Fachhochschule Aachen

Campus Jülich

Fachbereich Medizintechnik und Technomathematik
Scientific Programming

Parametrisierung hochdimensionaler Verteilungen für Monte Carlo Anwendungen

Bachelorprojekt von Eugen Schmidt

Jülich, den 15. September 2010

Eigenständigkeitserklärung

Diese Arbeit ist von mir selbstständig angefertigt und verfasst. Es sind keine anderen als die angegebenen Quellen und Hilfsmittel benutzt worden.

Ort, Datum	Eugen Schmidt

Diese Arbeit wurde betreut von:

1. Prüfer: Prof. Dr. rer. nat. Gerhard Dikta
2. Prüfer: Prof. Dr. rer. nat. Detlev Reiter

Zusammenfassung

Das Institut für Energieforschung 4 - Plasmaphysik beschäftigt sich mit Forschungsarbeiten zur Kernfusion. Dazu betreibt es das Fusionsexperiment TEXTOR. TEXTOR ist ein Tokamak[1], der speziell zur Untersuchung der Wechselwirkung zwischen Plasmagas und den Wänden des Experiments ausgelegt ist. Begleitend zu den Arbeiten am Experiment werden physikalische Modelle zur Beschreibung des Plasmaverhaltens entwickelt und Simulationsrechnungen durchgeführt. Bei Fusionsexperimenten ist neben dem Verhalten der geladenen Teilchen auch das Verhalten der neutralen Teilchen von Bedeutung, da diese das Plasmaverhalten beeinflussen können. Die Beschreibung neutraler Teilchen ist nur durch eine 6-dimensionale Boltzmann-Gleichung möglich. Aufgrund der Komplexität dieser Gleichung wird die Simulation des Neutralgastransports mit Monte-Carlo-Verfahren durchgeführt. Dazu wird die Experimentgeometrie im Rechner nachgebildet, Neutralteilchen an durch Plasma-Wand-Kontakt definierten Stellen gestartet und deren Flugbahn verfolgt und ausgewertet. Dabei ist die Reflektion von Teilchen an den umgebenden Wänden wesentlicher Bestandteil. Weil die räumlichen und zeitlichen Skalen für Prozesse in den Festkörpern um viele Größenordnungen kleiner beziehungsweise schneller sind als die im Plasma, kann man diesen Vorgang von der eigentlichen Simulation im Plasma abtrennen und vorab durch Simulation eine Datenbank aufbauen, mit der diese Reflektions- und Zerstäubungsprozesse der Plasmasimulation zugänglich gemacht werden. Auch diese Vorabberechnung geschieht mit Monte-Carlo-Simulation. Das Ergebnis eines solchen Simulationslaufs für Wandprozesse (Reflektion, Erosion) sind Energie- und Raumwinkelverteilungen (Energie E, Polarwinkel θ und Azimutalwinkel ϕ) von zurück ins Plasma emittierten Neutralteilchen. Dabei ist es besonders effizient, diese in ein Produkt von drei univariaten bedingten Verteilungsfunktionen zu zerlegen und die Inversen dieser Verteilungen zu tabellieren. Anhand dieser Tabellen lässt sich die ursprüngliche Verteilung beispielsweise mit Hilfe von Monte-Carlo-Prozeduren reproduzieren. Dieses Verfahren wird seit rund 20 Jahren in unveränderter Form praktiziert. Nun stellt sich die Frage, ob es in seiner Form ausreichend genau und effizient ist oder optimiert werden muss. Zur Lösung dieses Problems werden statistische Tests wie der Kolmogorov-Smirnov-Test angewandt und diskutiert. Mit Hilfe des Kolmogorov-Smirnov-Tests wird ein Gütemaß definiert, mit dem man reproduzierte Verteilungen bewerten kann.

[1]Tokamak (russ.): Toroidale Kammer mit Magnetspulen

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Problemstellung und Motivation

Die Energie- und Raumwinkelverteilungen (drei Dimensionen), die bei einer Monte-Carlo-Simulation von Neutralgastransport entstehen, enthalten Informationen darüber, wie sich Neutralteilchen bei einer Kollision mit den Experimentwänden verhalten, das heißt wieviel Energie sie verlieren beziehungsweise behalten und in welcher Richtung sie ihre Flugbahn fortsetzen. Insgesamt ergibt solch eine Simulation relativ große Datenmengen. Anhand einer stochastischen Methode, der Inversionsmethode, ist es möglich diese Verteilungen auf relativ kleine Datenmengen zu reduzieren. Mit einer weiteren Monte-Carlo-Prozedur kann man die Ursprungsverteilungen reproduzieren.

Nun stellt sich die Frage, ob die Qualität der reproduzierten Verteilungen genügt, oder ob Verbesserungsmaßnahmen getroffen werden müssen. Zur Klärung dieser Frage wird das bisherige Verfahren zu verschiedenen Spezialfällen mit veränderten Parametern durchgeführt und Unterschiede bei den Ergebnissen erläutert. Für den Vergleich zweier Verteilungen, und zwar der Ursprungsverteilung mit der reproduzierten Verteilung, wird der Kolmogorov-Smirnov-Test verwendet.

Ziel dieser Arbeit ist es, ein eigenes Gütemaß zu definieren, mit dessen Hilfe eine Aussage darüber getroffen werden kann, wie man die Parameter wählen muss, um eine möglichst gute Reproduktion der Ursprungsverteilung zu erreichen.

Kapitel 2

Das Inversionsverfahren

2.1 Einleitung

In Monte-Carlo-Verfahren für Transportprobleme ist es allgemein üblich multivariate Verteilungen mit Zufallszahlen zu generieren. Je nachdem wie hoch die Genauigkeitsanforderungen sind, können dazu sehr große Datenmengen erforderlich sein. Hier kommt zum Beispiel das *Inversionsverfahren* zum Zuge.
Bei einer Monte-Carlo-Neutralgastransportsimulation von Wandprozessen entstehen dreidimensionale Wahrscheinlichkeitsverteilungen der re-emittierten und zerstreuten Teilchen, und zwar für die Reflektionsenergie E_r und die beiden Reflektionswinkel θ (polar) und ϕ (azimutal). Um diese Verteilung innerhalb einer Monte-Carlo-Simulation wieder zu "erwürfeln" und aus Effizienzgründen bietet es sich an, diese Daten in ein Produkt von drei univariaten bedingten Verteilungen zu zerlegen. In diesem Kapitel wird das Verfahren erläutert und anhand von Beispielen veranschaulicht.

2.2 Allgemeine Verfahrensbeschreibung

Es sei eine auf einem Intervall $[a; b)$ gleichverteilte Zufallsvariable U mit der Verteilungsfunktion $F(\xi)$ gegeben. Die Verteilungsfunktion F von U sieht folgendermaßen aus:

$$F(U \leq \xi) = \begin{cases} 0 & \text{für } \xi < a, \\ \frac{\xi - a}{b - a} & \text{für } a < \xi \leq b, \\ 1 & \text{für } \xi \geq b \end{cases}$$

Mit Hilfe der gleichverteilten Zufallsvariablen U können Realisierungen y_i, $i = 1, n$ der Zufallsvariablen Y mit [L5] erzeugt werden.

$$y_i = F^{-1}(u_i) \tag{2.1}$$

Dieses Verfahren ist für beliebige invertierbare Verteilungsfunktionen anwendbar. Liegt eine nicht streng monoton wachsende, also eine nicht invertierbare, Funktion F, beispielsweise eine Stufenfunktion, vor, wird die Inverse von F formal definiert durch:

$$F^{-1}(y) = inf\{x \in \mathbb{R} | F(x) \geq y\} \text{ für } a \leq y \leq b \tag{2.2}$$

Beweis: Anhand von 2.1 gilt:

$$P(Y \leq \xi) = P(F^{-1}(U) \leq \xi) = P(U \leq F(\xi)) \tag{2.3}$$

Da U gleichverteilt ist, gilt zudem:

$$P(U \leq F(\xi)) = F(\xi) \tag{2.4}$$

Aus 2.3 und 2.4 folgt schließlich:

$$P(Y \leq \xi) = F(\xi)$$

Durch dieses Verfahren ist es im Prinzip möglich mit gleichverteilten Zufallsvariablen beliebige Verteilungen auf $\mathbb{R}$ zu simulieren.

2.3 Beispiel - stetige Funktion

Als Beispiel für eine stetige Verteilung nehmen wir die Exponentialverteilung Exp(λ). Sei $F : \mathbb{R} \to [0; 1]$. Exp($\lambda$) ist für $\lambda > 0$ definiert durch[L6]:

$$F(x) = \begin{cases} 1 - e^{-\lambda * x} & \text{für } x \geq 0, \\ 0 & \text{für } x < 0 \end{cases}$$

Setze $F(x) = u$. Dann kommt man durch folgende Umformungen an die Inverse:

$$u = 1 - e^{-\lambda * x} \Leftrightarrow e^{-\lambda * x} = 1 - u \Leftrightarrow -\lambda * x = log(1 - u) \Leftrightarrow x = -\frac{log(1-u)}{\lambda}$$

Falls U und somit auch $(1 - U)$ gleichverteilt sind auf $[0; 1]$, gilt für die Umkehrfunktion von F:

$$F^{-1}(u) = -\frac{log(u)}{\lambda}$$

Somit gilt für die Zufallsvariable Y und deren Realisierungen y_i:

$$Y = -\frac{log(U)}{\lambda} \text{ und } y_i = -\frac{log(u_i)}{\lambda}$$

2.4 Beispiel - Stufenfunktion

Die folgenden Daten sind rein fiktiv und wurden für dieses Beispiel erfunden. Zur Beschreibung des Verfahrens für diskrete Verteilungen wird die Herstellung eines Spielwarenartikels herangezogen. Bei der Produktion müssen 30% aller Artikel einer weiteren Sicherheitsprüfung unterzogen werden. Die restlichen 70% können direkt in die Verpackungsabteilung überbracht werden. Beim Überprüfen der Artikel wird bei einem einwandfreien Zustand die Zahl 1 vergeben und bei Fehlern die 2. Die Produktion von 20 Artikeln wird laut genannter Vorschrift simuliert. Dazu wird eine auf dem Intervall $[0; 1)$ gleichverteilte Zufallsvariable U verwendet. Es wird also 20 mal gewürfelt. Dabei gelten die Wahrscheinlichkeiten:

- $P(Y = 1) = 0.7$
- $P(Y = 2) = 0.3$

Es wird nun so vorgegangen, dass 20 Zufallszahlen $u_i, i = 1, 20$ zwischen 0 (inklusive) und 1 (exklusive) erzeugt werden. Liegt die eine Zahl unter 0.7, so bekommt die Zahl y_i den Wert 1 zugewiesen, ansonsten eine 2. Unten stehende Tabelle soll einen solchen Vorgang simulieren.

