

Datenanalyse in der Physik

Vorlesung 9 Erzeugung von Pseudo-Zufallszahlen und die Monte Carlo Methode

Prof. Dr. J. Mnich

DESY und Universität Hamburg



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 1

Zufallszahlengenerator

- Ein wichtiges, modernes Hilfsmittel in der Physik ist die Monte Carlo Methode

Sie wird eingesetzt zur numerischen Berechnung sehr komplexer Integrale insbesondere bei komplizierten Integrationsgrenzen

- Monte Carlo Methoden basieren auf numerisch generierten Pseudozufallszahlen

Zunächst diskutieren wir Algorithmen zur Erzeugung solcher Pseudozufallszahlen

- Gleichverteilte Zufallszahlen u_j im Intervall $[0, 1]$

Die einfachste Möglichkeit ist der linear kongruente Generator

$$u_j = I_j/m \quad \text{mit} \quad I_j = (a I_{j-1} + c) \mod m$$

Der rekursive Algorithmus benötigt 3 ganzzahlige Konstanten a, c, m und eine Saat I_0

- Die Folge wiederholt sich nach maximal m Zahlen
- Maximum der Periode wird erreicht durch geeignete Wahl von a, c, m



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 2

Einfaches Beispiel Pseudozufallszahlengenerator

```
/* Programm zur Berechnung von gleichverteilten Zufallszahlen nach dem
/* linearen kongruenten Generator (siehe Blobel & Lohrmann)

#include <math.h>
#include <stdlib.h>

/* Saat bzw. letzte Zufallszahl, mit Default vorbesetzt */
/* kann durch Saat(...) geaendert werden */
unsigned long LetzteSaat = 987654321;

/* Algorithmus des linearen kongruenten Generators */
/* Achtung: ohne Vorkehrung gegen Ueberschreitung des Wertebereichs */
double Zufall()
{
/* definiere Konstanten des Algorithmus nach (Blobel & Lohrmann) */
int m = 259200;
int a = 421;
int c = 54773;
unsigned long I;
I = (LetzteSaat*a+c)%m;
LetzteSaat = I;
return((double) I / (double) m);
}

void Saat(int I)
{
LetzteSaat = (unsigned long) I;
return;
}
```



Einfaches Beispiel Pseudozufallszahlengenerator

Die ersten 10 Zahlen des Algorithmus:

0.775841
0.840397
0.018611
0.046593
0.827122
0.429641
0.090262
0.211763
0.363588
0.281848

Häufigkeitsverteilung nach 10000 Zahlen

Bin	Häufigkeit
[0.0, 0.1]	1009
[0.1, 0.2]	1017
[0.2, 0.3]	996
[0.3, 0.4]	1006
[0.4, 0.5]	999
[0.5, 0.6]	1007
[0.6, 0.7]	985
[0.7, 0.8]	971
[0.8, 0.9]	992
[0.9, 1.0]	1018



Statistischer Test

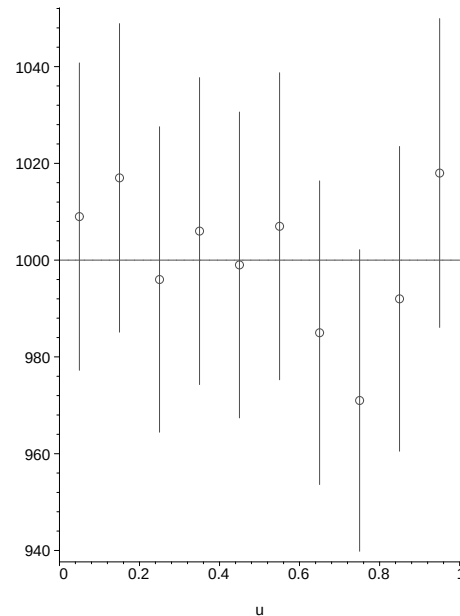
Diese Häufigkeitsverteilung ist hier gezeigt

Bei einer erwarteten mittleren Besetzungszahl von $\mu = 1000$ sollten die Einträge im Histogramm näherungsweise einer Gaußverteilung folgen

