Statistische Methoden der Datenanalyse

Wintersemester 2012/2013

Albert-Ludwigs-Universität Freiburg

Dr. Stan Lai und Prof. Markus Schumacher

Physikalisches Institut Westbau 2 OG Raum 008

Telefonnummer 07621 203 8408 (SL) / 7612 (MS)

E-Mail: Stan.Lai@physik.uni-freiburg.de

Markus.Schumacher@physik.uni-freiburg.de

http://terascale.physik.uni-freiburg.de/lehre/ws_1213/statmethoden_ws1213

Kapitel 4

Rechnererzeugte Zufallszahlen
Die Monte-Carlo-Methode

Die Monte-Carlo-(MC)-Methode

MC-Methode ist eine numerische Technik zur Bestimmung von

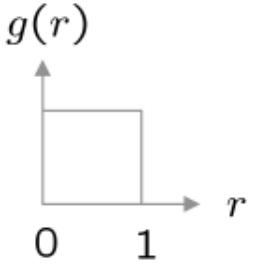
- Wahrscheinlichkeitsdichtefuntionen
- Transformation von Zufallsvariablen
- Bestimmung von Integralen
- Erwartunsgwerte
- Faltungen

mit Hilfe von Zufallszahlen

Anwendungen: - Generierung von Ereignissen/Messungen gemäß eines theoretischen Modells
- Simulation des Ansprechverhaltens eines Nachweisapparates/ einer Messapratur
- ....

Die Monte-Carlo-(MC)-Methode (2)

Die einzelnen Schritte sind:



(1) Generiere eine Sequenz von Zufallszahlen
$r_1, r_2, ..., r_m$ gemäß Gleichverteilung im Intervall [0, 1].

(2) Verwende diese um eine weitere Sequenz $x_1, x_2, ..., x_n$
zu erzeugen, die gemäß einer vorgegebenen WDF $f(x)$
verteilt sind. (x kann Vektor sein).

(3) Verwende x_i Werte um Eigenschaften von $f(x)$ zu bestimmen,
z.B. Erwartungswerte oder
Anteil x Werten mit $a < x < b$ ergibt

→ MC-Berechung = Integration (zumindest formal) $\int_a^b f(x)\,dx$.

MC generierte Werte = "simulierte Daten"

→ oft verwendet um Gültigkeit statistischer Methoden zu testen

Besondern nützlich bei:- vieldimensionalen f(x)
- komplizierten Unterräumen des "x"^{n}

Zufallszahlengenerator

Ziel: Erzeugung gleichmäßig verteilter Werte im Intervall [0, 1].

Werfe Münze für z.B. 32 Bitnummer... (too tiring).

→ 'Zufallszahlengenerator'

= Computeralgorithmus zur Erzeugung von $r_1, r_2, ..., r_n$.

Beispiel: Multiplikativer linear kongruenter Generator (MLCG)

$n_{i+1} = (a\, n_i) \bmod m$, wobei

n_i = ganze Zahl

a = Multiplikator (ganze Zahl)

m = Modulus (ganze Zahl)

n_0 = "Seed" ="Saatzahl" (Startwert)

Bemerkung: mod = Modulus z.B. 27 mod 5 = 2.

Diese Regel erzeugt eine Sequenz von Zahlen $n_0, n_1, ...$

Zufallszahlengenerator (2)

Die Sequenz ist (unglücklicherweise) periodisch!

Beispiel: (siehe Brandt Kapitel 4): $a = 3$, $m = 7$, $n_0 = 1$

$$
\begin{aligned}
n_1 &= (3 \cdot 1) \mathrm{mod} 7 = 3 \\
n_2 &= (3 \cdot 3) \mathrm{mod} 7 = 2 \\
n_3 &= (3 \cdot 2) \mathrm{mod} 7 = 6 \\
n_4 &= (3 \cdot 6) \mathrm{mod} 7 = 4 \\
n_5 &= (3 \cdot 4) \mathrm{mod} 7 = 5 \\
n_6 &= (3 \cdot 5) \mathrm{mod} 7 = 1
\end{aligned}
$$

← Folge wiederholt sich

Wähle *a*, *m* so, dass eine lang Periode (Maximum = *m* - 1) erreicht wird; *m* normalerweise nahe an der größten “Integer”-Zahl, die auf dem Computer repräsentiert werden kann.

Verwende nur ein Untermenge der einzelnen Periode der Sequenz.

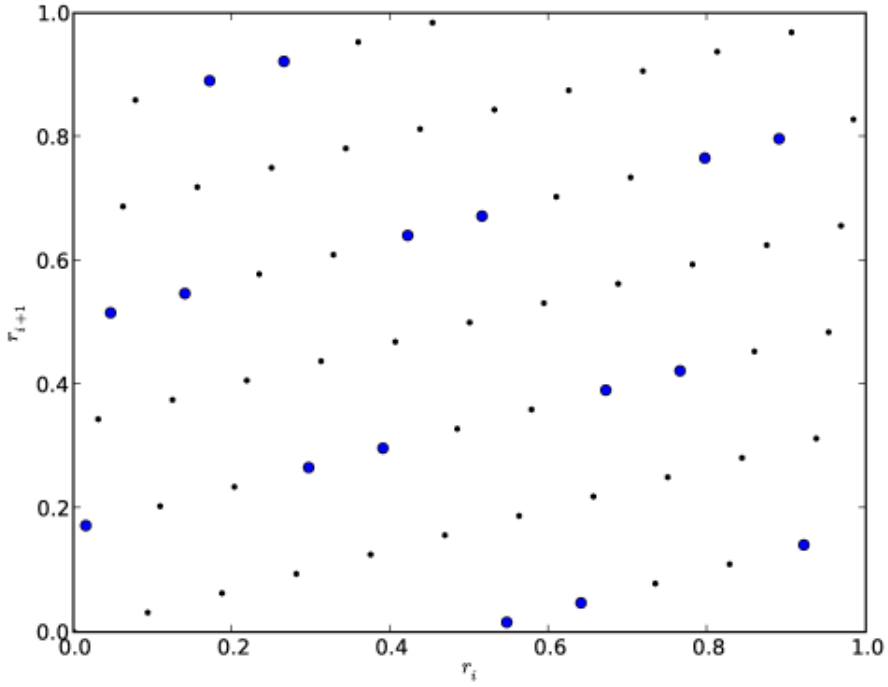
Zufallszahlengenerator (3)

$r_i = n_i/m$ sind [0, 1] aber sind sie “zufällig”?

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

Schlechtes Beispiel

a = 11
b = 0
m = 64



Liefert 2 Sequenzen der Periodenlänge 16

Für gerade Startwerte sind Perioden noch kürzer.

Zufallszahlengenerator (3)

$r_i = n_i/m$ sind in [0, 1] aber sind sie “zufällig”?