Nr.	U	Klasse	Nr.	U	Klasse
1	0.8	2	11	0.4	1
2	0.2	1	12	0.8	2
3	0.4	1	13	0.9	2
4	0.2	1	14	0.3	1
5	0.8	2	15	0.5	1
6	0.3	1	16	0.3	1
7	0.9	2	17	0.6	1
8	0.0	1	18	0.6	1
9	0.6	1	19	0.3	1
10	0.1	1	20	0.9	2

Tabelle 2.1: 1. Spalte: Nr. des Artikels, 2. Spalte: gleichverteilte Zufallszahl U, 3. Spalte: Klassifizierung des Artikels

Die gleichverteilte Zufallsvariable erzeugte somit 14 Einsen - entspricht 70% - und 6 Zweien - entspricht 30%. Auf der Abbildung 2.1 ist die Zuordnung der ersten beiden Fälle dargestellt. Die erste Zufallszahl ist 0.8, also erhält der erste Artikel die Bewertung 2 und muss einer weiteren Prüfung unterzogen werden. Die zweite Zufallszahl ist 0.2, also erhält der zweite Artikel die Bewertung 1.

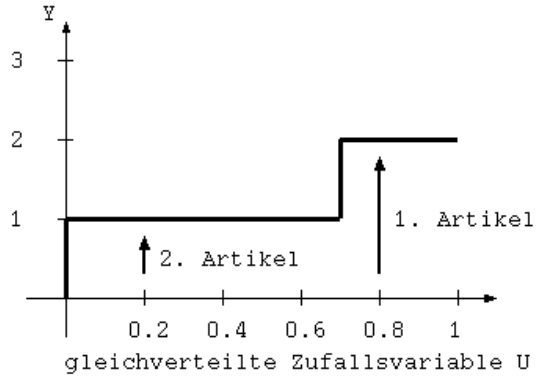


Abbildung 2.1: Zuordnung der ersten beiden Zufallszahlen

2.5 TRIM-Code Datenbank

Mit Hilfe des sogenannten TRIM-Codes, angewandt auf mehrdimensionale Datensätze beziehungsweise Verteilungen, werden Datenreduktionen durchgeführt. Diese reduzierten Daten stellen geglättete (getrim'te) Verteilungen der Ursprungsdaten dar. Anhand dieser Datenbanken und Monte-Carlo-Verfahren ist es möglich die Ursprungsdaten zu reproduzieren. Wie diese Datenbanken erzeugt werden, wird in den folgenden Unterkapiteln erklärt.

2.5.1 Einleitung

In einer Simulation werden zu verschiedenen Kombinationen aus Einfallswinkel $\alpha_i = 0°, 30°, 45°, 60°, 70°, 80°, 85°$ mit i = 1,...,7 und Einfallsenergie E_j = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000 eV mit j = 1,...,12 insgesamt also 84 (7*12) Datensätze erzeugt. Jeder Datensatz enthält bis zu n = 200000 dreidimensionale Punkte $(E, \theta, \phi)_i, i = 1, n$. Diese Größe n ist variabel und lässt sich im Prinzip noch beliebig erhöhen. Insgesamt ist es also eine große Menge an Daten, die in einer Simulation generiert werden. Um eine Datenreduktion durchzuführen und somit Speichereffizienz zu gewährleisten, wird das *Inversionsverfahren* auf dreidimensionale Daten erweitert und angewandt.

2.5.2 Verfahrensbeschreibung

Gegeben sei ein Satz von dreidimensionalen Daten (x_i, y_i, z_i) für $i = 1, n$, beispielsweise n = 200000. Um eine Datenreduktion durchzuführen, ist eine Sortierung der Daten in der ersten Richtung nötig. Bei der Sortierung der x-Werte werden die jeweils anderen zugehörigen y- und z-Werte mit verschoben.

Je nach gewünschter Präzision, legt man eine bestimmte Anzahl Quantile b (b steht für *bins*) fest. Quantile sind Lagemaße in der Statistik. Das p-Quantil Q_p gibt an, dass $p * 100\%$ der Werte einer geordneten Wertereihe unter Q_p liegen. Je nach gewünschter Anzahl bins, werden bei einem Datenumfang von n die Punkte x_p mit

$$p = \frac{n}{2*b} + (i-1) * \frac{n}{b}, i = 1, 2, ..., b$$

abgespeichert. i ist eine Laufvariable über alle Quantile. Die Punkte x_p sind Repräsentanten für die jeweiligen Teilbereiche, die somit geschaffen wurden. Diese Intervalle haben jeweils die Breite $\frac{n}{b}$. Bei $n = 200000$ Punkten und $b = 5$ wären demnach x_{20000}, x_{60000}, x_{100000}, x_{140000} und x_{180000} die abzuspeichernden Quantile in der ersten Dimension. Diese gehören zu den Wahrscheinlichkeiten 0.1, 0.3, 0.5, 0.7 und 0.9. Die 5 Teilbereiche haben jeweils die Breite $\frac{200000}{5} = 40000$. Die Punkte x_p stellen die Sprunghöhen der ersten inversen Verteilungsfunktion dar.
Der nächste Schritt ist die Sortierung der Daten in zweiter Dimension, also in y-Richtung. Diese geschieht allerdings immer nur teilweise, da die Daten in erster Dimension bereits in Teilbereiche aufgeteilt worden sind. In diesen Teilfeldern befinden sich jeweils $\frac{n}{b}$ Punkte. Hier werden die zugehörigen z-Werte jeweils mit verschoben. Je Teilbereich werden wiederum b Quantile ermittelt und abgespeichert. Bei n Punkten sind das alle y_p mit

$$p = \frac{n}{2*b^2} + (i-1) * \frac{n}{b} + (j-1) * \frac{n}{b^2}, i, j = 1, 2, ..., b.$$

i und j sind Laufvariablen jeweils über die Teilbereiche in erster Dimension (i) und über die Quantile in den Bereichen (j). Alle Quantile in einem Teilbereich i bilden zusammen die b bedingten Verteilungsfunktionen für die zweite Dimension.
Schließlich wird in dritter Dimension sortiert. Dies geschieht wiederum in den Teilbereichen, die bei der Sortierung in zweiter Dimension definiert worden sind. Die Teilbereiche bestehen jeweils aus $\frac{n}{b^2}$ Punkten. Zum Bestimmen der bedingten Verteilungsfunktionen in dritter Dimension, müssen die jeweiligen b Quantile z_p mit

$$p = \frac{n}{2*b^3} + (i-1) * \frac{n}{b} + (j-1) * \frac{n}{b^2} + (k-1) * \frac{n}{b^3}, i, j, k = 1, 2, ..., b$$

der Teilbereiche ermittelt werden. i, j und k sind Laufvariablen jeweils über die Teilbereiche in erster (i) Dimension, über die Teilbereiche in zweiter Dimension (j) und über die jeweiligen Quantile(k).
Anhand dieser tabellierten Quantile kann man die ursprünglichen Rohdaten reproduzieren. Wie dies im Einzelnen funktioniert, wird im Kapitel **Resampling** (siehe Kapitel 4) erläutert.

Kapitel 3

Kolmogorov-Smirnov-Test

3.1 Einleitung

Der Kolmogorov-Smirnov-Test ist ein statistischer Test, der es ermöglicht zu prüfen, ob zwei Wahrscheinlichkeitsverteilungen verschieden sind oder zu einem bestimmten Signifikanzniveau Gleichheit möglich ist. Dabei kann anhand von Stichproben einerseits geprüft werden, ob zwei Zufallsvariablen einer gemeinsamen, unbekannten Wahrscheinlichkeitsverteilung entnommen wurden oder ob eine Zufallsvariable von einer bestimmten Wahrscheinlichkeitsverteilung bezogen wurde. Da in unserem Fall nur ersteres, das heißt ob zwei Zufallsvariablen einer gemeinsamen, unbekannten Wahrscheinlichkeitsverteilung entnommen wurden, von Belang ist, gehen wir auch nur auf diesen Fall ein. Dabei lautet die Nullhypothese $H_0 : F_1(x) = F_2(x)$, die der Test zu widerlegen versucht, und die Alternativhypothese $H_1 : F_1(x) \neq F_2(x)$. Die Wahl des Stichprobenumfangs spielt bei diesem Test eine wesentliche Rolle. Wächst die Anzahl, so wird der Test schärfer und die Abweichungen umso deutlicher. Werden die Stichproben kleiner, so wird der Test weniger empfindlich, und Abweichungen sind seltener als "signifikant" einzustufen.

3.2 Verfahrensbeschreibung - 1-d

Der Vergleich zweier eindimensionaler Wahrscheinlichkeitsverteilungen mit dem Kolmogorov-Smirnov-Test ist recht trivial [L2]. Es seien zwei Stichproben A mit $a_1, a_2, ..., a_n$ und B mit $b_1, b_2, ..., b_m$ gegeben. Diese werden zunächst in aufsteigender Reihenfolge sortiert. Anhand der sortierten Stichproben werden jeweils Verteilungsfunktionen aufgestellt.

Für die erste Stichprobe gilt:

$$A(x) = \begin{cases} 0 & \text{für } x < a_1, \\ \frac{i}{n} & \text{für } a_i \leq x < a_{i+1}, i = 1, ..., n, \\ 1 & \text{für } a_n \leq x \end{cases} \tag{3.1}$$

Und für die zweite Stichprobe gilt:

$$B(x) = \begin{cases} 0 & \text{für } x < b_1, \\ \frac{i}{m} & \text{für } b_i \leq x < b_{i+1}, i = 1, ..., m, \\ 1 & \text{für } b_m \leq x \end{cases} \tag{3.2}$$

Die beiden Stichproben werden zu einer neuen Gesamtliste s vereint, doppelte Werte gelöscht und anschließend sortiert. Die Werte werden jeweils in $A(x)$ (3.1) und in $B(x)$ (3.2) ausgewertet, sowie die Differenz dieser Auswertungen berechnet. Die Kolmogorov-Smirnov-Statistik d bezeichnet den maximalen Wert aller $n * m$ Differenzen. Anhand des Stichprobenumfangs beider Verteilungen und des maximalen Abstands d kann mit Hilfe einer Formel eine Aussage darüber getroffen werden, ob die Nullhypothese zu einem bestimmten Signifikanzniveau α beibehalten oder abgelehnt werden kann. Das Signifikanzniveau gibt die Fehlerwahrscheinlichkeit an, mit der man eine Nullhypothese H_0 fälschlicherweise annimmt. Wird beispielsweise ein Signifikanzniveau von $\alpha = 0.05 = 5\%$ gewählt, so ergibt sich ein kritischer Wert $c(\alpha)$ laut folgender Formel [L2]:

$$c(\alpha) = k_\alpha * \sqrt{\frac{1}{n} + \frac{1}{m}} \tag{3.3}$$

Für kleine Stichprobengrößen ($n < 40$) gibt es Tabellen, denen k_α entnommen werden kann. Für große n lässt sich k_α annähern mit der Formel [L2]:

$$k_\alpha = \sqrt{-\frac{1}{2} * ln(\frac{\alpha}{2})} \tag{3.4}$$

Vergleicht man d mit $c(\alpha)$, kann schließlich eine Entscheidung getroffen werden. Ist $d < c(\alpha)$, kann die Annahme, die Verteilungen wären gleich, akzeptiert werden. Erreicht d den kritischen Wert, so wird die Nullhypothese abgelehnt.