⇒ Standardabweichung (Fehlerbalken) ist $\sigma = \sqrt{\mu}$

Aber der χ^2 -Test ergibt $\chi^2/n = 1,9/9$ also viel zu klein

⇒ der Pseudozufallzahlgenerator ist schlecht!
Es gibt offensichtlich Korrelationen zwischen den Zahlen



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 5

Das gleiche Beispiel für drand48()

Die ersten 10 Zahlen des Algorithmus:

0.000000
0.000985
0.041631
0.176643
0.364602
0.091331
0.092298
0.487217
0.526750
0.454433

Häufigkeitsverteilung nach 10000 Zahlen

Bin	Häufigkeit
[0.0, 0.1]	1033
[0.1, 0.2]	995
[0.2, 0.3]	1003
[0.3, 0.4]	951
[0.4, 0.5]	990
[0.5, 0.6]	1038
[0.6, 0.7]	999
[0.7, 0.8]	971
[0.8, 0.9]	1051
[0.9, 1.0]	969

Der χ^2 -Test ergibt $\chi^2/n = 9,5/9$



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 6

Zufallszahlengenerator

Pseudozufallszahlen sollen bestimmte Eigenschaften haben

- große Periode
- gleichförmige Verteilung
- keine Korrelationen
- ...

Es existieren viele Tests für Zufallszahlengeneratoren

Einfache Algorithmen (wie der hier gezeigte!) sind nicht befriedigend

- benutzen Sie Generatoren aus mathematischen Bibliotheken
- seien sie aber sehr kritisch bei den Anwendungen



Universität Hamburg

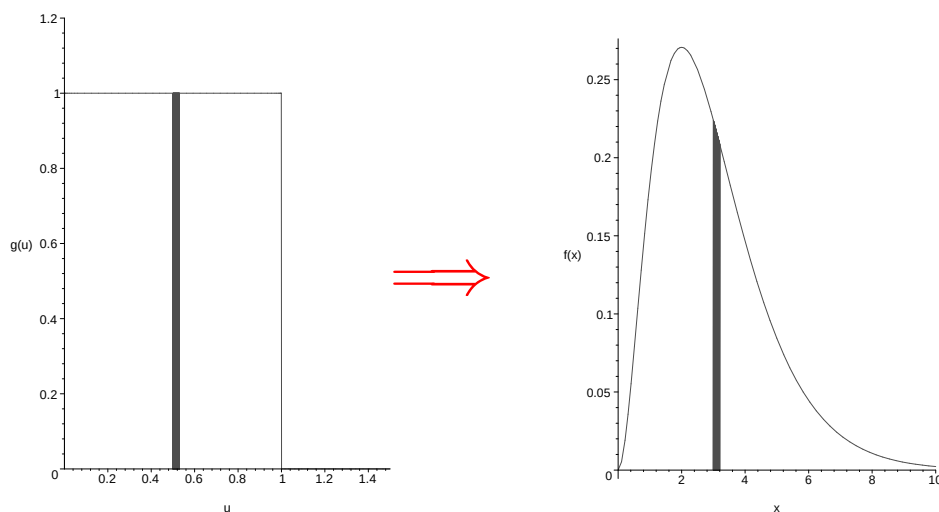
Datenanalyse in der Physik

Vorlesung 9 – p. 7

Zufallszahlen für beliebige Verteilungen

Aufgabe: Erzeugung von Zufallszahlen x_j für beliebige pdf $f(x)$

- Eine Möglichkeit ist, im Intervall $[0, 1]$ gleichverteilte Zufallszahlen zu transformieren:



Aus der Erhaltung der Wahrscheinlichkeit $g(u) du = f(x) dx$ folgt für die integrierte Wahrscheinlichkeitsdichte

$$F(x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^u g(v) dv = \int_0^u 1 dv = u$$



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 8

Zufallszahlen für beliebige Verteilungen

Die Gleichung $F(x) = u$ wird gelöst durch die Umkehrfunktion $x = F^{(-1)}(u)$

⇒ falls die Zufallszahlen u_j im Intervall $[0, 1]$ gleichverteilt sind, so folgen die Zufallszahlen

$$x_j = F^{(-1)}(u_j)$$

der Wahrscheinlichkeitsdichtefunktion $f(x)$

- Diese Methode ist anwendbar, falls die integrierte Wahrscheinlichkeitsdichte $F(x) = \int_{-\infty}^x f(t) dt$ und deren Umkehrfunktion $x = F^{(-1)}(u)$ analytisch berechnet werden können