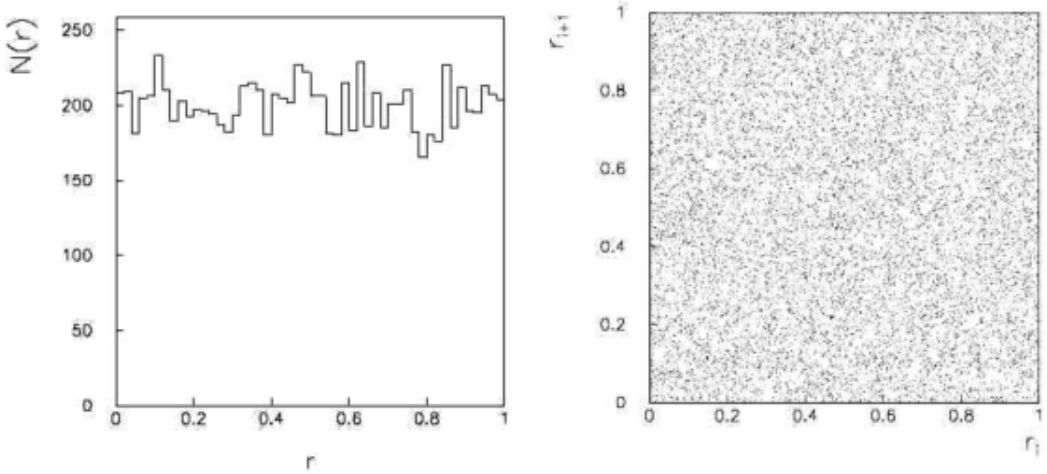
Wähle *a*, *m* so, dass die r_i eine Reihe von Tests für Zufälligkeit erfüllen:

gleichförmig verteilt in [0, 1],

alle Werte unabhängig (keine Korrelationen zwischen Paaren),

z.B. L’Ecuyer, Commun. ACM **31** (1988) 742 schlägt

a = 40692
m = 2147483399



Es gibt weit bessere Generatoren, z.B. **TRandom3**, basierend auf “Mersenne twister”-Algorithmus, Periode = $2^{19937} - 1$.

Siehe F. James, Comp. Phys. Comm. 60 (1990) 111; Brandt Ch. 4

Die Transformationsmethode

Anwendung der Transformationsmethode für Zufallsvariablen

bisher:	$f(x)$	$a(x)$	→	$g(a)$
	WDF für x gegeben	Funktion gegeben		WDF für a gesucht
jetzt:	$g(r)$	$x(r)$	→	$f(x)$
	Gleichverteilung in r gegeben	Transformation gesucht		WDF für x gegeben

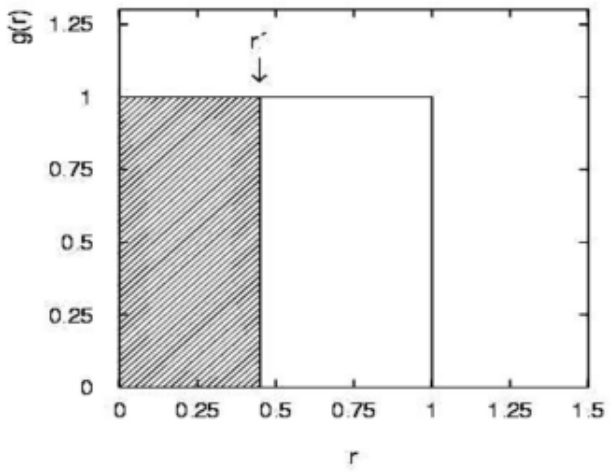
Ziel: gegeben Sequenz $r_1, r_2, \ldots, r_n$ gleichförmig in [0, 1],
erzeuge $x_1, x_2, \ldots, x_n$ die $f(x)$ folgen durch Auffinden einer geigneten Transformation $x(r)$.

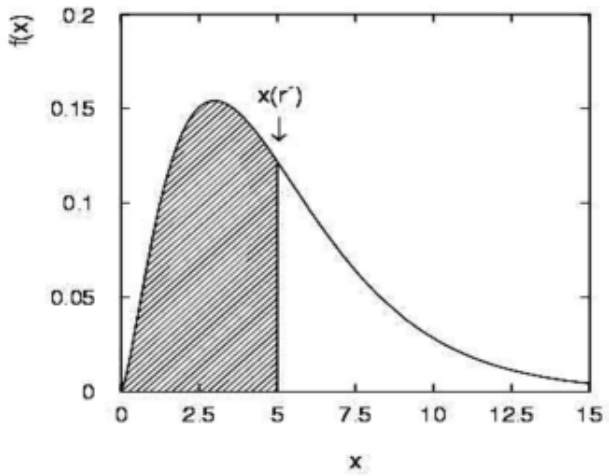
Die Transformationsmethode (2)

Verlange: Wkt, dass r in [r,r+dr] = g(r)dr= dr
= Wkt., dass x in [x(r), x(r)+dx(r)] = f(x) dx

Oder äquivalent: $P(r \leq r') = P(x \leq x(r'))$

$$\int_{-\infty}^{r'} g(r)\,dr = r' = \int_{-\infty}^{x(r')} f(x')\,dx' = F(x(r'))$$