3.3 Verfahrensbeschreibung - 2-d

Der Kolmogorov-Smirnov-Test für zweidimensionale Verteilungen ist nicht ganz so trivial wie im eindimensionalen Fall [L4]. Es seien zwei zweidimensionale Verteilungen X mit x_i, i=1,...,n, und Y mit y_j, i=j,...,m gegeben,

das heißt jeweils Datenmengen mit Wertepaaren (x, y). Gesucht wird die Kolmogorov-Smirnov-Statistik d also der maximale, kumulierte Abstand beider Verteilungen. Um diesen zu ermitteln, werden die Punkte beider Verteilungen X und Y jeweils als Ursprung eines Koordinatensystems betrachtet, alle umliegenden Punkte von X und Y quadrantweise gezählt und in Relation zur Gesamtzahl gesetzt abgespeichert. Im Folgenden ist $Q_{a,b,c}$ die Menge aller Punkte im a-ten Quadranten mit den Punkten aus der b-ten Verteilung als Ursprung für umliegende Punkte aus der c-ten Verteilung. Konkret bedeutet das für alle Zählungen:

- für alle Punkte von X als Ursprung und umliegende Punkte aus X:
  - $Q_{1,X,X} = \{(x, y) \in X | (x > x_i, y > y_i), i = 1, ..., n\}$
  - $Q_{2,X,X} = \{(x, y) \in X | (x < x_i, y > y_i), i = 1, ..., n\}$
  - $Q_{3,X,X} = \{(x, y) \in X | (x < x_i, y < y_i), i = 1, ..., n\}$
  - $Q_{4,X,X} = \{(x, y) \in X | (x > x_i, y < y_i), i = 1, ..., n\}$
- Analog zu den Punkten von X als Ursprung und umliegenden Punkten aus Y:
  - $Q_{1,X,Y} = \{(x, y) \in Y | (x > x_i, y > y_i), i = 1, ..., n\}$
  - $Q_{2,X,Y} = \{(x, y) \in Y | (x < x_i, y > y_i), i = 1, ..., n\}$
  - $Q_{3,X,Y} = \{(x, y) \in Y | (x < x_i, y < y_i), i = 1, ..., n\}$
  - $Q_{4,X,Y} = \{(x, y) \in Y | (x > x_i, y < y_i), i = 1, ..., n\}$
- Ebenso für alle Punkte von Y als Ursprung und umliegende Punkte aus X:
  - $Q_{1,Y,X} = \{(x, y) \in X | (x > x_i, y > y_i), i = 1, ..., m\}$
  - $Q_{2,Y,X} = \{(x, y) \in X | (x < x_i, y > y_i), i = 1, ..., m\}$
  - $Q_{3,Y,X} = \{(x, y) \in X | (x < x_i, y < y_i), i = 1, ..., m\}$
  - $Q_{4,Y,X} = \{(x, y) \in X | (x > x_i, y < y_i), i = 1, ..., m\}$
- Und schließlich alle Punkte von Y mit umliegenden Punkten aus Y:
  - $Q_{1,Y,Y} = \{(x, y) \in Y | (x > x_i, y > y_i), i = 1, ..., m\}$
  - $Q_{2,Y,Y} = \{(x, y) \in Y | (x < x_i, y > y_i), i = 1, ..., m\}$
  - $Q_{3,Y,Y} = \{(x, y) \in Y | (x < x_i, y < y_i), i = 1, ..., m\}$
  - $Q_{4,Y,Y} = \{(x, y) \in Y | (x > x_i, y < y_i), i = 1, ..., m\}$

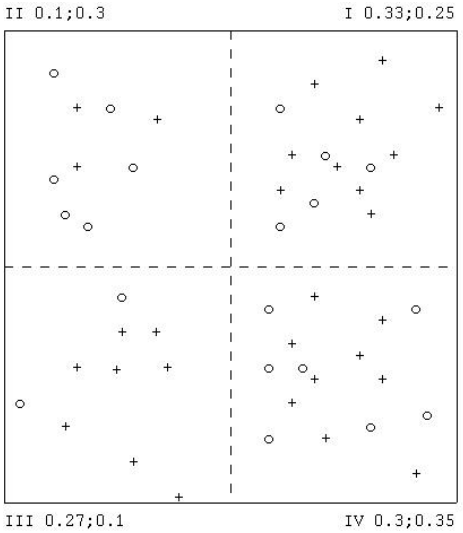
Es ergeben sich folgende Werte:

- $f_{1,X,X} = \frac{|Q_{1,X,X}|}{n}, f_{2,X,X} = \frac{|Q_{2,X,X}|}{n}, f_{3,X,X} = \frac{|Q_{3,X,X}|}{n}, f_{4,X,X} = \frac{|Q_{4,X,X}|}{n}$
- $f_{1,X,Y} = \frac{|Q_{1,X,Y}|}{m}, f_{2,X,Y} = \frac{|Q_{2,X,Y}|}{m}, f_{3,X,Y} = \frac{|Q_{3,X,Y}|}{m}, f_{4,X,Y} = \frac{|Q_{4,X,Y}|}{m}$
- $f_{1,Y,X} = \frac{|Q_{1,Y,X}|}{n}, f_{2,Y,X} = \frac{|Q_{2,Y,X}|}{n}, f_{3,Y,X} = \frac{|Q_{3,Y,X}|}{n}, f_{4,Y,X} = \frac{|Q_{4,Y,X}|}{n}$
- $f_{1,Y,Y} = \frac{|Q_{1,Y,Y}|}{m}, f_{2,Y,Y} = \frac{|Q_{2,Y,Y}|}{m}, f_{3,Y,Y} = \frac{|Q_{3,Y,Y}|}{m}, f_{4,Y,Y} = \frac{|Q_{4,Y,Y}|}{m}$

Nun werden jeweils für die X- und die Y-Punkte als Ursprung die maximalen Abstände $d1$ und $d2$ ermittelt. Diese ergeben sich folgendermaßen:

- $d1 = max(|f_{i,X,X} - f_{i,X,Y}|), i = 1, 4$
- $d2 = max(|f_{i,Y,X} - f_{i,Y,Y}|), i = 1, 4$

Da $d1$ und $d2$ unterschiedliche Ergebnisse haben können, je nachdem welche Verteilung man als Basis für die Koordinatensysteme wählt, errechnet sich das entgültige d als Mittel von $d1$ und $d2$: $d = \frac{d1+d2}{2}$. Abbildung 3.1 soll das zweidimensionle Verfahren veranschaulichen.



	+	o
I	10	5
II	3	6
III	8	2
IV	9	7
Gesamt	30	20

Abbildung 3.1: zweidimensionale Verteilungen von "+"- und "o"-Symbolen. Der größte Abstand befindet sich im II. Quadranten. Er liegt bei 0.3-0.1=0.2

Die Untersuchung des Ergebnisses d ist leider nicht genau definiert wie im eindimensionalen Fall. Es existieren approximative Formeln mit deren Hilfe man zweidimensionale Verteilungen miteinander vergleichen kann. Diese sind

abhängig vom Korrelationskoeffizienten der einzelnen Verteilungen von zweidimensionalen Verteilungen. In [L4], einem Kapitel der "numerical recipes", kann man Genaueres nachlesen. Aus Zeitgründen wird nicht näher auf dieses Verfahren eingegangen. Dies wird in den folgenden Tagen geschehen.

3.4 Verfahrensbeschreibung - 3-d

Aus Analogiegründen wird der dreidimensionale Kolmogorov-Smirnov-Test (KS-Test) erwähnt. Der KS-Test für zweidimensionale (2-d) Verteilungen lässt sich auf dreidimensionale (3-d) Fälle, das heißt für Verteilungen X und Y, die jeweils eine Sammlung von dreidimensionalen Daten - Tripeln (x, y, z) - beinhalten, erweitern. Dabei werden analog zum 2-d-Verfahren zunächst alle Punkte der ersten und zweiten Verteilung jeweils als Ursprung eines drei-dimensionalen Koordinatensystems betrachtet und alle umliegenden X-Punkte sowie alle umliegenden Y-Punkte in den einzelnen Oktanten gezählt, die relativen Häufigkeiten berechnet, die maximalen Abstände bestimmt und schließlich d berechnet. Verwendet man ein dreidimensionales Koordinatensystem nach der Linke-Hand-Regel (Abbildung 3.2), so lassen sich die Punkte beispielsweise aus X jeweils folgendermaßen den einzelnen Oktanten zuordnen:

- 1. Quadrant: $\{(x, y, z) \in X | (x > x_i, y > y_i, z > z_i)\}$
- 2. Quadrant: $\{(x, y, z) \in X | (x < x_i, y > y_i, z > z_i)\}$
- 3. Quadrant: $\{(x, y, z) \in X | (x < x_i, y < y_i, z > z_i)\}$
- 4. Quadrant: $\{(x, y, z) \in X | (x > x_i, y < y_i, z > z_i)\}$
- 5. Quadrant: $\{(x, y, z) \in X | (x > x_i, y > y_i, z < z_i)\}$
- 6. Quadrant: $\{(x, y, z) \in X | (x < x_i, y > y_i, z < z_i)\}$
- 7. Quadrant: $\{(x, y, z) \in X | (x < x_i, y < y_i, z < z_i)\}$
- 8. Quadrant: $\{(x, y, z) \in X | (x > x_i, y < y_i, z < z_i)\}$

Eine Näherungsformel zur Beurteilung der Kolmogorov-Smirnov-Statistik wie im zweidimensionalen Fall war leider nicht auffindbar.