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 9

Beispiel: exponentialverteilte Zufallszahlen

Wir möchten Zufallszahlen gemäß der Wahrscheinlichkeitsdichtefunktion

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

erzeugen

Berechnung der inversen, integrierten Wahrscheinlichkeitsdichte:

$$F(x) = \int_{-\infty}^x f(t) dt = \int_0^x \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t} \right]_0^x = -e^{-\lambda x} + 1 = u$$

$$\Rightarrow e^{-\lambda x} = 1 - u$$

$$x = -\frac{\ln(1 - u)}{\lambda}$$

Weil die Zufallszahl $1 - u$ genau wie u im Intervall $[0, 1]$ gleichverteilt ist, folgt auch die Zahlenreihe

$$x_j = -\frac{\ln u_j}{\lambda}$$

der Wahrscheinlichkeitsdichte $f(x; \lambda)$.



Universität Hamburg

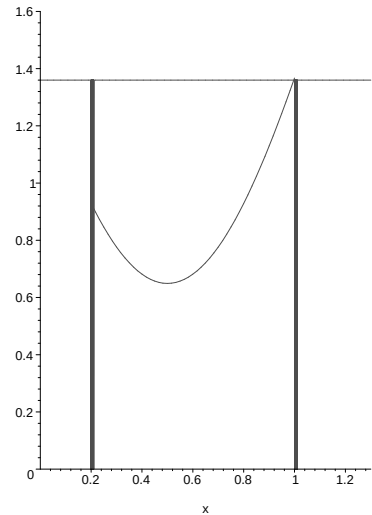
Datenanalyse in der Physik

Vorlesung 9 – p. 10

Zufallszahlen für beliebige Verteilungen

Wenn $F^{(-1)}(u)$ nicht berechnet werden kann, hilft das *Hit and Miss* Verfahren:

- Sei $f(x)$ eine pdf, die im Intervall $[a, b]$ definiert ist und das Maximum f_m hat
- Man erzeuge eine im Intervall $[a, b]$ gleichverteilte Zufallszahl x_j
- Diese Zufallszahl wird akzeptiert, falls eine zweite, im Intervall $[0, f_m]$ gleichverteilte Zufallszahl y_j , kleiner als $f(x_j)$ ist
- sonst wird x_j verworfen



- Man kann die Effizienz des Algorithmus dadurch erhöhen, dass man eine Funktion $f'(x) \geq f(x)$ findet, zu der man die umgekehrte integrierte Funktion $F'(u)$ bilden kann
- Die Zufallszahl x_j wird dann nach $f'(x)$ erzeugt und akzeptiert falls für die zweite Zufallszahl y_j gilt $y_j \leq f(x_j)/f'(x_j)$

- Für viele häufig gebrauchte pdf (z.B Gaußverteilung) existieren optimierte Algorithmen



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 11

Monte Carlo Methode

Monte Carlo Methoden sind wichtige Hilfsmittel der modernen Physik (und anderer Natur- und Ingenieurwissenschaften)

- Ihr Gebrauch wird durch Verfügbarkeit schneller Computer gefördert
- Im wesentlichen erleichtert bzw. ermöglicht die Monte Carlo (MC) Methode die Berechnung komplizierter Integrale bzw. Integrale mit komplizierten Grenzen
- Der Gebrauch von MC ist heute weitverbreitet

In der Hochenergiephysik werden hoch komplexe MC Simulationen eingesetzt um physikalische Prozesse, Detektoren und Experimente zu beschreiben



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 12

Monte Carlo Integration

Wir beginnen mit der MC Integration einer eindimensionalen Funktion $g(x)$:

$$I = \int_a^b g(x) dx$$

Dieses Integral kann ausgedrückt werden über den Erwartungswert $E[g(x)]$ einer gleichförmigen Wahrscheinlichkeitsdichtefunktion $f(x)$ im Intervall $[a, b]$

$$\begin{aligned} f(x) &= \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{sonst} \end{cases} \\ E[g(x)] &= \int_{-\infty}^{+\infty} f(x) \cdot g(x) dx = \frac{1}{b-a} \int_a^b 1 \cdot g(x) dx \\ \int_a^b g(x) dx &= (b-a) E[g(x)] \end{aligned}$$