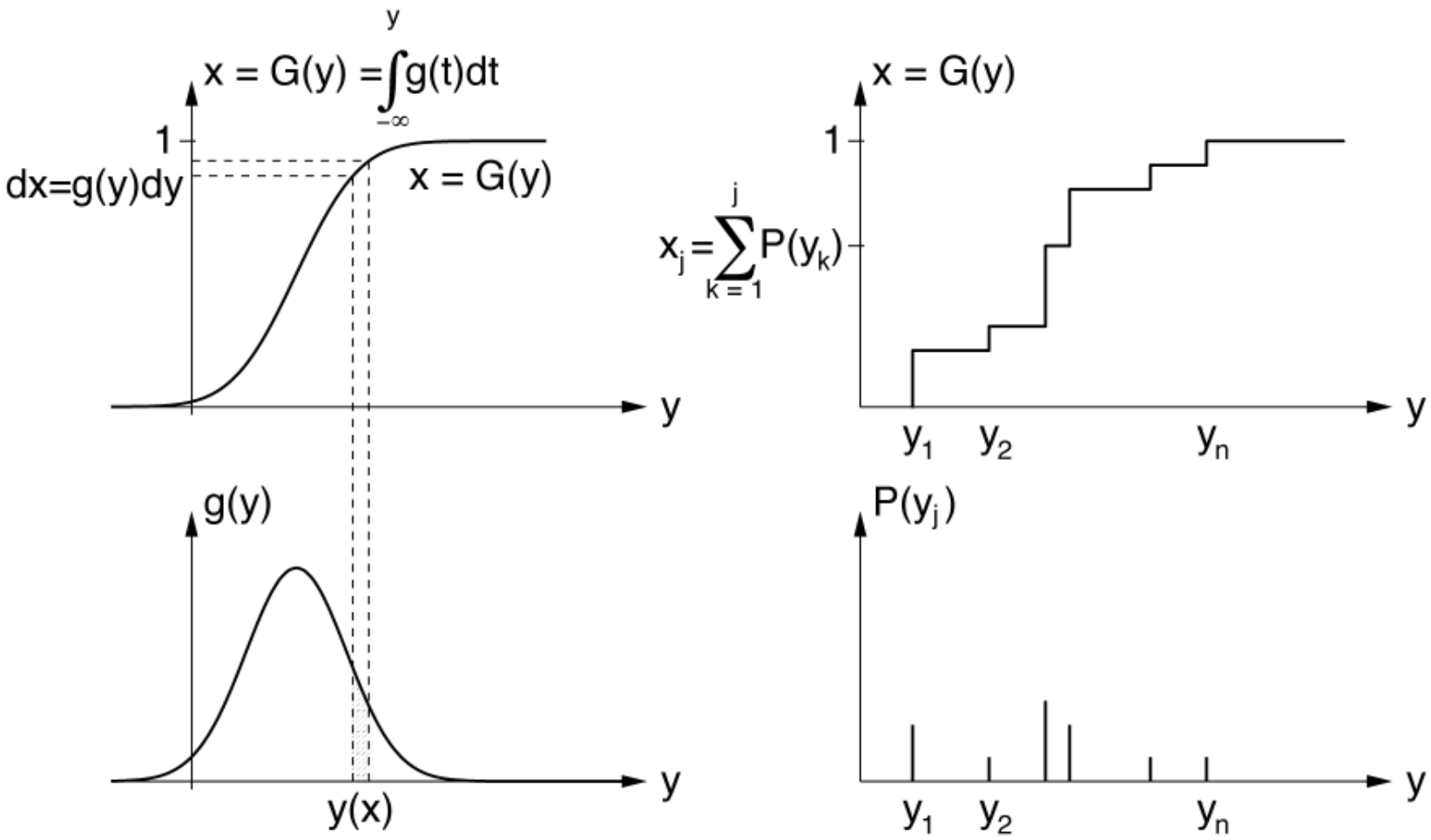




Da g(r)=1 gilt: $F(x)= r$ d.h. $x = F^{-1}(r)$

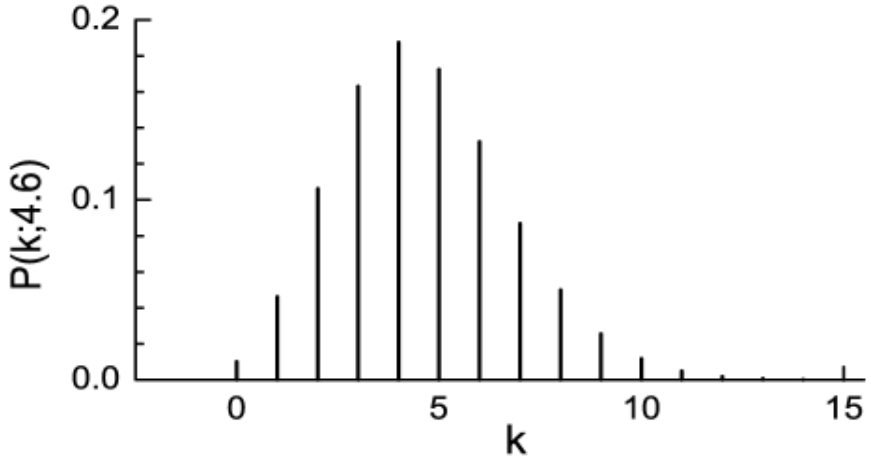
Benötigt: Kumulativfunktion analytisch bestimmbar und invertierbar.

Trafomethode für kontinuierliche u. diskrete ZV



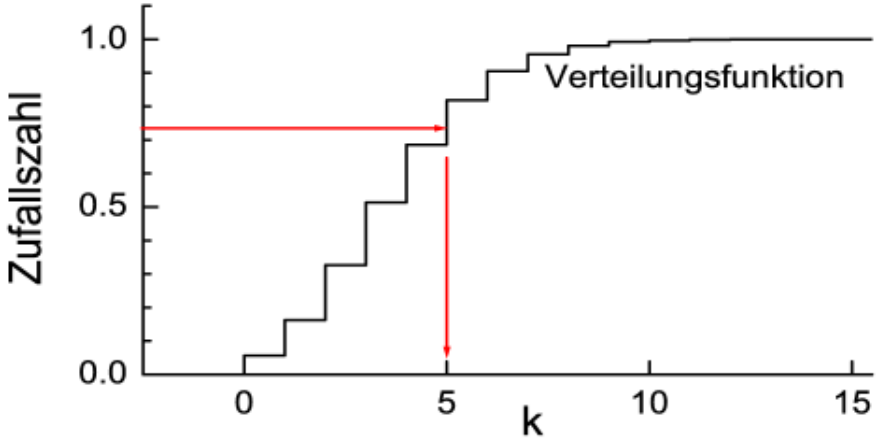
Transformatsmethode für Histogramme

Ordne r die Zahl n zu,
die dem kleinsten S(k) entspricht,
welches S>r erfüllt.
S(k) ist Kumulativverteilung
für diskrete Zufallsvariable.



Kann für empirische Verteilung
in Form eines Histogrammes
angewendet werden.

Rest r-S(j-1) wird für lineare
Interpolation verwendet.



Die Transformationsmethode (3)