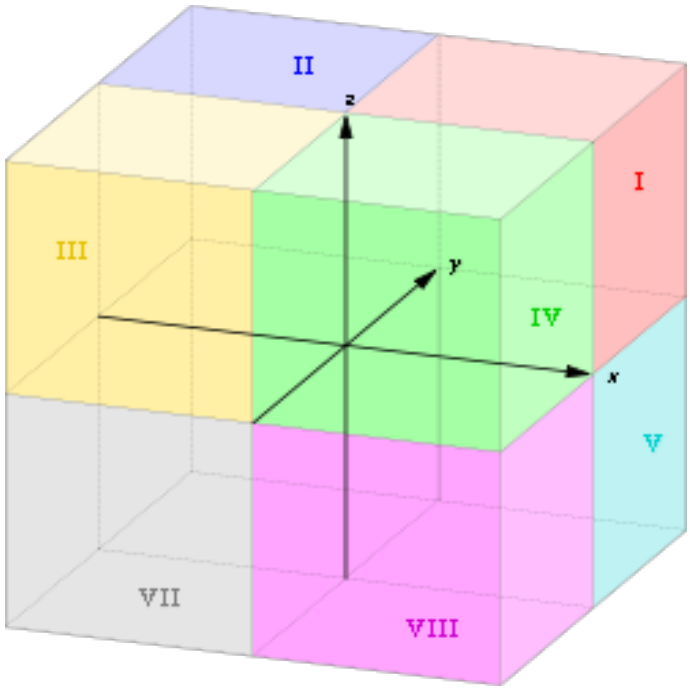


Abbildung 3.2: dreidimensionales Koordinatensystem nach der Linke-Hand-Regel. Quelle: Wikipedia [L7]

Kapitel 4

Resampling

4.1 Allgemein

Hat man auf eine mächtige Datenmenge eine Datenreduktion angewandt, so wie es in Kapitel 1.5 erklärt ist, kann die Verteilung der Rohdaten approximativ reproduziert werden. Dazu werden aus den tabellierten b Quantilen stufenförmige Verteilungsfunktionen F zusammengestellt. Die Quantile stellen die Sprungstellen von F dar. Die y-Achse der Funktion erstreckt sich über die Wahrscheinlichkeiten von 0 bis 1, also von 0% bis 100%. Da die Quantile den Ursprungsdaten äquidistant entnommen wurden, gelten für diese gleiche Wahrscheinlichkeiten. Somit ergeben sich äquidistante Sprunghöhen für F. Diese haben jeweils die Höhe $\frac{1}{b}$. Als nächstes wird F invertiert. F^{-1} ist eine Stufenfunktion und setzt sich folgendermaßen zusammen:

$$F^{-1}(\xi) = \begin{cases} Q_{\frac{1}{b}} & \text{für } 0 \leq \xi < \frac{1}{b}, \\ Q_{\frac{2}{b}} & \text{für } \frac{1}{b} \leq \xi < \frac{2}{b}, \\ ... & \\ Q_{\frac{b-1}{b}} & \text{für } \frac{b-2}{b} \leq \xi < \frac{b-1}{b}, \\ Q_{\frac{b}{b}=1} & \text{für } \frac{b-1}{b} \leq \xi < 1 \end{cases}$$

Durch Erzeugung gleichverteilter Zufallszahlen ξ und linearer Interpolation zwischen den Quantilen, erhält man eine Annäherung für die Rohdaten. Liegt ξ zwischen den Sprungstellen $x_{\frac{b-i}{b}}$ und $x_{\frac{b-i+1}{b}}, i = 1, ..., b$, so wird zwischen $F^{-1}(x_{\frac{b-i}{b}})$ und $F^{-1}(x_{\frac{b-i+1}{b}})$ linear interpoliert nach folgender Vorgabe:

$$F^{-1}(\xi) = F^{-1}(x_{\frac{b-i}{b}}) + \frac{F^{-1}(x_{\frac{b-i+1}{b}}) - F^{-1}(x_{\frac{b-i}{b}})}{x_{\frac{b-i+1}{b}} - x_{\frac{b-i}{b}}} * (\xi - x_{\frac{b-i}{b}}) \qquad (4.1)$$

4.1.1 Beispiel mit 5 bins

Es sei eine Datenreduktion mit $b = 5$ Quantilen durchgeführt worden und die Quantile gegeben mit $Q_p = \{1, 4, 5, 8, 10\}$, $p = 0.1, 0.3, 0.5, 0.7$ und 0.9. Die Verteilungsfunktion $F(x)$ für diese Werte sieht folgendermaßen aus:

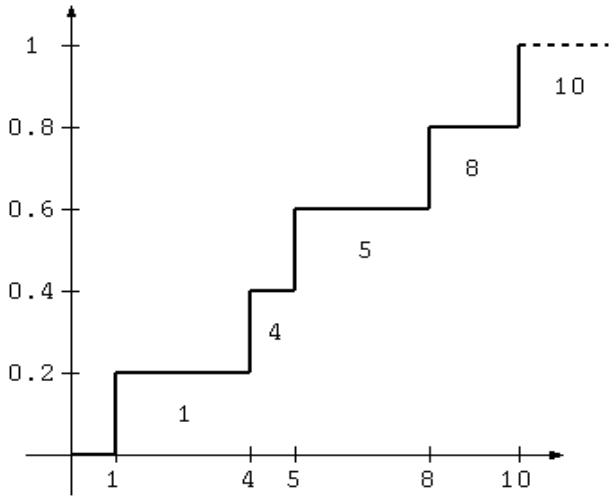


Abbildung 4.1: empirische Verteilungsfunktion $F(x)$ mit 5 bins

Die Inverse zu $F(x)$ und die durch Zufallszahlen und Interpolationen approximierte Verteilung sehen folgendermaßen aus:

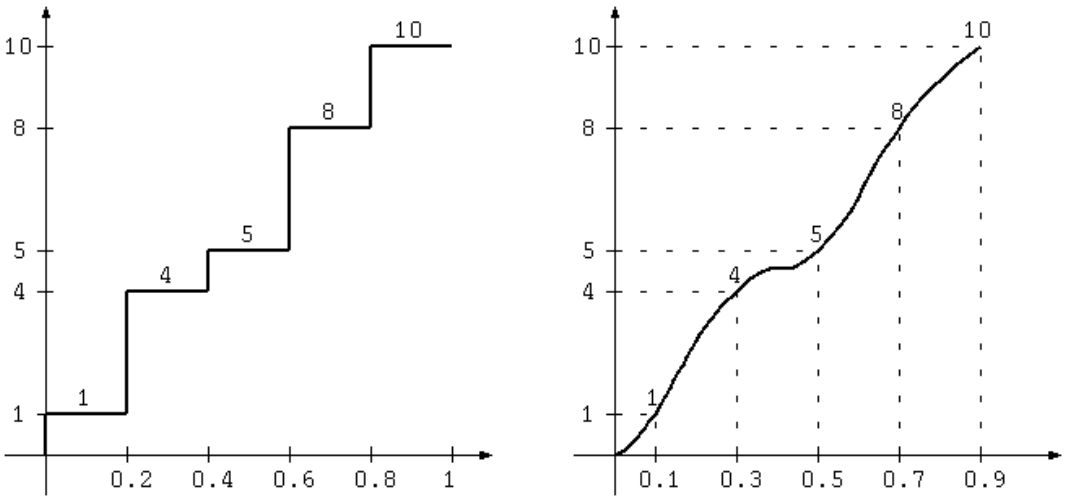


Abbildung 4.2: Inverse $F^{-1}(x)$ und angenäherte Verteilung

4.1.2 Beispiel mit 10 bins

Es sei eine Datenreduktion mit $b = 10$ Quantilen durchgeführt worden sein und die Quantile gegeben mit $Q_p = \{0.4, 2, 3.7, 4.2, 4.5, 5.8, 7.3, 8.5, 9.4, 10.3\}$, $p = 0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85$ und 0.95. Die Verteilungsfunktion $F(x)$ für diese Werte sieht folgendermaßen aus:

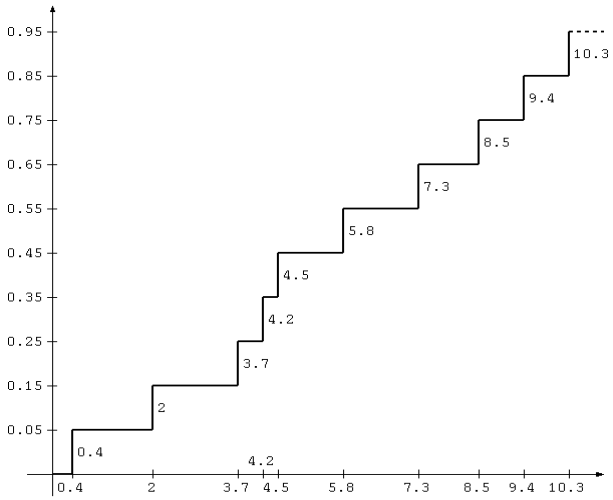


Abbildung 4.3: empirische Verteilungsfunktion $F(x)$ mit 10 bins

Die Inverse zu $F(x)$ und die durch Zufallszahlen und Interpolationen approximierte Verteilung sehen folgendermaßen aus:

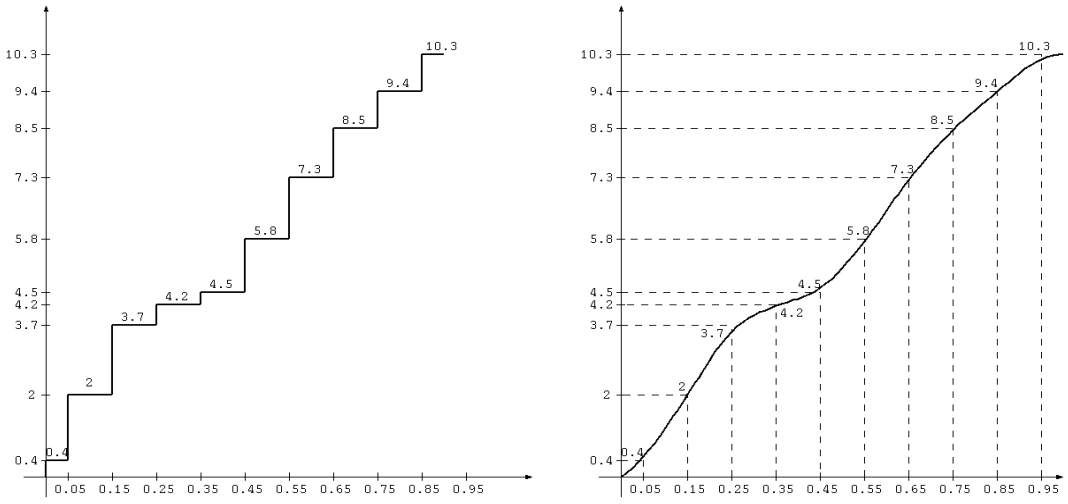


Abbildung 4.4: Inverse $F^{-1}(x)$ und angenäherte Verteilung

4.2 Einleitung - TRIM-Code

Die mit dem beschriebenen Verfahren erzeugten empirischen Verteilungsfunktionen werden in tabellarischer Form gespeichert. Für jede Kombination von Projektilsorte und Wandmaterial werden 12 verschiedene Teilchenenergien, 7 verschiedene Einfallswinkel untersucht und 84 Dateien erzeugt. Jede Datei wird für je einen bestimmten Einfallswinkel und eine bestimmte Energie des einfallenden Teilchens erzeugt und speichert Daten in drei Dimensionen, nämlich die Reflektionsenergie E_r und die beiden Reflektionswinkeln θ und ϕ. Die verwendeten α_i sind 0°, 30°, 45°, 60°, 70°, 80°, 85° (n=7) und die Energien der einfallenden Teilchen E_j sind 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000 eV (m=12). In jeder Dimension ist eine feste Zahl b (b steht im folgenden für *bins*) an Intervallen vorhanden. Wie die Daten reproduziert werden können, soll an einem Beispiel gezeigt werden. Angenommen wir wollen die Verteilung für einen Einfallswinkel von $\alpha 0 = 50°$ und eine Einfallsenergie von $E0 = 150$ eV reproduzieren. Die einzelnen Schritte werden in den nachfolgenden Unterkapiteln erläutert.