Der Erwartungswert $E[g(x)]$ kann näherungsweise durch eine Summe berechnet werden mit n im Intervall $[a, b]$ gleichverteilten Zufallszahlen x_i

$$E[g(x)] \approx \frac{1}{n} \sum_{i=1}^n g(x_i)$$

⇒ das Integral I wird näherungsweise durch die MC Integration berechnet

$$I \approx I_{MC} = \frac{b-a}{n} \sum_{i=1}^n g(x_i)$$



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 13

Varianz der Monte Carlo Integration

Der Fehler der Näherung I_{MC} hängt von der Varianz der Verteilung der $g(x_i)$ ab

Zur Erinnerung:

$$V[x] = E[(x - \langle x \rangle)^2] = \int_{-\infty}^{+\infty} (x - \langle x \rangle)^2 f(x) dx = E[x^2] - \langle x \rangle^2$$

Varianz des MC Integrals:

$$V[I_{MC}] = V \left[\frac{b-a}{n} \sum_{i=1}^n g(x_i) \right] = \left(\frac{b-a}{n} \right)^2 V \left[\sum_{i=1}^n g(x_i) \right] \approx \frac{(b-a)^2}{n} V[g(x_i)]$$

Dabei ist im letzten Schritt der zentrale Grenzwertsatz $n \rightarrow \infty$ benutzt worden

Die Varianz der Näherung I_{MC} nimmt also mit $1/n$ ab,
bzw. der Fehler mit $1/\sqrt{n}$

Beispiel: Um den Fehler zu halbieren muss die Zahl n der berechneten Funktionswerte vervierfacht werden



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 14

Methoden zur Reduzierung der Varianz

Importance Sampling:

- Sei $f(x)$ eine im Intervall $[a, b]$ normierte Wahrscheinlichkeitsdichtefunktion. Dann lässt sich das Integral der Funktion $g(x)$ schreiben als Erwartungswert von $r(x) = g(x)/f(x)$ bezüglich der pdf $f(x)$:

$$I = \int_a^b g(x) dx = \int_a^b \left(\frac{g(x)}{f(x)} \right) f(x) dx = E \left[\frac{g(x)}{f(x)} \right] = E[r(x)]$$

- Dieser Erwartungswert lässt sich nun mit der MC Methode näherungsweise mit Zufallszahlen x_i berechnen, die nach $f(x)$ verteilt sind

$$I \approx I_{MC} = \frac{b-a}{n} \sum_{i=1}^n r(x_i)$$

- Das ist sinnvoll, falls man zu $f(x)$ die Umkehrung der Stammfunktion $F^{(-1)}$ analytisch angeben kann um die Zufallszahlen zu erzeugen
- Die Varianz von $r(x)$ sollte klein sein:

$$V[r(x_i)] = E[(r(x_i) - \langle r \rangle)^2]$$

Das wird erreicht für $r(x) \approx \text{const}$, also wähle man $f(x) \approx g(x)$ in $[a, b]$



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 15

Methoden zur Reduzierung der Varianz

- Die Verwendung einer Wahrscheinlichkeitsdichtefunktion $f(x)$ erlaubt auch die Behandlung von Singularitäten $g(x) \rightarrow \pm\infty$
- Falls die Integrationsgrenzen $\pm\infty$ einschließen, muss zunächst Variablentransformation durchgeführt werden
- Verwendung von Importance Sampling schwierig falls $f(x)$ im Intervall sehr kleine Werte annimmt und für mehrdimensionale Integrale

Kontrollfunktion:

- Genauigkeit der MC Integration kann auch durch eine analytisch integrierbare Funktion $f(x) \approx g(x)$ verbessert werden

$$I = \int_a^b g(x) dx = \int_a^b f(x) dx + \int_a^b [g(x) - f(x)] dx$$

Hier wird das linke Integral analytisch und das rechte (kleine!) durch MC berechnet



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 16

Methoden zur Reduzierung der Varianz

Partitionierung:

- Die Varianz kann reduziert werden durch Aufteilen des Integrationsintervalls in mehrere Teilintervalle

