Stochastische Simulation
Skript zur Vorlesung im SS 06

basierend auf einer Mitschrift und Ausarbeitung
der Vorlesung im WS 99/00 durch Isabelle Flückiger

Hansruedi Künsch
Seminar für Statistik, ETH Zürich

April 2006

Inhaltsverzeichnis

Kapitel 1

Einführung, Beispiele

1.1 Was ist stochastische Simulation?

(stochastische) Simulation	=	Nachbildung eines (stochastischen) Systems auf einem Computer zwecks Untersuchung der Eigenschaft dieses Systems.

Damit unterscheidet sich also eine Simulation sowohl von einer mathematisch-analytischen Untersuchung als auch von einem echten Experiment oder einer Beobachtungsstudie, wo man einen Vorgang in der Natur oder in der Ökonomie beobachtet und Daten erhebt.

Gemeinsam mit einem empirischen Experiment ist:

Der empirische Ansatz (d.h. das Zählen, Messen etc.)

Gemeinsam mit einer mathematischen Untersuchung ist:

Ein mathematisches Modell (= Abbild der Wirklichkeit)

Der Vorteil von Simulationen gegenüber einem echten Experiment ist vor allem der kleinere Aufwand (an Zeit und Geld). Die Zeit in einer Simulation läuft meist wesentlich schneller ab als in der Realität, und man kann Parameter im System leicht ändern, was in der Realität oft unmöglich oder sehr kompliziert ist.

Der Vorteil von Simulationen gegenüber einer mathematischen Analyse ist, dass man komplexe Systeme untersuchen kann, die man analytisch nicht oder nur approximativ behandeln kann. Insbesondere ist man weniger auf asymptotische Näherungen angewiesen.

Der Rest dieses Kapitels wird anhand von Beispielen das Spektrum von Problemen illustrieren, die mit Simulation gelöst werden können. Bevor wir beginnen, geben wir noch eine kurze generelle Beschreibung einer stochastische Simulation. Man interessiert sich für eine (oder mehrere) Zufallsvariablen Y, die man mit Hilfe einer Funktion h aus gewissen Inputgrössen $\mathbf{X} = (X_1, \ldots, X_p)$ erhält. Die Funktion h und die Verteilung von X sind bekannt, und man möchte die Verteilung von Y oder Kenngrössen wie Erwartungwert, Standardabweichung oder Quantile berechnen. Wenn man Zufallsvektoren $\mathbf{X}$ mit der (richtigen) geforderten Verteilung erzeugen kann (wie man das macht, wird in dieser

Vorlesung besprochen), dann kann man wie folgt vorgehen: Man erzeugt N (unabhängige) Realisierungen von $\mathbf{X}$, d.h.

$$\mathbf{X}_i = (X_{i1}, \ldots, X_{ip}) \quad i = 1, \ldots, N$$

und approximiert

$$\begin{aligned} \mathrm{P}\left[h(\mathbf{X}) \leq b\right] &\approx \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathbf{1}_{[h(\mathbf{X}_i) \leq b]} \\ \text{bzw.} \quad \mathbf{E}\left[h(\mathbf{X})\right] &\approx \tfrac{1}{N} \textstyle\sum_{i=1}^{N} h(\mathbf{X}_i). \end{aligned}$$

Die Rechtfertigung dafür ist das Gesetz der Grossen Zahl. Mit Hilfe des Zentralen Grenzwertsatzes (ZGS) ist es sogar möglich, die Genauigkeit dieser Approximation anzugeben.

1.2 Randomisierte Algorithmen

Es gibt viele Algorithmen, welche den Zufall verwenden, um ein deterministisches Problem zu lösen. Da man solche Algorithmen ganz auf dem Computer implementieren will, muss man Zufallszahlen auf dem Computer zur Verfügung haben. Meist genügen uniforme Zufallszahlen, deren Erzeugung wir im nächsten Kapitel behandeln. Die Laufzeit und die Genauigkeit eines solchen Algorithmus' sind dann im Allgemeinen auch zufällig. Die Verteilung dieser Zufallsgrössen kann man auch mit Simulationen bestimmen.

Der vielleicht älteste randomisierte Algorithmus ist das Ziehen einer Stichprobe, um etwas über eine ganze Population zu erfahren. Die folgenden Beispiele sind moderner und sollen illustrieren, was für Vorteile es bringen kann, den Zufall zur Lösung eines deterministischen Problems zu verwenden.