4.3 Reflektionswahrscheinlichkeit

Die erste Zeile eines vom TRIM-Code erzeugten Datensatzes enthält Input-Daten wie Ordnungszahl und Masse des einfliegenden Teilchens und des Wandmaterials, die Teilchenenergie und den Einfallswinkel sowie die Reflektionswahrscheinlichkeit R_N etc.. Für den Einfall von Deuterium auf Kohlenstoff im 45°-Winkel und einer Energie von 100 eV sieht die Zeile folgendermaßen aus:

```
1  2.01  6 12.01 0.10E+03 45.0 3.37E-01 1.58E-01  50.0 200000
```

R_N basiert auf den Eingabedaten $\alpha 0$ und $E0$ und wird linear interpoliert. Dazu werden $\alpha 0$ und $E0$ den passenden Intervallen zugeordnet. Die einzelnen Intervalle beziehungsweise deren Grenzen werden durch die in der Einleitung genannten α_i und E_j dargestellt. Es werden α_{i1} und α_{i2}, sowie E_{j1} und E_{j2} gesucht für die gilt: $\alpha_{i1} < \alpha 0 \leq \alpha_{i2}$ und $E_{j1} < E0 \leq E_{j2}$, $i2 \in \{1, 2, ..., 7\}, i2 = i1 + 1$, $j2 \in \{1, 2, ..., 12\}$, $j2 = j1 + 1$. Für das Beispiel wären die Intervallgrenzen $\alpha_{i1} = 45°$ und $\alpha_{i2} = 60°$, sowie $E_{j1} = 100$ eV und $E_{j2} = 200$ eV. Die Reflektionswahrscheinlichkeit wird zwischen den R_N aus den Datensätzen mit (α_i, E_j): $(45°, 100eV)$, $(45°, 200eV)$, $(60°, 100eV)$ und $(60°, 200eV)$ interpoliert. Konkret ergibt die Interpolation zwischen den R_N von $(45°, 100eV)$ und $(45°, 200eV)$ den Wert $rf1$. $rf2$ entsteht durch Interpolation der R_N aus $(60°, 100eV)$ und $(60°, 200eV)$. *rprob* wird schließlich

durch $rf1$ und $rf2$ interpoliert und stellt die gesuchte Reflektionswahrscheinlichkeit für $\alpha 0$ und $E0$ dar. Eine Grafik soll diesen Vorgang veranschaulichen:

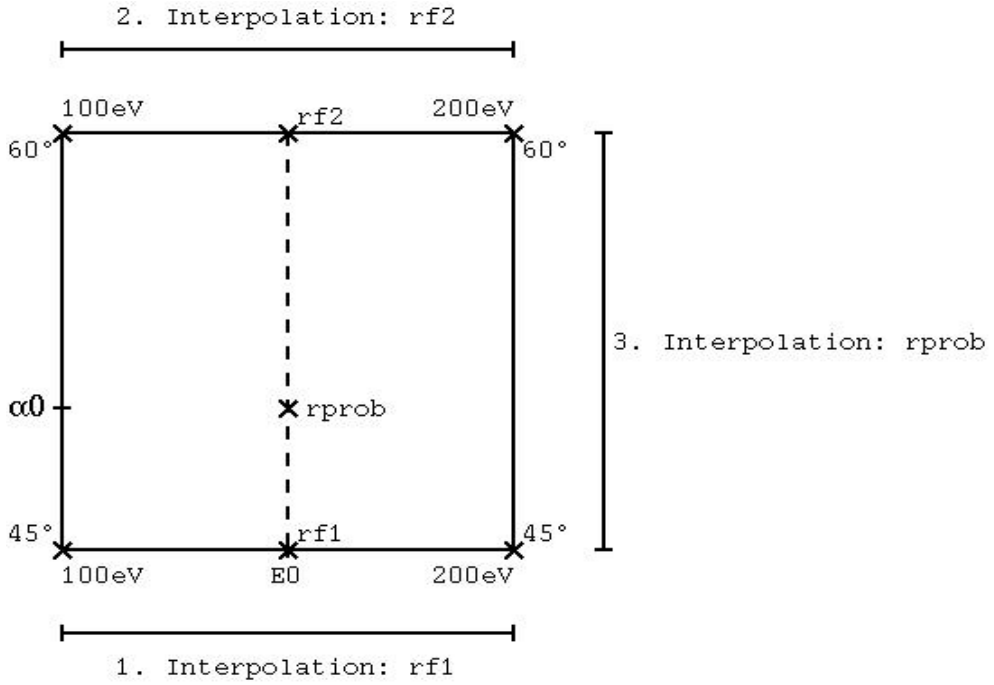


Abbildung 4.5: Interpolation der Reflektionswahrscheinlichkeit

4.4 Reflektionsenergie

Zur Bestimmung der Reflektionsenergie E_r wird das Verfahren um eine Dimension erweitert. Da das Resampling der Daten ebenfalls eine Monte-Carlo-Prozedur ist, basiert die Berechnung von E_r sowohl auf den Eingabedaten $\alpha 0$ und $E0$, als auch auf einer Zufallszahl. Bei $b = 5$ werden die Sprungstellen der Verteilungsfunktion in erster Dimension - beispielsweise für den Einfall von Deuterium auf Kohlenstoff im 45°-Winkel und einer Energie von 100 eV - wie folgt tabelliert abgespeichert:

```
1.36920E+01  3.29970E+01  4.75550E+01  6.03200E+01  7.55430E+01
```

Wie auch bei der Interpolation von R_E werden die vier Datensätze für (α_i, E_j): $(45°, 100eV)$, $(45°, 200eV)$, $(60°, 100eV)$ und $(60°, 200eV)$ herangezogen. Bis zum entgültigen Wert für E_r wird eine Zufallszahl generiert und fünf Mal interpoliert:

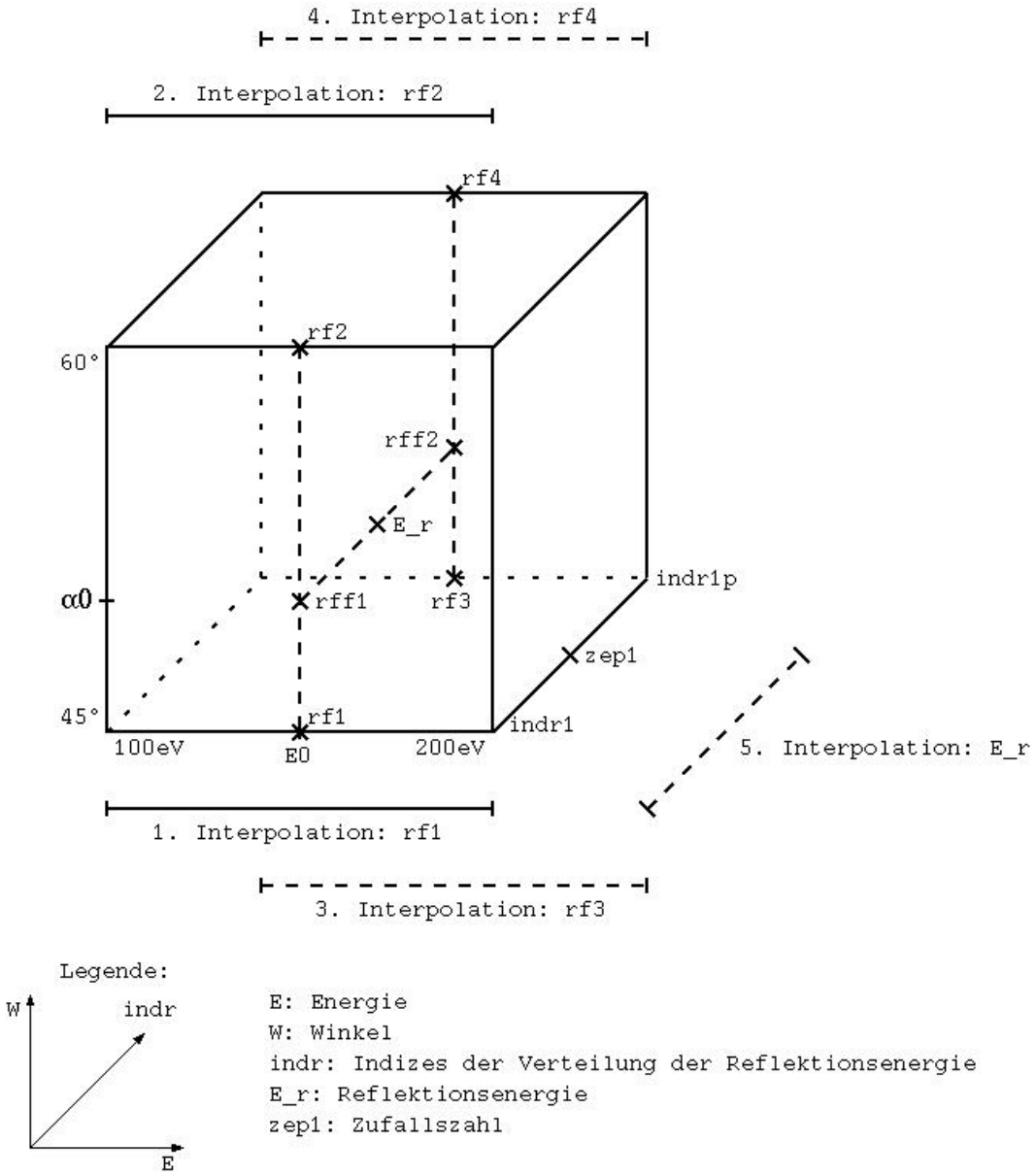


Abbildung 4.6: Interpolation der Reflektionsenergie

- $zep1$: Zufallszahl ξ generieren, $0 < \xi \leq 1$
  Aufgrund dieser Zufallszahl wird entschieden zwischen welchen Werten der Verteilungsfunktion interpoliert wird. Man merke sich die Indizes $indr1$ und $indr1p$ der beiden möglichen Reflektionsenergien. Diese werden weiterhin bei der Berechnung des Polar- und des Azimutalwinkels in den nächsten Unterkapiteln benötigt. Es sei $\xi = 0.2$, dann sind

$indr1 = 1$ und $indr1p = 2$ bei $b = 5$.