Im Grenzfall des Aufteilens in n gleiche Teilintervalle, d.h. eine Zufallszahl pro Teilintervall, erreicht man die Genauigkeit der numerischen Integration ($\rightarrow$ siehe unten)

Gewichtung:

- Falls man ein Integral über eine Funktion $g(x)$ mit MC berechnet hat, kann man die Integration über eine modifizierte Funktion $g'(x)$ auch wie folgt ausführen

$$I' = \int_a^b g'(x) dx \approx \frac{b-a}{n} \sum_{i=1}^n g'(x_i) = \frac{b-a}{n} \sum_{i=1}^n g(x_i) w_i$$

mit den Gewichten

$$w_i = \frac{g'(x_i)}{g(x_i)}$$



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 17

Mehrdimensionale Integrale

Die Stärke der MC Methode liegt in der Berechnung mehrdimensionaler Integrale über einen komplizierten Integrationsbereich $F(x_1, x_2, \dots, x_n)$

$$I = \int \int \dots \int_{F(x_1, x_2, \dots, x_n)} g(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

Sorgfalt ist geboten bei der Wahl der Verteilung der Zufallsvariablen!

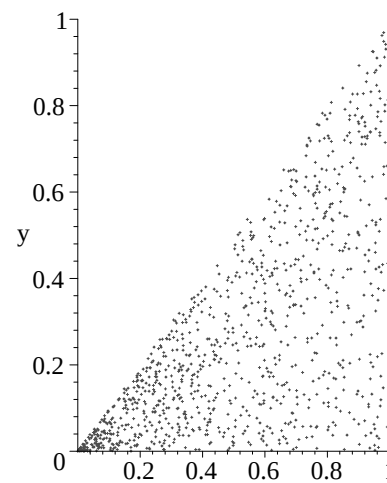
Wir betrachten ein Beispiel in zwei Dimensionen

$$I = \int_0^1 \int_0^x g(x, y) dx dy$$

Ein beliebter Fehler ist:

- würfel x_i gleichverteilt in $[0, 1]$
- würfel y_i gleichverteilt in $[0, x]$

ergibt keine Gleichverteilung in der (x, y) -Ebene



Universität Hamburg

Datenanalyse in der Physik

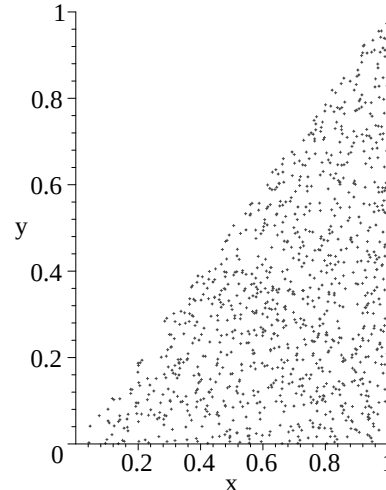
Vorlesung 9 – p. 18

Mehrdimensionale Integrale

Ein richtiger Algorithmus ist:

- würfel x_i gleichverteilt in $[0, 1]$
- würfel y_i gleichverteilt in $[0, 1]$
- und verwirfe das Paar falls $y_i > x_i$ ist

ergibt Gleichverteilung in der (x, y) -Ebene



Ein alternativer und effizienterer Algorithmus für dieses Problem ist:

- würfel u_i gleichverteilt in $[0, 1]$
- würfel v_i gleichverteilt in $[0, 1]$
- und setze $x_i = \max(u_i, v_i)$ und $y_i = \min(u_i, v_i)$



Universität Hamburg

Datenanalyse in der Physik

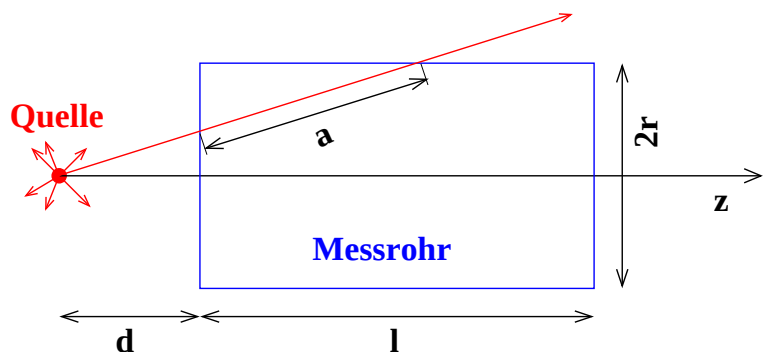
Vorlesung 9 – p. 19

Beispiel zur Monte Carlo Methode

Wir behandeln hier und in den Übungen ein einfaches physikalisches Beispiel zur Anwendung der Monte Carlo Methode