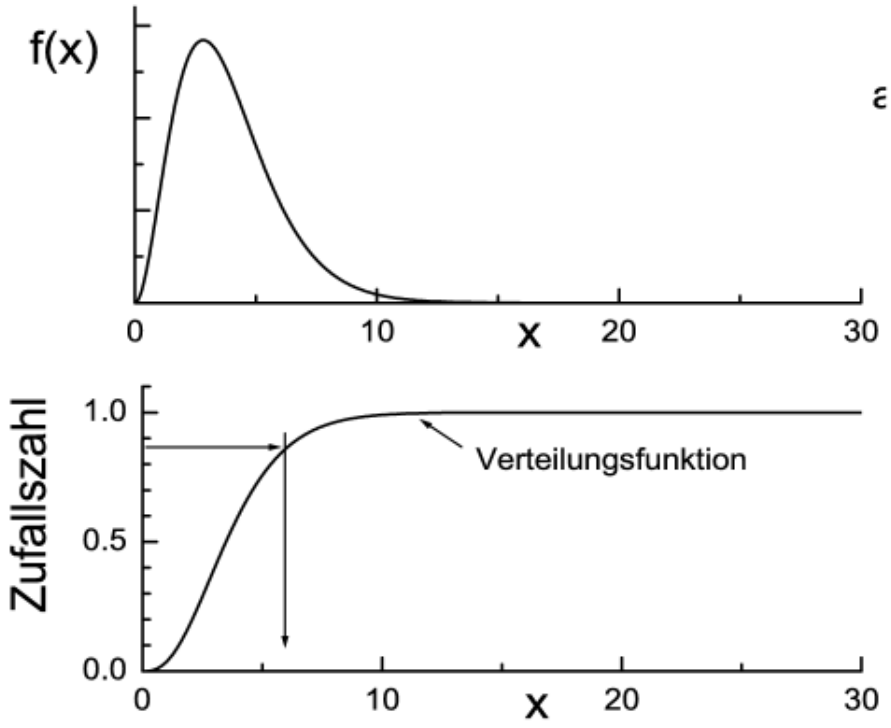
Aus r gleichförmig in [0,1] erzeuge x, die WDF f(x) folgt, gemäß:

$$f(x)\mathrm{d}x = u(r)\mathrm{d}r,$$

$$\int_{-\infty}^{x} f(x')\mathrm{d}x' = \int_{0}^{r(x)} u(r')\mathrm{d}r' = r(x)$$

$$F(x) = r,$$

$$x(r) = F^{-1}(r)\,.$$



Voraussetzung: Kumulativverteilung analytisch integrierbar u. invertierbar

Effizienz des Verfahrens 100% (jedes r erzeugt ein x)

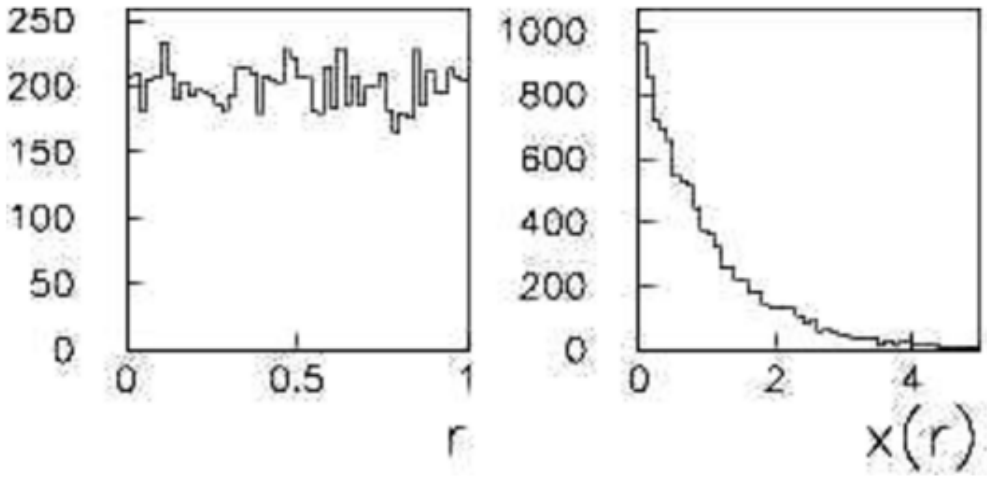
Beispiel für die Transformationsmethode

Exponential-WDF: $f(x;\xi) = \frac{1}{\xi} e^{-x/\xi} \quad (x \geq 0)$

Setze $\int_0^x \frac{1}{\xi} e^{-x'/\xi}\, dx' = r$ und löse nach x (r) auf.

$$-e^{(-x/\xi)} + 1 = r$$

→ $x(r) = -\xi \ln(1-r)$ ($x(r) = -\xi \ln r$ geht auch)



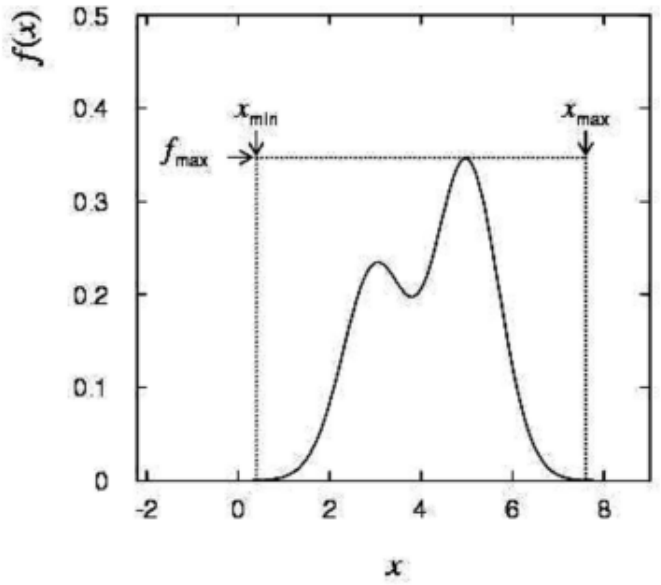
Vorteil: 100% Effizienz, d.h. aus jedem r_i wird ein x_i erzeugt.

Weitere Beispiele für die Transformationsmethode

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

Die von-Neumannsche-Zurückweisungsmethode

Schliesse WDF in Box ein



(1) Generiere Zufallszahl x, gleichförmig in $[x_{min}, x_{max}]$, i.e.

$$x = x_{\text{min}} + r_1(x_{\text{max}} - x_{\text{min}}) \ , \ r_1 \text{ ist gleichverteilt in [0,1].}$$

(2) Generiere eine 2te unabhängige Zufallszahl u gleichverteilt zwischen 0 und f_{max}, i.e. $u = r_2 f_{\text{max}}$.