- lineare Interpolation zwischen den je 1. ($indr1$) und 2. ($indr1p$) E_r aus den folgenden Datensätzen ergeben folgende Werte:
  - Zwischenwerte, die sich aus linearer Interpolation ergeben:
    - $rf1$: 1. E_r aus $(45°, 100eV)$ und $(45°, 200eV)$
    - $rf2$: 1. E_r aus $(60°, 100eV)$ und $(60°, 200eV)$
    - $rf3$: 2. E_r aus $(45°, 100eV)$ und $(45°, 200eV)$
    - $rf4$: 2. E_r aus $(60°, 100eV)$ und $(60°, 200eV)$
  - weitere Zwischenwerte, die aus linearer Interpolation entstehen:
    - $rff1 : rf1$ und $rf2$
    - $rff2 : rf3$ und $rf4$
- Die lineare Interpolation zwischen $rff1$ und $rff2$ ergibt schließlich die reflektierte Energie des Teilchens. Abbildung 4.6 dient zur Veranschaulichung.

4.5 Polarwinkel

Die Komplexität der Berechnung des Polarwinkels θ des reflektierten Teilchens erhöht sich wieder um eine Stufe. Zusätzlich zu weiteren Interpolationen basiert die Bestimmung des Polarwinkels auf einer zweiten Zufallszahl η. Die Sprungstellen der Verteilungsfunktion in zweiter Dimension werden tabelliert als ein $b \times b$-Block gespeichert. Dieser Block sieht beispielsweise für den Einfall von Deuterium auf Kohlenstoff im 45°-Winkel und einer Energie von 100 eV so aus:

```
4.96950E-01  6.90340E-01  8.03770E-01  8.92570E-01  9.66460E-01
4.76970E-01  6.61790E-01  7.83520E-01  8.82550E-01  9.63810E-01
4.52720E-01  6.38840E-01  7.62850E-01  8.66800E-01  9.56930E-01
4.07560E-01  5.76370E-01  7.03260E-01  8.20380E-01  9.32880E-01
2.76260E-01  4.02950E-01  4.97780E-01  5.98230E-01  7.40260E-01
```

Die Indizes $indr1$ und $indr1p$ geben immer die beiden Zeilen an, in denen interpoliert wird. Das Ergebnis von η bestimmt die Wahl der Spalten. Zur Berechnung des entgültigen Werts finden mehrere Zwischeninterpolationen statt und eine zweite Zufallsvariable wird generiert:

- Es sei $\eta = 0.7$. Somit sind die Spaltenindizes bei $b = 5$ $indr2 = 3$ und $indr2p = 4$. Diese Werte sowie $indr1$ und $indr1p$ werden später auch für die Berechnung des Azimutalwinkels ϕ benötigt.

- Zwischenwerte, die sich aus linearer Interpolation ergeben:
  - $rf1 : \theta$ aus der 1. Zeile und 3. Spalte der Datensätze $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf2 : \theta$ aus der 1. Zeile und 3. Spalte der Datensätze $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf3 : \theta$ aus der 2. Zeile und 3. Spalte der Datensätze $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf4 : \theta$ aus der 2. Zeile und 3. Spalte der Datensätze $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf5 : \theta$ aus der 1. Zeile und 4. Spalte der Datensätze $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf6 : \theta$ aus der 1. Zeile und 4. Spalte der Datensätze $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf7 : \theta$ aus der 2. Zeile und 4. Spalte der Datensätze $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf8 : \theta$ aus der 2. Zeile und 4. Spalte der Datensätze $(60°, 100eV)$ und $(60°, 200eV)$
- Weitere Zwischenwerte, die durch lineare Interpolation aus entstehen:
  - $rff1 : rf1$ und $rf2$
  - $rff2 : rf3$ und $rf4$
  - $rff3 : rf5$ und $rf6$
  - $rff4 : rf7$ und $rf8$
- Weitere Zwischenwerte, die aus linearer Interpolation entstehen:
  - $rfff1 : rff1$ und $rff2$
  - $rfff2 : rff3$ und $rff4$

Der entgültige Wert für den Polarwinkel θ geht schließlich aus der linearen Interpolation zwischen $rfff1$ und $rfff2$ hervor.

4.6 Azimutalwinkel

Die Komplexität bei der Berechnung des Azimutalwinkels ϕ nimmt wieder um eine Dimension zu. Das bringt mehr Interpolationen und eine weitere Zufallszahl mit sich. Die Werte für ϕ werden in einem Block der Größe

$(b * b) \times b$ gespeichert. Diesen Block kann man sich vorstellen als eine Aneinanderreihung von $b \times b$-Blöcken. Dieser sieht beispielsweise für den Einfall von Deuterium auf Kohlenstoff im 45°-Winkel und einer Energie von 100 eV mit 5 bins folgendermaßen aus:

```
-9.37740E-01 -5.13370E-01  7.25400E-02  5.92080E-01  9.39070E-01
-9.39950E-01 -6.09920E-01 -2.07590E-02  5.79210E-01  9.57290E-01
-9.52570E-01 -5.81780E-01 -3.08810E-02  5.64880E-01  9.37180E-01
-9.37250E-01 -5.62670E-01  2.94250E-02  5.65240E-01  9.55160E-01
-9.51410E-01 -5.84080E-01 -1.23710E-02  5.86230E-01  9.44940E-01
-9.58560E-01 -6.46780E-01 -6.73100E-02  5.12850E-01  9.53030E-01
-9.55590E-01 -5.95790E-01  5.25020E-03  5.56710E-01  9.41550E-01
-9.58500E-01 -6.43460E-01 -4.76590E-02  5.58000E-01  9.41870E-01
-9.44380E-01 -6.11800E-01 -1.19610E-02  6.15630E-01  9.51870E-01
-9.45870E-01 -5.96780E-01 -2.86020E-03  5.68340E-01  9.41670E-01
-9.60920E-01 -6.28610E-01  7.72800E-02  6.51340E-01  9.70830E-01
-9.50610E-01 -6.10580E-01 -3.68370E-02  6.43300E-01  9.65260E-01
-9.43010E-01 -6.19940E-01 -3.80200E-02  5.91760E-01  9.50520E-01
-9.48720E-01 -5.91810E-01  1.73900E-02  6.00340E-01  9.61300E-01
-9.56690E-01 -6.30530E-01 -4.40530E-02  5.59690E-01  9.44840E-01
-4.44510E-01 -5.35190E-02  4.14520E-01  7.89640E-01  9.72970E-01
-3.75160E-01  8.82550E-02  4.56900E-01  8.10200E-01  9.80150E-01
-3.26500E-01  1.52520E-01  5.48780E-01  8.40820E-01  9.78420E-01
-3.26390E-01  2.20730E-01  5.92190E-01  8.55190E-01  9.86260E-01
-5.43900E-01  1.16290E-01  5.63400E-01  8.41820E-01  9.79960E-01
 3.90030E-01  6.31790E-01  7.90890E-01  9.23300E-01  9.91390E-01
 4.36020E-01  6.94530E-01  8.53260E-01  9.49370E-01  9.93600E-01
 4.90430E-01  7.55210E-01  8.80380E-01  9.58200E-01  9.95680E-01
 4.64230E-01  7.76450E-01  8.92620E-01  9.61270E-01  9.95690E-01
 4.78350E-01  7.69640E-01  9.05100E-01  9.64870E-01  9.96210E-01
```

Innerhalb der vier Datensätzen für (α_i, E_j): (45°, 100 eV), (45°, 200eV), (60°, 100eV) und (60°, 200eV) geben $indr1$ und $indr1p$ die Teilblöcke, $indr2$ und $indr2p$ die Zeilen in den Teilblöcken und die neuen Indizes $indr3$ und $indr3p$, beide abhängig von der neuen Zufallszahl, die Spalten an, in denen gesucht wird. Die Berechnung geht wieder über mehrere Zwischenschritte, die Interpolationen beinhalten:

- Zufallszahl ζ generieren. Es sei $\zeta = 0.4$, dann sind $indr3 = 2$ und $indr3p = 3$.

- lineare Interpolationen ergeben:

  - $rf1 : \phi$ aus 1. Teilblock, 3. Zeile im Teilblock, 2. Spalte, Datensatz: $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf2 : \phi$ aus 1. Teilblock, 3. Zeile im Teilblock, 2. Spalte, Datensatz: $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf3 : \phi$ aus 2. Teilblock, 3. Zeile im Teilblock, 2. Spalte, Datensatz: $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf4 : \phi$ aus 2. Teilblock, 3. Zeile im Teilblock, 2. Spalte, Datensatz: $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf5 : \phi$ aus 1. Teilblock, 4. Zeile im Teilblock, 2. Spalte, Datensatz: $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf6 : \phi$ aus 1. Teilblock, 4. Zeile im Teilblock, 2. Spalte, Datensatz: $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf7 : \phi$ aus 2. Teilblock, 4. Zeile im Teilblock, 2. Spalte, Datensatz: $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf8 : \phi$ aus 2. Teilblock, 4. Zeile im Teilblock, 2. Spalte, Datensatz: $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf9 : \phi$ aus 1. Teilblock, 3. Zeile im Teilblock, 3. Spalte, Datensatz: $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf10 : \phi$ aus 1. Teilblock, 3. Zeile im Teilblock, 3. Spalte, Datensatz: $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf11 : \phi$ aus 2. Teilblock, 3. Zeile im Teilblock, 3. Spalte, Datensatz: $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf12 : \phi$ aus 2. Teilblock, 3. Zeile im Teilblock, 3. Spalte, Datensatz: $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf13 : \phi$ aus 1. Teilblock, 4. Zeile im Teilblock, 3. Spalte, Datensatz: $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf14 : \phi$ aus 1. Teilblock, 4. Zeile im Teilblock, 3. Spalte, Datensatz: $(60°, 100eV)$ und $(60°, 200eV)$
  - $rf15 : \phi$ aus 2. Teilblock, 4. Zeile im Teilblock, 3. Spalte, Datensatz: $(45°, 100eV)$ und $(45°, 200eV)$
  - $rf16 : \phi$ aus 2. Teilblock, 4. Zeile im Teilblock, 3. Spalte, Datensatz: $(60°, 100eV)$ und $(60°, 200eV)$

- weitere Interpolationen ergeben:

  - $rff1 : rf1$ mit $rf2$

  - $rff2 : rf3$ mit $rf4$
  - $rff3 : rf5$ mit $rf6$
  - $rff4 : rf7$ mit $rf8$
  - $rff5 : rf9$ mit $rf10$
  - $rff6 : rf11$ mit $rf12$
  - $rff7 : rf13$ mit $rf14$
  - $rff8 : rf15$ mit $rf16$

- weitere Interpolationen ergeben:
  - $rfff1 : rff1$ mit $rff2$
  - $rfff2 : rff3$ mit $rff4$
  - $rfff3 : rff5$ mit $rff6$
  - $rfff4 : rff7$ mit $rff8$

- weitere Interpolationen ergeben:
  - $rffff1 : rfff1$ mit $rfff2$
  - $rffff2 : rfff3$ mit $rfff4$

Der entgültige Wert für den Azimutalwinkel ϕ ergibt sich aus der linearen Interpolation zwischen $rffff1$ und $rffff2$.