Beispiel 1.1 (Solovay-Strassen Primzahlentest)**.** *R. Solovay und V. Strassen, SIAM J. Comput, 6 (1977), 84-85, und 7 (1978), 118, schlugen den folgenden randomisierten Algorithmus vor, um zu testen, ob eine grosse ungearde Zahl N eine Primzahl ist:*

- *Bestimme das grösste 2^k, welches $N-1$ ohne Rest teilt.*
- *Wähle ein a zufällig aus $\{1, 2, \ldots, N-1\}$ und prüfe, ob $a^{N-1} \equiv 1 (\mod N)$? Falls nicht, dann gibt man das Ergebnis "N ist zusammengesetzt" aus.*
- *Für $1 \leq j \leq k$ mache das Folgende:*
  Wenn $a^{(N-1)/2^j} \neq \pm 1 (\mod N)$, gibt man das Ergebnis "N ist zusammengesetzt" aus.
  Wenn $a^{(N-1)/2^j} = -1 (\mod N)$, gibt man das Ergebnis "N ist prim" aus.
- *Wenn die Schleife ganz durchlaufen wird, gibt man das Ergebnis "N ist prim" aus.*

Die Autoren bewiesen, dass der Algorithmus mit Wahrscheinlichkeit mindestens 3/4 das Ergebnis "N ist zusammengesetzt" ergibt, falls N tatsächlich zusammengesetzt ist, und sonst stets das Ergebnis "N ist prim". Eine Irrtumswahrscheinlichkeit von 0.25 ist natürlich grösser als was man normalerweise bereit ist zu akzeptieren. Durch (unabhängige) Wiederholungen des Algorithmus lässt sich diese jedoch leicht verkleinern.

Beispiel 1.2 ((Randomisiertes) Quicksort)**.** *(siehe Abbildung 1.1):*

Der Algorithmus ist rekursiv definiert. Er hat als Input die zu sortierende Gesamtmenge S und als Output die gleiche Menge S sortiert in aufsteigender Reihenfolge. Wenn S nur aus

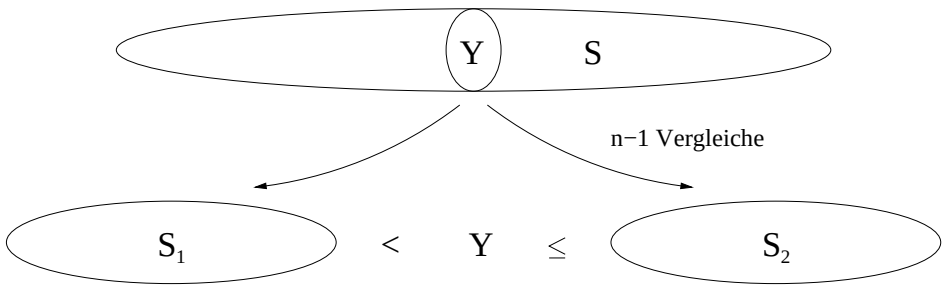

Abbildung 1.1: Quicksort

einem Element besteht, wird abgebrochen. Anderenfalls wählt der Algorithmus ein Element Y aus S und teilt S in zwei Mengen S_1 und S_2 auf, und zwar so, dass alle Elemente in S_1 kleiner als Y, und alle Elemente von S_2 grösser oder gleich Y sind. Dann ruft man den gleichen Algorithmus rekursiv zweimal auf, einmal mit S_1 als Input und einmal mit S_2 als Input.

Dieser Algorithmus ist dann schnell, wenn die beiden Mengen S_1 und S_2 etwa gleich gross werden. Wenn wir z.B. Y stets als das erste Element von S wählen, ist das ungünstig in den Fällen, wo S schon fast geordnet ist. Es zeigt sich, dass man durch eine zufällige Wahl von Y diese Probleme vermeiden kann.