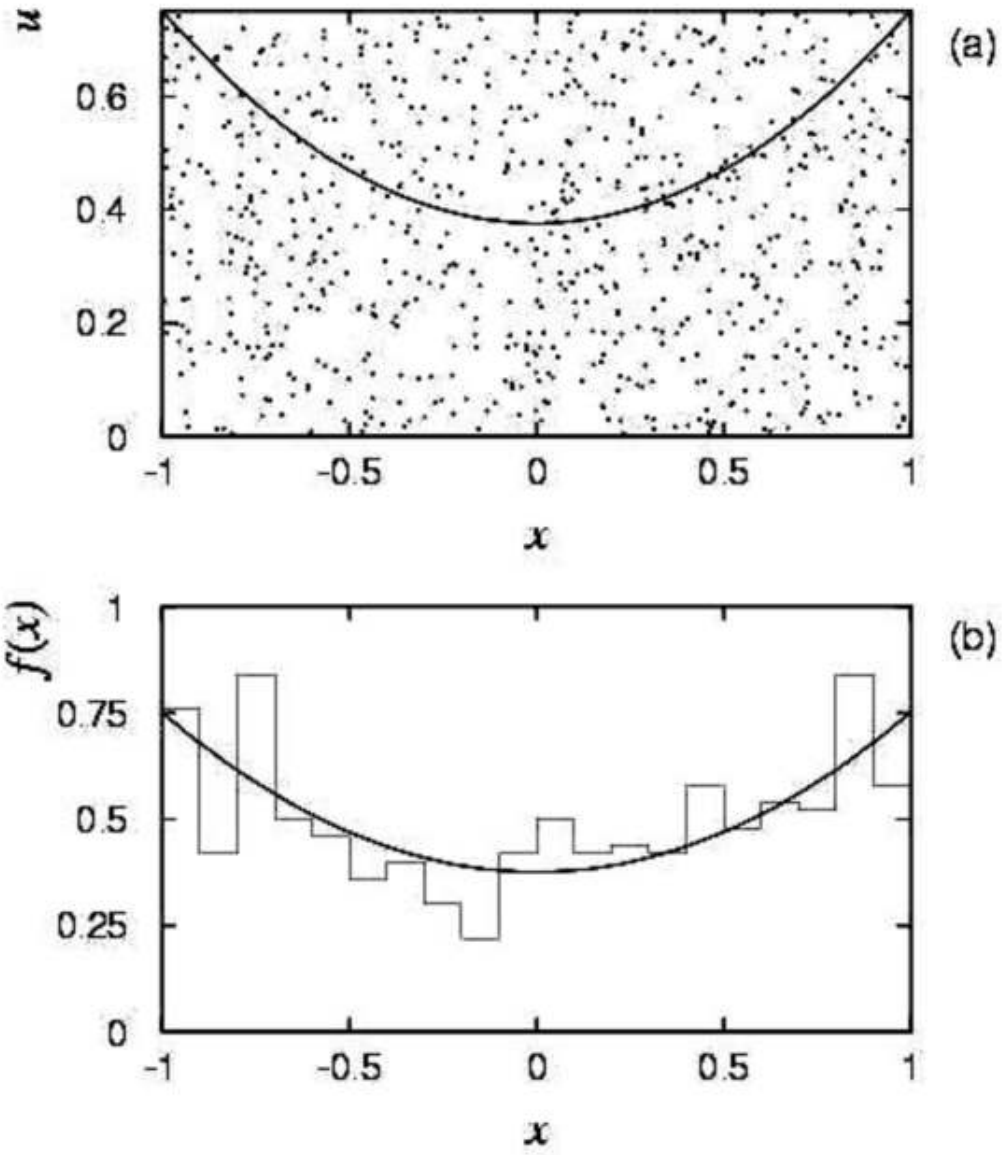
(3) Wenn $u < f(x)$, dann akzeptiere x. Wenn nicht, verwerfe x and versuche es erneut.

Die von-Neumannsche-Zurückweisungsmethode

$$f(x) = \frac{3}{8}(1 + x^2)$$

$$(-1 \leq x \leq 1)$$

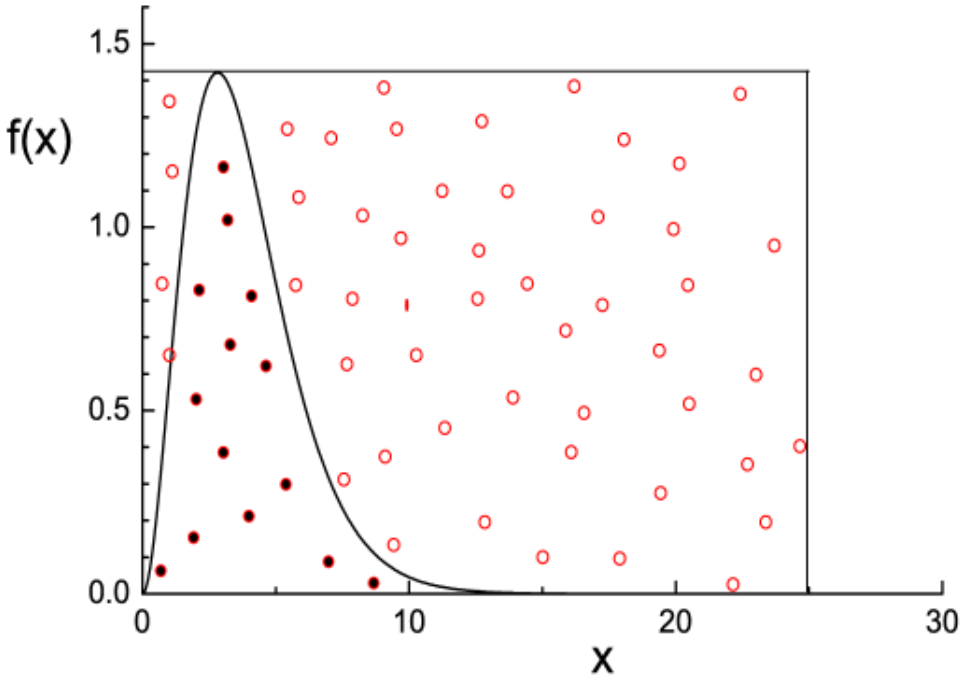
Wenn Punkt unterhalb der Kurve, dann behalte ihn und Fülle *x*-Wert in Histogramm.



Effiizienzerhöhung für Zurückweisungmethode

Einfaches Beispiel:
Plankspektrum

$$f(x) = c\frac{x^3}{e^x - 1}$$



Effizienz: hier ca. 10%
Weiteres Problem: Wertebereich x geht bis unendlich
Box wird nur bis x_{max} gewürfelt.

Majorantenmethode (importance sampling)

Suche geignete Funktion (Majorante) $m \geq f$ für alle x

Mit invertierbarer Stammfuktion M(x) $M(x) = \int_{-\infty}^{x} m(x')\mathrm{d}x'$

Zufallszahlen gemäß m(x) werden über erzeugt $x = M^{-1}(r)$

Erzeuge zweite Zufallszahl in Abhängigkeit von x im Bereich: null und $m(x)$

Behalte (verwerfe) diese wenn sie ≤ (>) f(x) ist

Bruchteil $[m(x) - f(x)]/f(x)$

der zweiten Zufallszahlen wird lokal verworfen

Besonders gut, wenn Majorante m(x) nahe an f(x) → große Effizienz
Vorteil: Generierung von Verteilungen die bis "unendlich gehen"

Majorantenmethode: Beispiel

Zielfunktion: $f(x) = c(\mathrm{e}^{-0.2x} \sin^2 x) \quad \text{für} \quad 0 < x < \infty$

Geeignete Majorante: $m(x) = c\,\mathrm{e}^{-0.2x}$.