Kapitel 5

Ergebnisse

5.1 Einleitung

In diesem Kapitel geht es darum, die Güte des Verfahrens, das in Kapitel 1.5 beschrieben ist, zu bestimmen. Zuerst werden dazu aus gegebenen Rohdaten multivariate bedingte Verteilungsfunktionen mit 5 und mit 10 bins aufgestellt. Aus diesen Verteilungsfunktionen werden unterschiedlich große Datensätze reproduziert. Anhand des Kolmogorov-Smirnov-Tests werden diese Datensätze mit den Rohdaten verglichen. Um die Güte der reproduzierten Verteilungen zu beurteilen, definieren wir ein eigenes Gütemaß. Hierfür werden die Formeln 3.3 und 3.4 - zu finden in Kapitel 3 - verwendet. Das Gütemaß sieht folgendermaßen aus (3.4 in 3.3 eingesetzt):

$$c(\alpha) = \sqrt{-\frac{1}{2} * ln(\frac{\alpha}{2})} * \sqrt{\frac{1}{n} + \frac{1}{m}} \tag{5.1}$$

Hierbei beschreiben c den kritischen Wert für die Kolmogorov-Smirnov-Statistik, n die Anzahl der Rohdaten, α das Signifikanzniveau, mit dem der Test durchgeführt wird, und m die Anzahl reproduzierter Daten. Das Signifikanzniveau α hängt davon ab, welche Fehlerwahrscheinlichkeit toleriert wird. Die Anzahl der Rohdaten ergibt sich aus einer Monte-Carlo-Prozedur, hängt von der Reflektionswahrscheinlichkeit ab und kann nicht ohne größere Umstände variiert werden. Somit sind α und n feste Werte. Es bleibt m als einzige Veränderliche. Abhängig von m können zu einem festgelegten α und einem bestimmten Fall von Neutralgastransport, also zu einem festen n, Funktionen aufgestellt werden, die die kritischen Werte beschreiben. Für verschiedene $\alpha_i = 0.1\%, 1\%, 5\%, 10\%, i = 1, ..., 4$ und $n = 5000$ sehen die Funktionen aus wie in Abbildung 5.1 dargestellt.
Für den Vergleich berechnen wir nun Kolmogorov-Smirnov-Statistiken d (KS-d) zu unterschiedlich mächtigen Reproduktionen eines bestimmten Falls. Zur

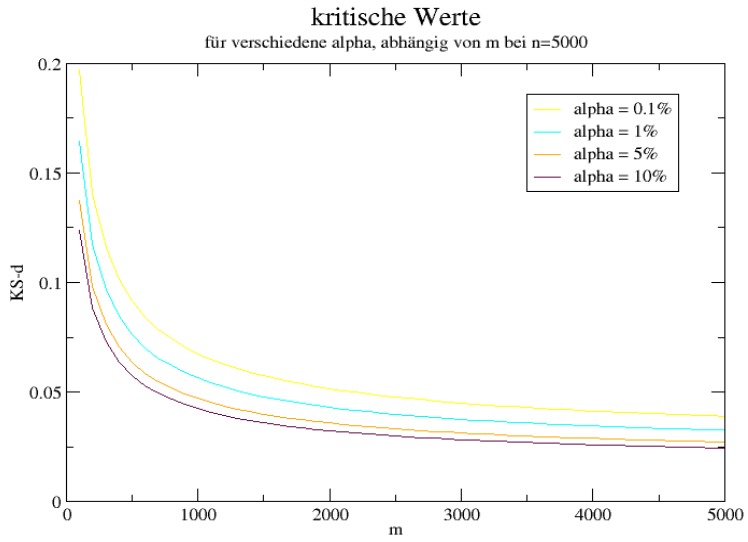


Abbildung 5.1: y-Achse: Kolmogorov-Smirnov d, x-Achse: m

Veranschaulichung nehmen wir den Einfall von Wasserstoff (H) auf Eisen (Fe) bei einem Winkel von 0° und einer Einfallsenergie von 1 eV. Die Rohdaten umfassen hierbei $n = 111467$ dreidimensionale Punkte. In Abbildung 5.2 sind die tatsächlich berechneten Kolmogorov-Smirnov-Statistiken d, sowohl für 5 als auch für 10 bins, sowie die kritischen Werte zu unterschiedlichen α dargestellt. Es handelt sich um die Reproduktion des Polarwinkels θ des oben genannten Falls. Zwischen den einzelnen Punkten (m, d) wurde linear interpoliert. Man erkennt, dass die KS-Distanzen der reproduzierten Verteilung mit 10 bins stets geringer sind als die der Verteilungen mit 5 bins. Das bedeutet gleichzeitig, dass die "5 bin"-Kurve die α-Kurven eher, also bei einem kleineren m, schneidet als die "10 bin"-Kurve. Die Schnittstelle m_s bedeutet, dass der Test bis zu dieser Größe, das heißt für alle $m < m_s$, nicht anspringt und die Verteilungen für gleich erklärt. Für alle $m \geq m_s$ sind ausreichend Daten vorhanden, sodass der Test feststellt, dass die Verteilungen ungleich sind. Für das Beispiel bedeutet dies, dass die "5 bin"-Kurve jeweils die Kurven

- $\alpha = 0.1\%$ bei $m_s \approx 2500$
- $\alpha = 1\%$ bei $m_s \approx 1600$
- $\alpha = 5\%$ bei $m_s \approx 1100$
- $\alpha = 10\%$ bei $m_s \approx 800$

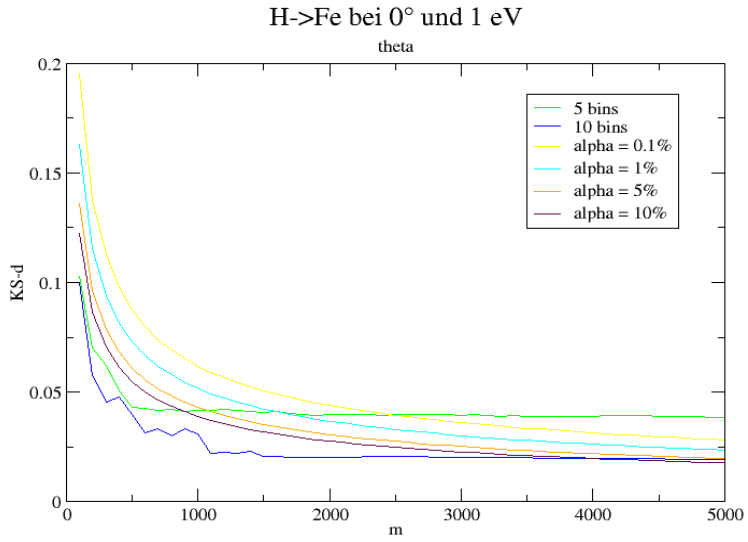


Abbildung 5.2: Vergleich kritische Werte - KS-d von “10 bin“- (blau) und “5 bin“-Verteilungsfunktionen(grün)

schneidet. Analog bedeutet dies für die “10 bin“-Kurve, dass diese jeweils die Kurven

- $\alpha = 0.1\%$ bei $m_s > 5000$
- $\alpha = 1\%$ bei $m_s > 5000$
- $\alpha = 5\%$ bei $m_s \approx 5000$
- $\alpha = 10\%$ bei $m_s \approx 4000$

schneidet. Da die “10 bin“-Kurve ab $m \approx 1500$ relativ konstant verläuft und die α-Kurven streng monoton fallen, sieht es so aus als würden sich die restlichen Kurven jenseits von $m = 5000$ schneiden. Abbildung 5.3 stellt den gleichen Fall nochmal dar, allerdings für $m = 5000, ..., 100000$. Hier wird die Vermutung bestätigt. Die “10 bin“-Kurve schneidet die “$\alpha = 0.1\%$“-Kurve bei ca. $m = 12000$ und die “$\alpha = 1\%$“-Kurve bei ca. $m = 9000$. Die Schnittpunkte sind allein durch Hinsehen also sehr grob geschätzt. Da für alle Funktionen gilt: $f : \mathbb{N}_{>0} \to \mathbb{R}$, muss auch gelten: $m \in \mathbb{N}_{>0}$. Den Schnittpunkt beziehungsweise den Punkt, der am nahesten Schnittpunkt liegt, könnte man demzufolge genau bestimmen, beispielsweise anhand des auf $\mathbb{N}$ beschränkten Bisektionsverfahrens.
Bei den Tests wurden alle Dimensionen jeweils separat getestet. Dass die Wahl des ersten Winkels beim Resamplen von der Wahl der Energie abhängt

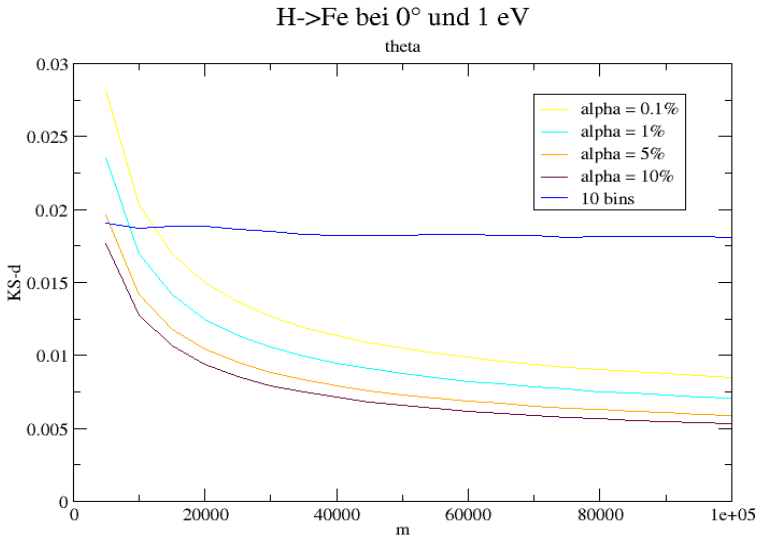


Abbildung 5.3: Vergleich kritische Werte mit KS-d von “10 bin“ - Verteilungsfunktion(blau)

und somit möglicherweise Abhängigkeiten bei den Testergebnissen bestehen könnten, wurde ausser Acht gelassen.