Beispiel 1.3 ("Damenproblem" (E.Welzl))**.** *Die Aufgabe sei die folgende: Auf einem $n \times n$ Schachbrett sollen n Damen so positioniert werden, dass keine der Damen eine andere bedroht. In Abbildung 1.2 ist eine Lösung für ein 6×6 Brett gezeichet. Offensichtlich muss in jeder Zeile genau eine Dame platziert werden. Zu Beginn hat man viele Möglichkeiten zur Auswahl, aber gegen Schluss kann man leicht in Situationen geraten, wo in einer Zeile keine Dame mehr platziert werden kann. Ein systematisches Vorgehen wählt in jeder Zeile das erste Feld, das möglich ist und noch nie versucht wurde. Wenn man in eine Sackgasse kommt, geht man soweit zurück, bis man ein anderes Feld wählen kann. Ein zufälliges Vorgehen wählt in jeder Zeile eines der möglichen Felder mit gleicher Wahrscheinlichkeit. Gerät man in eine Sackgasse geht man ebenfalls zurück. In der untenstehenden Tabelle steht die Anzahl der benötigten Platzierungen beim systematischen und beim zufälligen Vorgehen, für einige n. Offensichlich braucht man bei der zufälligen Positionierung mit grosser Wahrscheinlichkeit weniger Versuche.*

Es gibt jedoch Fälle, wo auch die zufällige Platzierung lange gebraucht hat, z.B. bei $n = 25$. Das liegt nicht daran, dass bei $n = 25$ das Problem schwieriger ist, sondern man hat dort einfach Pech gehabt. Eine weitere Verbesserung ergibt sich durch die Idee, ganz von vorne zu beginnen, wenn man z.B. nach 100 Versuchen noch keine Lösung gefunden hat. Führt man das Verfahren bei $n = 25$ mehrmals durch, benötigt man ohne Neustart im Mittel etwa 9000 Versuche und bei Neustart nach 100 Versuchen im Mittel etwa 500 Versuche.

Zur Analyse dieser Idee "Neustart" machen wir folgende Betrachtung. Die Laufzeit $T > 0$ eines randomisierten Algorithmus sei eine Zufallsvariable mit Verteilung F. $T_1,\ T_2, \ldots$

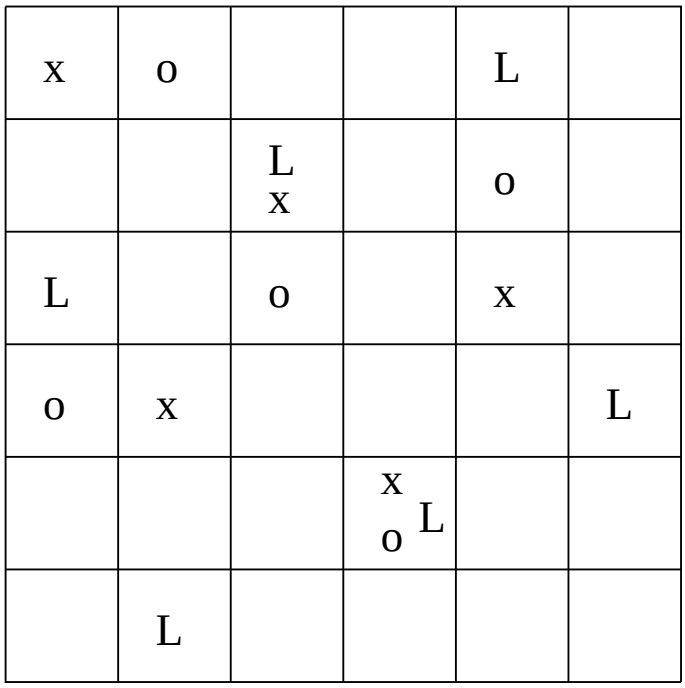

Abbildung 1.2: Damenproblem. Mit L ist die Lösung gekennzeichnet, x und o zeigen zwei falsche Versuche.

seien i.i.d. Kopien von T. Wenn man neu startet nach der Zeit c, dann startet man offensichtlich den Algorithmus N Mal, wobei $N = \min\{i;\ T_i \leq c\}$ ist. Also ist $W = (N-1)\ c + T_N \leq N\ c$ die gesamte Laufzeit des Algorithmus mit Neustart. Da N geometrisch verteilt ist mit $p = \mathrm{P}[T_i \leq c] = F(c)$, folgt

$$\mathbf{E}[W] \leq c \cdot \mathbf{E}[N] = \frac{c}{p} = \frac{c}{F(c)} < \infty,$$

während $\mathbf{E}[T]$ auch ∞ sein kann. Bei der Exponentialverteilung ist $c/\mathbf{E}[T] = c\lambda > 1 - \exp(-c\lambda) = 1 - F(c)$, also bringt dann Neustarten nichts. Dies ist auch anschaulich klar wegen der Gedächtnislosigkeit der Exponentialverteilung.