Bedingung für Kumulativfkt.:
$$r = \int_0^x \frac{1}{0.2} \mathrm{e}^{-0.2x'} \mathrm{d}x'$$
$$= 1 - \mathrm{e}^{-0.2x} .$$

Transformation von gleichverteilten r_1 auf x gemäß Majorante m(x) $\quad x = -\frac{1}{0.2} \ln(1 - r_1)$

Würfele zweite Zufallszahl r_2 und bilde Produkt: $r_2\, m(x)$

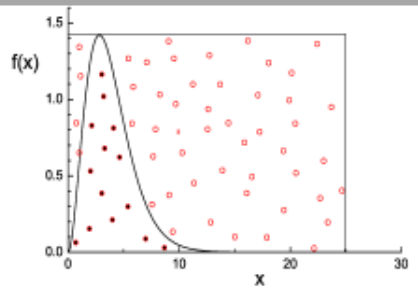
Akzeptanz/Zurückweisung gemäß:

für $r_2 < \sin^2 x \quad \rightarrow$ behalte x,

für $r_2 > \sin^2 x \quad \rightarrow$ verwerfe x

Majorantenmethode: Beispiel 2

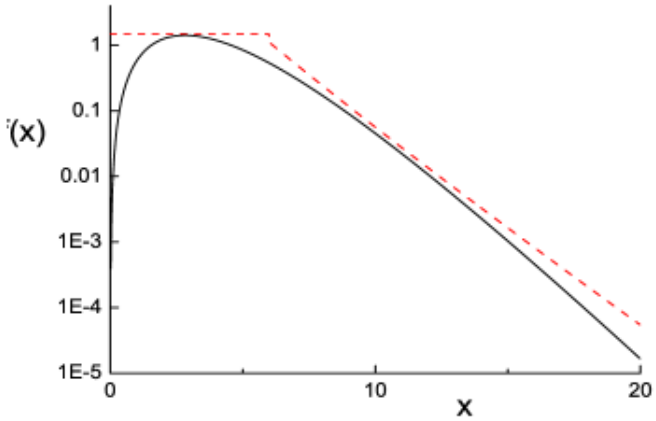
Zielfunktion: $$f(x) = c\frac{x^3}{\mathrm{e}^x - 1}$$



Stückweise Majorante: $x < x_1 \qquad m_1(x) = 6\,c$

$X^{-0.1}$ ermöglicht analytische Integration $\qquad x > x_1$

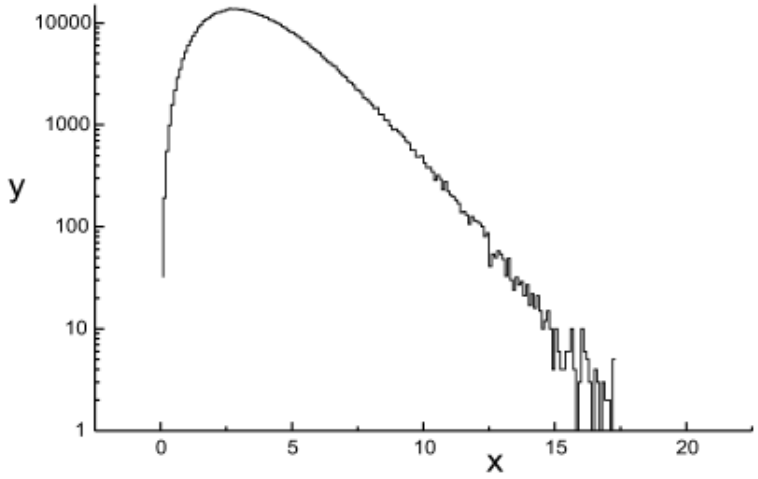
$$m_2(x) = 200\,c\,x^{-0.1}\mathrm{e}^{-x^{0.9}}$$



Für $x > x_1$:

$$M_2(x) = \int_{x_1}^{x} m_2(x')\mathrm{d}x',$$

$$= \frac{200c}{0.9}\left[\mathrm{e}^{-x_1^{0.9}} - \mathrm{e}^{-x^{0.9}}\right]$$



Behandlung additiver WDFs

Oft ist WDF Summe von mehreren Termen: $f(x) = f_1(x) + f_2(x)$

mit $S_1 = \int_{-\infty}^{\infty} f_1(x)\mathrm{d}x, \quad S_2 = \int_{-\infty}^{\infty} f_2(x)\mathrm{d}x \quad S_1 + S_2 = 1$

Wähle mit Wkt S_1 bzw S_2 ein Zufallszahl die nach $f_1(x)$ bzw. $f_2(x)$ verteilt ist.

Falls Stammfunktionen $F_1(x) = \int_{-\infty}^{x} f_1(x')\mathrm{d}x' \quad F_2(x) = \int_{-\infty}^{x} f_2(x')\mathrm{d}x'$ invertierbar

Dann generiere x aus gleichverteilten r gemäß

$x = F_1^{-1}(r)$ für $r < S_1$ $\qquad$ $x = F_2^{-1}(r - S_1)$ für $r > S_1$

Behandlung additiver WDFs: Beispiel

Ziel-WDF: $$f(x) = \varepsilon \frac{\lambda \mathrm{e}^{-\lambda x}}{1 - \mathrm{e}^{-\lambda a}} + (1 - \varepsilon)\frac{1}{a} \quad \text{für} \quad 0 < x < a$$

Bestimme Stammfunktion und deren Inverse für beide Summanden.

Transformiere gleichverteilte Zufallszahl r gemäß:

$$r < \varepsilon \qquad x = \frac{-1}{\lambda} \ln\left(1 - \frac{1 - \mathrm{e}^{-\lambda a}}{\varepsilon} r\right)$$

$$r > \varepsilon \qquad x = a \frac{r - \varepsilon}{1 - \varepsilon}$$

Methode liefert 100% Effizienz.
Ohne Separation wäre Inverse hier nicht analytisch berechenbar gewesen.

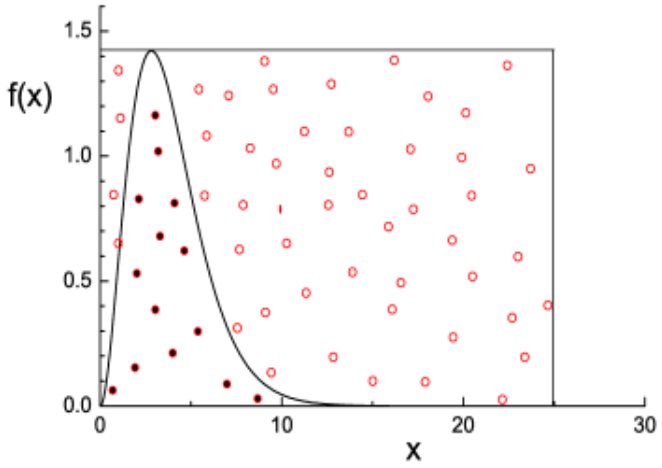
Seperation auch meist sinnvoll, wenn Terme nicht analytisch invertierbar.