5.2 Variation der bin-Zahl

Der TRIM-Code hat bislang immer mit 5 bins gearbeitet. Es besteht allerdings die Möglichkeit diese Zahl zu variieren. Das kann je nach Fall Vor- oder Nachteile mit sich bringen. Der Vorteil liegt klar auf der Hand. Wie man an den meisten Grafiken erkennen kann, sind die “10 bin“-Kurven *besser*. Dies lässt sich darauf zurückführen, dass man mit Verteilungsfunktionen mit mehr bins mehr Informationen zur Verfügung hat und die Originalverteilungen somit präziser reproduzieren kann. Die Erhöhung der bin-Zahl kann sich aber auch negativ auswirken. Die Grafik 5.4 zeigt, dass die “10 bin“-Kurve *schlechter* ist, der Kolmogorov-Smirnov-Test also bei 10 bins früher anspringt. Grund dafür ist die geringe Reflektionswahrscheinlichkeit bei dem Fall D→C bei $\alpha 0 = 85°$ und $E0 = 1$ eV. Stellt man die Verteilungsfunktionen anhand der Rohdaten auf, hat man bei 10 bins entsprechend weniger Informationen in zweiter und dritter Dimension. Tabelle 5.1 zeigt die Anzahl Informationen je Dimension für $n = 20000$. Für wieviele bins man sich schließlich entscheidet, sollte also auch davon abhängen, wie hoch die Reflektionswahrscheinlichkeit ist beziehungsweise wieviele Rohdaten man zur

bins	1. Dimension	2. Dimension	3. Dimension
5	20000	4000	800
10	20000	2000	200

Tabelle 5.1: Anzahl Informationen in 1., 2. und 3. Dimension

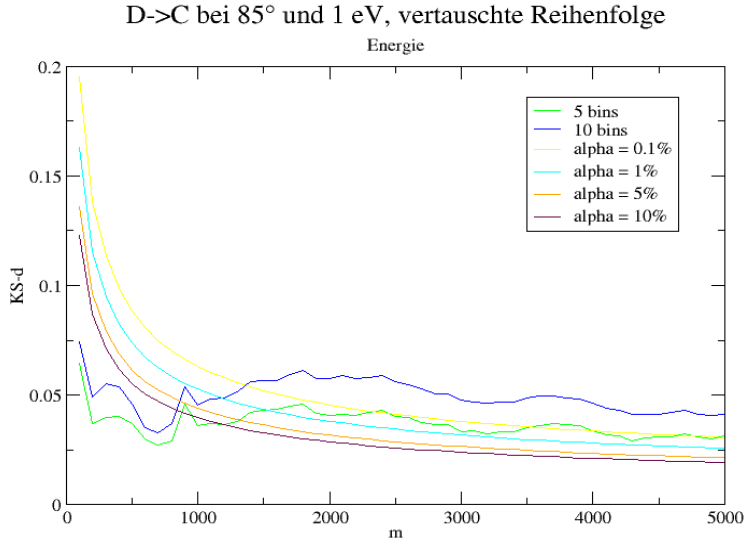


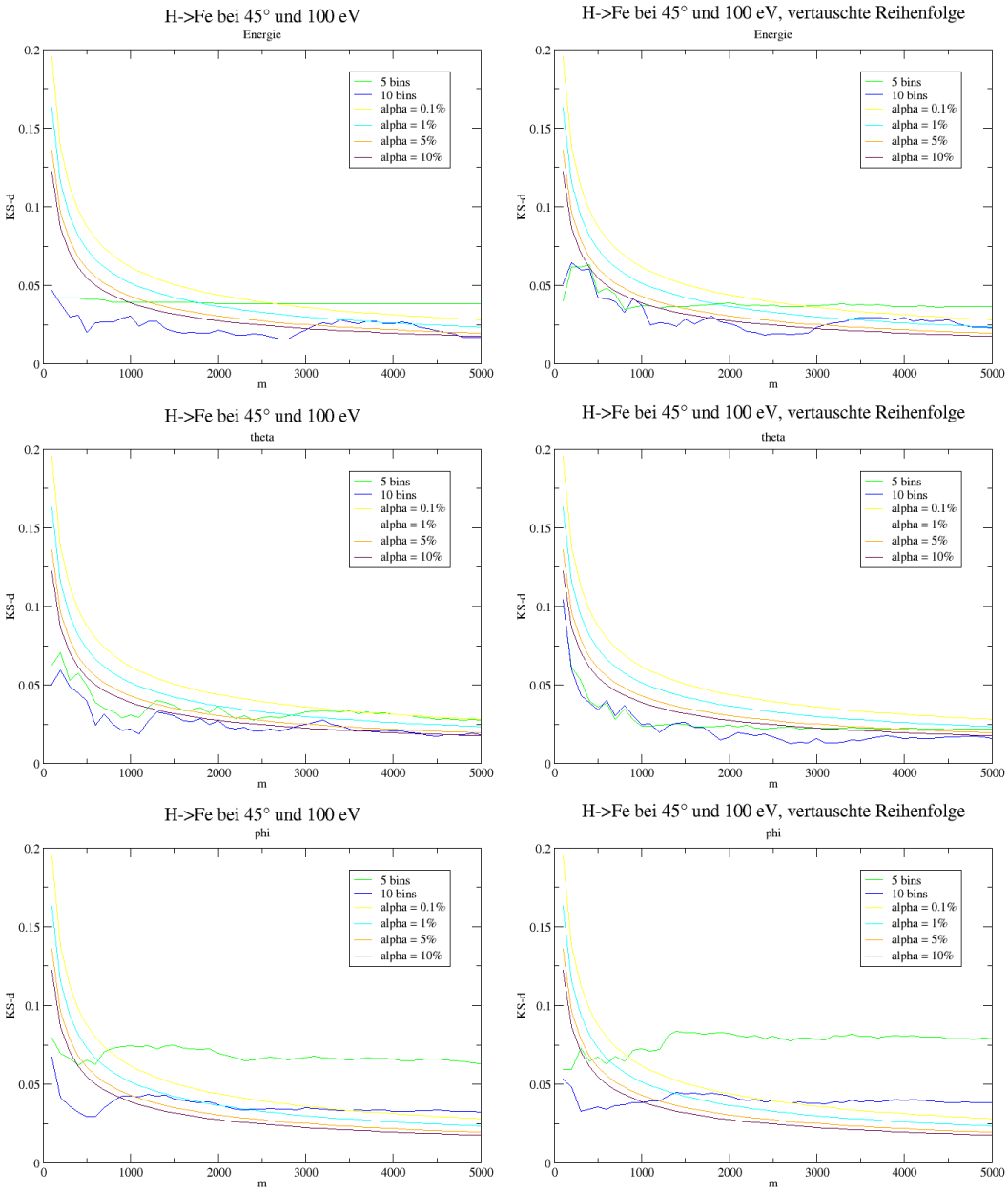
Abbildung 5.4: vertauschte Reihenfolge

Verfügung hat.

5.3 Variation der bin-Reihenfolge

Die Energie- und Raumwinkelverteilungen können durchaus sehr unterschiedlich sein. Bei einem Einfall eines Teilchens in einem Winkel von 0° gegen die Oberflächennormale ist der Reflektionswinkel um den Einfallspunkt gleichverteilt. Hingegen ist die Reflektionsrichtung bei einem schrägen Einfall, wie beispielsweise 85°, ziemlich eindeutig. Je nachdem welche Verteilungen vorliegen, kann es sinnvoll sein die Reihenfolge beim Aufstellen der bedingten Verteilungsfunktionen zu vertauschen. Dementsprechend muss die Reihenfolge beim *Resamplen* angepasst werden. Stichprobenartig wurden einige Fälle ausgewählt und in normaler sowie in vertauschter Reihenfolge ausgewertet. Vorgestellt wird hier der Einfall von Wasserstoff (H) auf Eisen (Fe) in einem 45°-Winkel und einer Einfallsenergie von 100 eV. An den Bildern auf der nächsten Seite erkennt man, dass die Unterschiede nicht groß sind. Auch

wenn die Kurven für kleine Energien teilweise etwas unterschiedlich wirken, ist der Trend doch recht ähnlich. Dass es gewinnbringend sein könnte, die Reihenfolge beim Aufstellen der bedingten Verteilungsfunktionen zu vertauschen, lässt sich also nicht ausschließen. Um zu erfahren, in welchen Fällen es tatsächlich Vorteile bringen könnte, müsste jeder Fall separat betrachtet werden.



Kapitel 6

Ausblick

Wie im Anschluss an Kapitel 3.3 und 3.4 erwähnt, sprengt die Behandlung des zwei- und dreidimensionalen Kolmogorov-Smirnov-Tests den Rahmen dieser Arbeit. Weitere Recherche über die Methode im zweidimensionalen Fall mit anschließender Implementierung der Testroutinen wäre eine weiterführende Aufgabe. Ist dieses Problem gelöst, müssen Überlegungen angestellt werden, ob und wie dieses Verfahren auf den dreidimensionalen Fall erweiterbar ist. Die Lösung dieses Problems würde ein in Kapitel 5.1 angesprochenes Problem mitlösen. Bei den mehrdimensionalen Kolmogorov-Smirnov-Tests würden Abhängigkeiten zwischen den Dimensionen mit beachtet werden.

In dieser Arbeit wurden für diverse m speziell die reproduzierten Energie- und Raumwinkelverteilungen aus bedingten Verteilungen mit 5 und 10 bins betrachtet. In Kapitel 5.2 wurde erklärt, dass die Wahl der bin-Zahl die Qualität der reproduzierten Verteilung beeinflussen kann. Mit dieser Information kann das in Kapitel 5.1 definierte Gütemaß um eine Dimension erweitert werden. Abhängig von der Menge der reproduzierten Daten m, was wir bereits haben, und der Wahl der bin-Zahl b, kann ein neues beziehungsweise. erweitertes Maß geschaffen werden. Das Maximum dieses Maßes stellt die optimale Kombination aus m und b: (m,b) dar.

Literaturverzeichnis

[L1] englische Wikipedia-Seite zu "Kolmogorov-Smirnov-Test", http://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

[L2] Artikel "Kolmogorov-Smirnov-Zwei-Sample-Test" von Uni-Konstanz, http://www.uni-konstanz.de/ZE/RZ/statistik/wnagl/downloads/nichtpara/np5.pdf

[L3] "Are Two Distributions Different?", Numerical Recipes, S. 614, Kap. 14.3

[L4] "Do Two-Dimensional Distributions Differ?", Numerical Recipes, S. 640, Kap. 14.7

[L5] Inversionsverfahren im Artikel "Techniken der stochastischen Simulation", http://www.fernuni-hagen.de/BWLQUAM/assets/errata/e00859-3_33ff.pdf

[L6] Inversionsmethode auf http://www.mathematik.uni-ulm.de/stochastik/lehre/ss03/markov/skript/node29.html

[L7] Wikipedia - Oktant (Geometrie), http://de.wikipedia.org/wiki/Oktant_%28Geometrie%29