Die Vorteile des Zufalls in diesen Beispielen sind:

- Vermeidet ungünstige Wahlen, meist am Anfang des Algorithmus.
- Aufwandreduktion durch in Kauf nehmen von einer falschen Antwort mit kleiner Wahrscheinlichkeit (Primzahlentest).
- Mit Neustarten, wenn Laufzeit zu gross wird, kann man die Fälle eliminieren, wo man am Anfang eine ungünstige Wahl getroffen hat.

n	systematische Platzierung	zufällige Platzierung	n	systematische Platzierung	zufällige Platzierung
5	5	5	18	41299	234
6	31	6	19	2545	145
7	9	7	20	199635	82
8	113	13	21	8562	165
9	41	26	22	1737188	29
10	102	19	23	25428	3091
11	52	32	24	411608	833
12	261	34	25	48683	25954
13	111	270	26	397699	231
14	1899	29	27	454213	95536
15	1359	63	28	3006298	2220
16	10052	241	29	1532239	1610
17	5374	698	30	56429619	269

Tabelle 1.1: Anzahl Platzierungen zur Lösung des Damenproblems auf einem $n \times n$ Schachbrett bei systematischem und zufälligen Vorgehen.

1.3 Monte Carlo Integration

Sei $f : [0,1]^p \rightarrow \mathbb{R}$. Das Integral

$$\theta = \int_0^1 \cdots \int_0^1 f(\mathbf{x})\, d\mathbf{x}$$

kann aufgefasst werden als $\mathbf{E}\left[f(U_1, \ldots, U_p)\right]$ mit $U_1, \ldots, U_p$ i.i.d. Uniform$(0,1)$. Also kann man dieses Integral approximieren, indem man zuerst $N \times p$ Werte $U_{i,1}, \ldots, U_{i,p}$, $i = 1, \ldots, N$ erzeugt und dann das arithmetische Mittel berechnet:

$$\widehat{\theta} = \frac{1}{N} \sum_{i=1}^{N} f(U_{i,1}, \ldots, U_{i,p}).$$

Falls $\int f^2(\mathbf{x}) d\mathbf{x} < \infty$, dann

$$\mathrm{P}\left[\sqrt{N}|\widehat{\theta} - \theta| \leq z\right] \longrightarrow \Phi\left(\frac{z}{\sigma}\right) - \Phi\left(-\frac{z}{\sigma}\right)$$

wobei $\sigma^2 = \int \left(f(\mathbf{x}) - \theta\right)^2 d\mathbf{x}$ (ZGS). D.h. die Konvergenzgeschwindigkeit ist $\frac{1}{\sqrt{N}}$. Sie ist unabhängig von der Dimension p und der Glattheit von f. Ausserdem ist die Methode extrem einfach zu programmieren, und man kann leicht eine (stochastische) Fehlerabschätzung machen.

Im Vergleich dazu haben wir die numerische Integration:

$$\widetilde{\theta} = \sum w_i f(\mathbf{x}_i)$$

wobei $\mathbf{x}_1, \ldots, \mathbf{x}_p$ deterministische Stützstellen sind und w_i die Gewichte. Im einfachsten Fall hat man konstante Gewichte $w_i = \frac{1}{N}$ und die Stützstellen liegen auf einem kubischen Gitter:

$$\mathbf{x}_i \in \left\{\frac{1}{2K}, \frac{3}{2K}, \ldots, \frac{2K-1}{2K}\right\}^p,$$

mit $N = K^p$ (für gleiche Abstände). Die Konvergenzgeschwindigkeit ist in dem Fall für glatte Funktionen gleich $N^{-1/p}$. Das heisst, für $p \geq 3$ hat die Monte Carlo Integration eine schnellere Konvergenzrate.

Als Kompromiss zwischen zufälligen und systematischen Stützstellen gibt es die sogenannte Quasi Monte Carlo Integration. Diese hat dann eine Konvergenzrate $N^{-1}\log(N)^p$.