Gegeben sei eine gleichförmig in alle Raumrichtungen strahlende radioaktive Quelle

Ihre Aktivität, d.h. Zerfälle pro Zeiteinheit, soll mit einem zylindrischen Messrohr (z.B. Geiger-Müller-Zähler) gemessen werden. Wir nehmen an, die Quelle sei punktförmig und am Ursprung des Koordinatensystems $(x, y, z) = (0, 0, 0)$



- Das Messrohr überdeckt nur einen Teil des gesamten Raumwinkels $\Omega = 4\pi$. Außerdem soll die Nachweiswahrscheinlichkeit P des Messrohrs davon abhängen, wie groß der Weg der Strahlung a darin ist

$$P = P(a) = 1 - e^{-\lambda a} \quad (\text{das ist z.B. für } \gamma\text{-Strahler der Fall})$$

- Aufgabe: Man bestimme die totale Nachweiswahrscheinlichkeit, also welcher Bruchteil der ausgesendeten Strahlung nachgewiesen wird, und den statistischen Fehler.



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 20

Beispiel zur Monte Carlo Methode

Dieses Problem kann mit Hilfe einer kleinen Monte Carlo Simulation gelöst werden

- Man „erzeuge“ Strahlung, die gleichförmig auf einer Kugeloberfläche verteilt ist (Kugelkoordinaten!)

- Azimuthalwinkel ϕ : gleichverteilte Zufallszahlen in $[0, 2\pi]$

- Polarwinkel θ : nach $f(\theta) = \sin \theta$ verteilte Zufallszahlen in $0, \pi/2$

Das Raumwinkelement ist $d\Omega = d\phi d\cos \theta = \sin \theta d\phi d\theta$

Die Simulation des Halbraumes $\cos \theta < 0$ ist nicht notwendig

→ ergibt globalen Faktor $1/2$

- Man berechne für jedes Teilchen die Strecke im Messrohr

$$\begin{aligned}\tan \theta &\geq \frac{r}{d} &\Rightarrow a &= 0 \\ \frac{r}{d} > \tan \theta &\geq \frac{r}{d+l} &\Rightarrow a &= \frac{r}{\sin \theta} - \frac{d}{\cos \theta} \\ \tan \theta &< \frac{r}{d+l} &\Rightarrow a &= \frac{l}{\cos \theta}\end{aligned}$$

und würfele, ob das Teilchen registriert wird oder nicht

- Die gesamte Nachweiswahrscheinlichkeit ist dann die Zahl der registrierten geteilt durch die Zahl der erzeugten Teilchen



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 21

Komplexes Beispiel zur Monte Carlo Methode

Moderne Experimente der Hochenergiephysik bestehen aus sehr vielen einzelnen Detektoren

- L3 am LEP Beschleuniger (CERN) hatte u.a. etwa 11 000 Kristalle zur Energiemessung
- CMS am LHC Beschleuniger wird ca. 15 000 Silizium-Streifendetektoren enthalten mit etwa 10^7 einzelnen Kanälen

Zur Analyse der Daten werden sehr detaillierte MC Simulationen benötigt

- Simulation der physikalischen Reaktion:
alle entstehenden Teilchen und deren erwartete Energie-, Impuls- und Winkelverteilungen
- Nachweiswahrscheinlichkeit für jedes Detektorelement
- Orts- und Energieauflösung jeder einzelnen Detektorkomponente

Am Ende der Simulation stehen digitalisierte Signale der einzelnen Detektorkomponenten, die sich nicht von echten Daten unterscheiden

Der simulierte Datensatz dient dann zur Optimierung der Selektion und Bestimmung der Akzeptanz



