berlin, 2 Auflage, 2002.

[Aro50] Aronszajn, N.: *Theory of reproducing kernels*. Transactions of the American Mathematical Society, 68, 1950.

[Bau92] Bauer, Heinz: *Maß- und Integrationstheorie*. de Gruyter, Auflage: 2., überarb. Auflage, 1992.

[Bau02] Bauer, Heinz: *Wahrscheinlichkeitstheorie*. de Gruyter, 5., durchges. und verb. Auflage, 2002.

[BG04] Bungartz, Hans-Joachim und Michael Griebel: *Sparse grids*. Acta Numerica, 13:1–123, 2004.

[BGRZ94] Bungartz, Hans-Joachim, Michael Griebel, Dierk Röschke und Christoph Zenger: *Pointwise Convergence of the Combination Technique for the Laplace Equation*. East-West Journal of Numerical Mathematics, 2(1):21–45, 1994.

[BL99] Berry, Michael und Gordon Linoff: *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley & Sons, Inc., New York, NY, USA, 1999.

[Bog98] Bogachev, Vladimir I.: *Gaussian Measures*, Band 62. American Mathematical Society, Rhode Island, 1 Auflage, 1998.

[Bun92] Bungartz, Hans-Joachim: *Dünne Gitter und deren Anwendung bei der adaptiven Lösung der dreidimensionalen Poisson-Gleichung*. Dissertation, Fakultät für Informatik, Technische Universität München, November 1992.

[Car97] Carreira, Miguel: *A Review of Dimension Reduction Techniques*. Technischer Bericht CS–96–09, Dept. of Computer Science, University of Sheffield, January 1997.

[EDD03] Elgammal, A., R. Duraiswami und L. Davis: *Efficient Kernel Density Estimation using the Fast Gauss Transform with Applications to Color Modeling and Tracking*, 2003.

[FE01] Frank Eibe, Ian H. Witten: *Data Mining, Praktische Werkzeuge und Techniken für das maschinelle Lernen*. Hanser Fachbuchverlag, 1 Auflage, 2001.

[Gar04] Garcke, Jochen: *Maschinelles Lernen durch Funktionsrekonstruktion mit verallgemeinerten dünnen Gittern.* Doktorarbeit, Institut für Numerische Simulation, Universität Bonn, 2004.

[Gar06] Garcke, Jochen: *Regression with the optimised combination technique.* In: *ICML '06: Proceedings of the 23rd international conference on Machine learning*, Seiten 321–328, New York, NY, USA, 2006. ACM Press.

[GH06] Griebel, Michael und Markus Hegland: *Gaussian priors for exponential families and infinite parameters - the maximum a-posteriori approach for probability densities.* 2006.

[Gir97] Girosi, Federico: *An Equivalence Between Sparse Approximation and Support Vector Machines.* Technischer Bericht AIM-1606, 1997.

[GM03] Gray, A. und A. Moore: *Rapid evaluation of multiple density models*, 2003.

[GSZ92] Griebel, Michael, Michael Schneider und Christoph Zenger: *A combination technique for the solution of sparse grid problems.* In: Groen, P. de und R. Beauwens (Herausgeber): *Iterative Methods in Linear Algebra*, Seiten 263–281, Elsevier, North Holland, 1992. IMACS.

[Hac89] Hackbusch, Wolfgang: *Integralgleichungen. Theorie und Numerik.* Teubner, 1989.

[HB02] Hanke-Bourgeois, Martin: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens.* Teubner, 2002.

[Heg06] Hegland, Markus: *An approximate Maximum a Posteriori Method with Gaussian Process Priors*, 2006.

[HGC05] Hegland, M., J. Garcke und V. Challis: *The combination technique and some generalisations.* 2005.

[HHR99] Hegland, Markus, Giles Hooker und Stephen Roberts: *Finite element thin plate splines in density estimation*, 1999.

[Hoo99] Hooker, Giles: *Developing a Spline-Smoothed Density.* School of Mathematical Sciences, Australian National University, November 1999. Honours Thesis.

[KS00] Katkovnik, Vladimir und Ilya Shmulevich: *Nonparametric Density Estimation with Adaptive Varying Window Size*, 2000.

[KS02] Katkovnik, Vladimir und Ilya Shmulevich: *Kernel density estimation with adaptive varying window size.* Pattern Recogn. Lett., 23(14):1641–1648, 2002.

[Lou] Louis, Alfred K.: *Inverse und schlecht gestellte Probleme.* Teubner Studienbuecher: Mathematik. Teubner, Stuttgart.

[MP95] Moghaddam, B. und A. Pentland: *Probabilistic Visual Learning for Object Detection.* In: *International Conference on Computer Vision (ICCV'95)*, Seiten 786–793, Cambridge, USA, June 1995.

[MV99] Mukherjee, Sayan und Vladimir Vapnik: *Multivariate Density Estimation: a Support Vector Machine Approach.* Technischer Bericht AIM-1653, 1999.

[Sco92] Scott, David W.: *Multivariate Density Estimation.* Wiley, New York, 1. Auflage, 1992.

[See04] Seeger, Matthias: *Gaussian Processes for Machine Learning*, 2004.

[Sil86] Silverman, B. W.: *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[Smo63] Smolyak, S. A: *Quadrature and interpolation formulas for Tensor Products of Certain Classes of Functions.* Soviet mathematics, 4:240–243, 1963.

[SS98] Smola, Alex J. und Bernhard Schölkopf: *From Regularization Operators to Support Vector Kernels.* In: Jordan, Michael I., Michael J. Kearns und Sara A. Solla (Herausgeber): *Advances in Neural Information Processing Systems*, Band 10. The MIT Press, 1998.

[SS01] Schölkopf, Bernhard und Alexander J. Smola: *Learning with Kernels: Support Vector Machines, Regularization,Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.

[SSM98] Smola, Alex J., Bernhard Schölkopf und Klaus-Robert Müller: *The connection between regularization operators and support vector kernels.* Neural Networks, 11(4):637–649, 1998.

[The06] The MathWorks: Matlab. `http://www.mathworks.de/products/matlab/`, November 2006.

[TS92] Terrell, George R. und David W. Scott: *Variable Kernel Density Estimation.* The Annals of Statistics, 20(3):1236–1265, 1992.

[Tur] Turlach, Berwin A.: *Bandwidth Selection in Kernel Density Estimation: A Review.*

[Uni06] University of California: UCI Machine Learning Repository. `http://www.ics.uci.edu/~mlearn/`, November 2006.

[Vap00] Vapnik, Vladimir N.: *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 2 Auflage, 2000.

[Wah90] Wahba, Grace: *Spline models for observational data.* SIAM [Society for Industrial and Applied Mathematics], 1990.

[WGS+99] Weston, J., A. Gammerman, M. Stitson, V. Vapnik, V. Vovk und C. Watkins: *Support vector density estimation*, 1999.

[WK71] Wahba, Grace und G. Kimeldorf: *Some results on tchebycheffian spline functions.* Journal of mathematical Aanalysis and Applications, 33:82–95, 1971.

[YA04] Yasemin Altun, Alex J. Smola und Thomas Hofmann: *Exponential Families for Conditional Random Fields.* In: *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, Seiten 2–9, Arlington, Virginia, 2004. AUAI Press.

[Zen91] Zenger, Christoph: *Sparse Grids.* In: Hackbusch, Wolfgang (Herausgeber): *Parallel Algorithms for Partial Differential Equations*, Band 31 der Reihe *Notes on Numerical Fluid Mechanics*, Seiten 241–251. Vieweg, 1991.