MC-Integration

Integrale bei der Integrand das Vorzeichen wechselt werden in Bereiche zwischen Nullstellen zerlegt. Dann gilt o.B.d.A:

$$I = \int_{x_a}^{x_b} y(x)\,\mathrm{d}x \quad \text{mit} \quad y > 0$$

Methode 1: Primitive Zurückweisungsmethode $\quad \widehat{I} = I_0 \frac{N}{N_0}$



N: Anzahl der Erfolge (akzeptierten Zufallszahlen)

N_0: Anzahl aller Zufallszahlen

I_0: Integral der constanten = Fläche der Box

Unsicherheit aus Binomialverteilung

$$\varepsilon = N/N_0 \qquad \delta N = \sqrt{N_0 \varepsilon (1-\varepsilon)}\,,$$

$$\frac{\delta I}{I} = \frac{\delta N}{N} = \sqrt{\frac{1-\varepsilon}{N}}$$

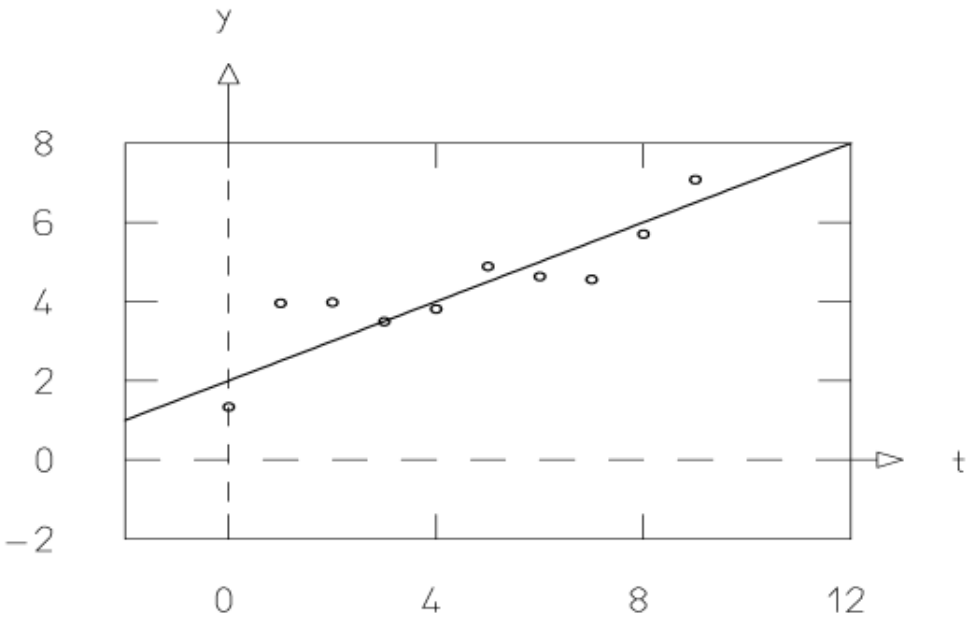
Simulation von Messungen

Betrachte naturwissenschaftliches Gesetz $y = at + b$

Messungen werden Stützstellen durchgeführt $t_0\ ,\ t_1 = t_0 + \Delta t\ ,\ t_2 = t_0 + 2\Delta t\ ,.$

Perfekte Messgenauigkeit liefert $y_i = at_i + b\ ,\ i = 0,1,\ldots,n-1$

Messung fehlerbehaftet aus Gauss-WDF $y_i' = y_i + \varepsilon_i$



Simulation von Messungen

Zerfallszeiten eines radiaktven Präparats

$$f(x) = \frac{1}{\tau}\exp(-x/\tau)\,,\quad x > 0\,,$$

Endliche Zeitauflösung der Messapparatur

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-y^2/2\sigma^2)$$

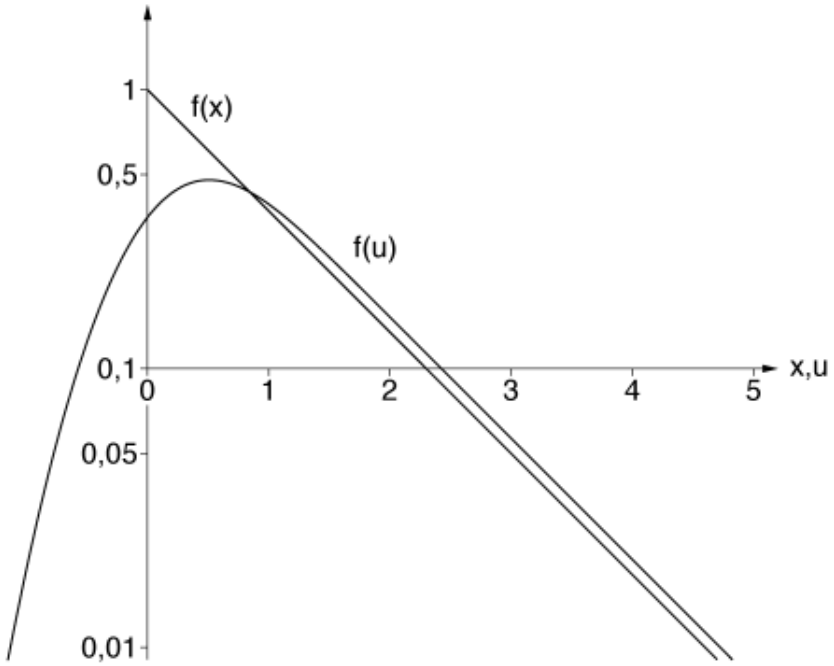
Analytische Faltung kompliziert

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma\tau}\exp\left\{\frac{\sigma^2}{2\tau^2} - \frac{u}{\tau}\right\}\int_{-\infty}^{u-\sigma^2/\tau}\exp\left(\frac{-v^2}{2\sigma^2}\right)\mathrm{d}v$$

1) Erzeuge Ereignisse gemäß Exponential-WDF

2) Addiere, subtrahiere Messfehler gemäß Gauss-WDF für jede simulierte Lebendsauer