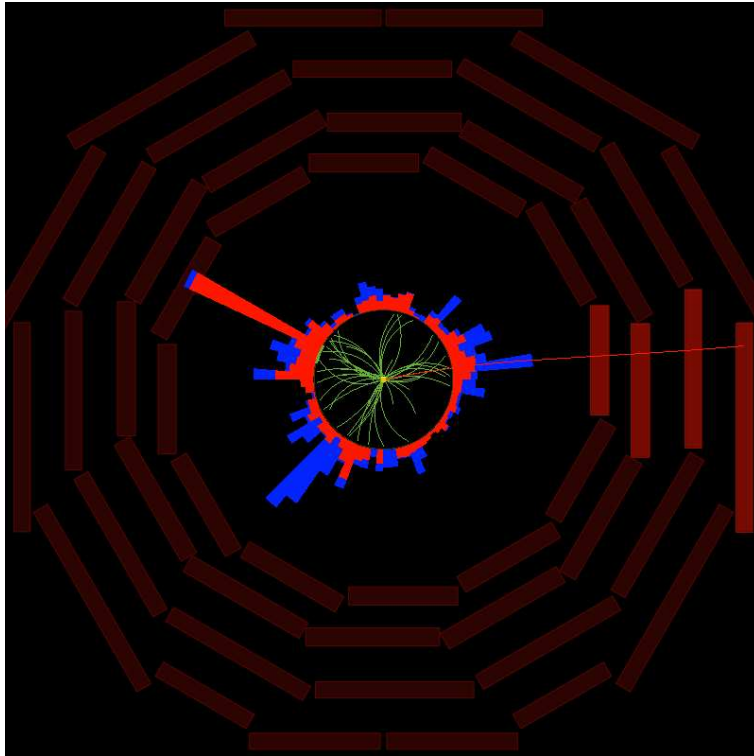
Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 22

Komplexes Beispiel zur Monte Carlo Methode

CMS Experiment am LHC Beschleuniger am CERN:
Simulation eines Top-Paar-Ereignisses $pp \rightarrow t\bar{t} + X$



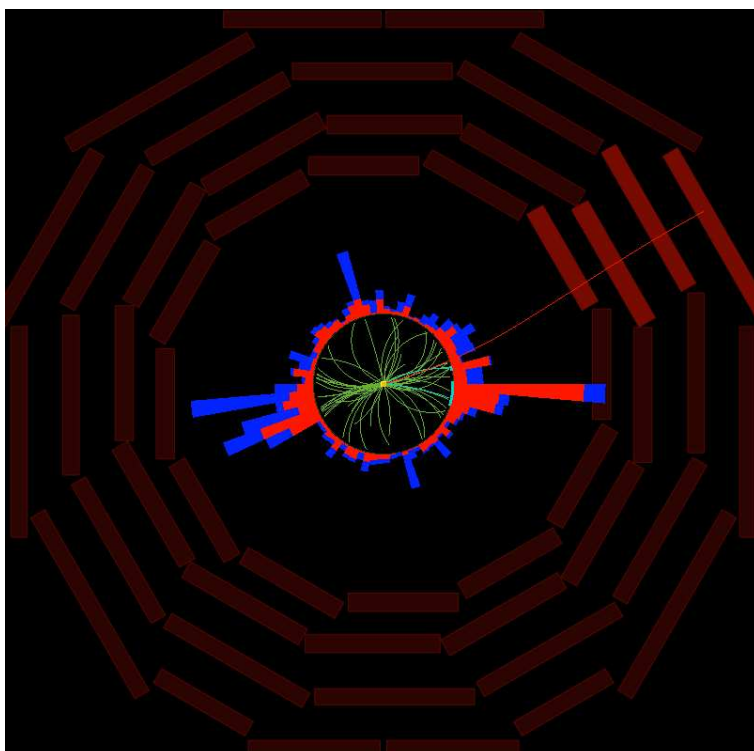
Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 23

Komplexes Beispiel zur Monte Carlo Methode

CMS Experiment am LHC Beschleuniger am CERN:
Im Vergleich zu einem realen Ereignisse in den Daten $pp \rightarrow t\bar{t} + X$



Universität Hamburg

Datenanalyse in der Physik

Vorlesung 9 – p. 24