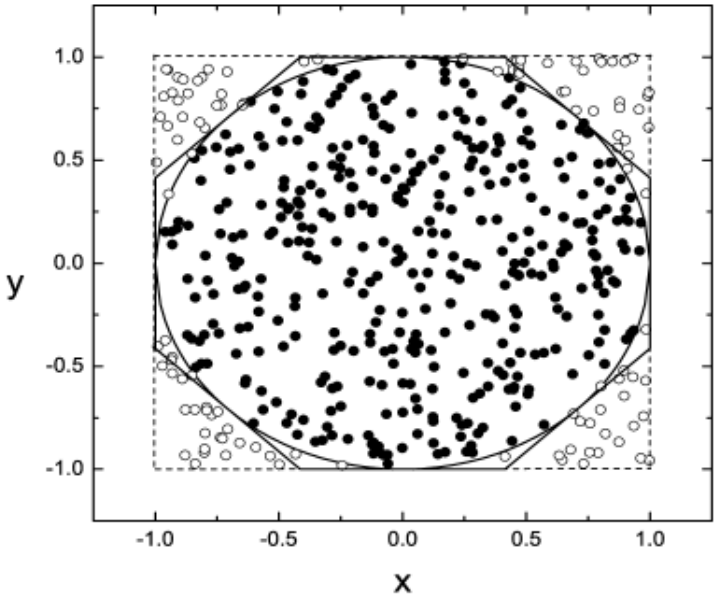
→ WDF für Experiment
= numerische Durchführung der Faltung



Verbesserte Zurückweisungsmethode

Verkleinerung des Referenzbereichs durch einführen einer Majorante

Beispiel: Bestimmung von “pi”



Referenzbereich = Quadrat: $\langle N \rangle = \frac{\pi}{4} N_0 \quad \widehat{\pi} = \frac{4N}{N_0} \quad \frac{\delta\widehat{\pi}}{\pi} = \frac{\sqrt{1-\pi/4}}{\sqrt{N_0 \pi/4}} \approx 0.52 \frac{1}{\sqrt{N_0}}$

Referenzbereich = Achteck → Reduzierung des Fehlers um Faktor ~2 bei gleicher Anzahl Versuche

Verbesserte Zurückweisungsmethode

Zu bestimmen: Integral über y(x) $I = \int_{x_a}^{x_b} y(x)\mathrm{d}x$

Majorante m(x) für y(x) mit invertierbarer Stammfunktion $M(x) = \int_{x_a}^{x} m(x')\mathrm{d}x'$

Generiere Zufallszahlen x_i gemäß m(x) durch Transformationsmethode

Erzeuge weitere gleichverteilte Zufallszahl y_i im Bereich $0 < y < m(x_i)$

Zähle Anzahl der Erfolge N, definiert über $y_i \leq y(x_i)$

Integral ist dann gegeben durch: $I = M(x_b)\frac{N}{N_0}$

$$\delta N = \sqrt{N_0 \varepsilon (1-\varepsilon)}\,,$$
$$\frac{\delta I}{I} = \frac{\delta N}{N} = \sqrt{\frac{1-\varepsilon}{N}}$$

Fehler reduziert sich mit “Anschmiegen” der Majorante m(x) an y(x).

Wichtungsmethode

Zu bestimmen: Integral über y(x) $I = \int_{x_a}^{x_b} y(x)\mathrm{d}x$

Würfele gleichverteilte Zufallszahlen im Intervall $x_a < x < x_b$

Bestimmte Mittelwert der Stichprobe für y(x): $\overline{y} = \sum_{i=1}^{N} y(x_i)/N$

Schätzwert für Integral gegeben durch: $\widehat{I} = (x_b - x_a)\overline{y}$

Entspricht numerischer Integration, aber Stützstellen zufällig.

Unsicherheit aus Varianz von y(x): $(\delta\overline{y})^2 = \frac{1}{N}\int_{x_a}^{x_b} (y(x) - \langle y\rangle)^2\, y(x)\mathrm{d}x / \int_{x_a}^{x_b} y(x)\mathrm{d}x$

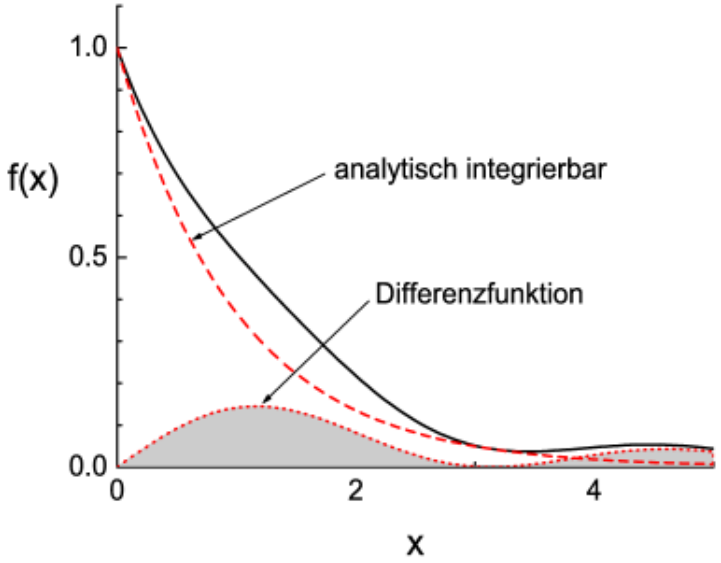
Bei MC-Integration abgeschätzt über:

$$(\delta\overline{y})^2 \approx \frac{1}{N(N-1)}\sum_i (y(x_i) - \overline{y})^2 \qquad \frac{\delta\hat{I}}{\hat{I}} = \frac{\delta\overline{y}}{\overline{y}}$$

Subtraktionsmethode

Wenn wir Funktion $\widetilde{y}(x)$ kenne, die analytisch integrierbar ist und nur wenig von $y(x)$ abweicht, lässt sich Integral umschreiben auf:

$$\int_{x_a}^{x_b} y(x)\mathrm{d}x = \int_{x_a}^{x_b} \widetilde{y}(x)\mathrm{d}x + \int_{x_a}^{x_b} \left(y(x) - \widetilde{y}(x)\right)\mathrm{d}x$$



Es muss nur noch die Differenzfunktion (schattierte Fläche) mit der Gewichtungsmethode integriert werden.

Rückführung auf Erwartungswerte

Oft Faktorisierung des Integranden sinnvoll: $y(x) = f(x)y_1(x)$

$f(x)$ WDF, nach der einfach Zufallszahlen erzeugt werden können

Integral ergibt sich als Erwartungswert in der Form:

$$\int_{x_a}^{x_b} y(x)\mathrm{d}x = \int_{x_a}^{x_b} f(x)y_1(x)\mathrm{d}x$$
$$= \langle y_1 \rangle \ .$$

Geschätzt wird der Wert durch den Stichprobenmittelwert:

$\widehat{I} = \frac{\sum_i y_1(x_i)}{N}$ mit x_i gemäß $f(x)$ erzeugt.

Fehler ergibt sich wiederum über: $(\delta\bar{y})^2 \approx \frac{1}{N(N-1)} \sum_i (y(x_i) - \bar{y})^2$ $\frac{\delta\hat{I}}{\hat{I}} = \frac{\delta\bar{y}}{\bar{y}}$