Die einfachste Art, Stützstellen für Quasi Monte Carlo zu generieren, ist das Verfahren von Halton. Im eindimensionalen Fall wählt man eine natürliche Zahl $b \geq 2$ und stellt zunächst die natürlichen Zahlen mit der Basis b dar:

$$k = \sum_{i=1}^{\infty} a_i(k) b^{i-1} \quad (a_i(k) \in \{0, 1, \ldots, b-1\}).$$

Das k-te Element der Halton-Folge ist dann

$$x_k = H(k, b) = \sum_{i=1}^{\infty} a_i(k) b^{-i} \in (0, 1).$$

Das heisst, man spiegelt die Zahl k, dargestellt zur Basis b, und setzt 0. davor. Im p-dimensionalen Fall wählt man für die j-te Komponente die Halton-Folge mit der Basis b_j, wobei b_j die j-te Primzahl bezeichnet:

$$\mathbf{x}_k = (H(k, 2), H(k, 3), H(k, 5), \ldots, H(k, b_p)).$$

Manchmal nimmt man als erste Komponente auch die gleichabständige Folge $(k - 0.5)/N$. Dies gibt zwar eine leicht bessere Konvergenzrate, dafür muss man alle Punkte neu erzeugen, wenn man die Gesamtzahl N erhöhen will.

1.4 Simulation im Unterricht

Praktisch jeder Taschenrechner/Computer hat zumindest uniforme Zufallszahlen. Damit kann man Simulation auch im Unterricht einsetzen.

Vorteile:
• Praktische Erfahrung der Variabilität des Zufalls (man kann empirisch die Variabilität des Zufalls prüfen).
• Keine Ermüdung, wie beim Werfen von echten Münzen/Würfeln.
• Auch Resultate, deren analytische Herleitung schwierig ist, können behandelt werden.

Nachteile:
• Es handelt sich um eine black box und kann als künstlich empfunden werden (man muss einfach glauben, dass ein zufälliges Experiment durchgeführt wird).
• Verwirrung kann entstehen, da es zwei Arten von Variabilität gibt, die Variabilität des Experiments, die einem interessiert *und* Variabilität der Simulation (da N endlich ist), welche man am liebsten eliminieren möchte.

Beispiel 1.4 (Unterscheidung Münzwurf von anderen binären Folgen)**.** *Nachfolgend sind fünf binäre Folgen je der Länge 100 aufgeführt. Bei welchen handelt es sich um den Wurf einer fairen Münze?*

***A**:*

1 1 1 0 1 1 0 0 1 1 0 1 0 1 1 0 1 1 0 1
0 1 1 1 1 1 1 0 1 1 0 0 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 0 0 1
0 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 0
1 0 1 1 1 1 1 0 1 1 0 1 1 1 0 0 1 0 1 1 0

***B**:*

0 1 1 0 1 0 0 1 1 0 1 1 1 0 0 0 1 0 1 1
1 0 0 0 0 1 0 0 0 1 0 1 0 0 1 1 0 0 1 1
1 0 0 0 0 0 0 1 0 1 0 1 0 1 0 1 0 0 1 0 1
1 0 1 1 0 0 1 1 0 0 0 1 0 0 0 0 0 1 1 0
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0

***C**:*

1 0 0 1 1 0 0 1 1 0 1 0 0 1 1 1 1 1 1 1
0 0 0 1 0 1 0 1 1 0 0 1 0 0 0 1 1 0 0 0
1 0 0 0 1 1 1 1 1 1 0 0 0 0 1 1 1 0 0 0 1
1 1 1 0 0 0 1 1 1 0 0 1 1 1 1 1 1 1 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0

***D**:*

1 0 0 1 0 1 0 1 0 1 1 0 1 0 0 1 1 1 1 1
0 1 1 0 1 1 0 1 0 0 0 1 1 1 1 0 1 0 0 0
0 0 1 0 0 1 1 0 1 1 0 1 0 1 1 1 1 0 1 0
0 1 0 1 0 1 1 0 0 0 0 0 0 1 1 0 0 0 1 0
1 0 1 1 0 1 0 1 0 0 0 0 1 0 0 0 1 0 1 0

***E**:*

1 0 1 1 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 1
1 0 0 1 1 0 1 1 0 0 1 0 0 1 1 0 1 0 0 1
0 1 0 1 1 0 0 1 0 1 0 1 1 0 1 1 1 1 1 1
1 1 1 1 1 0 1 0 1 1 0 0 1 1 0 1 1 0 1 1
0 1 0 0 1 0 1 1 1 1 0 0 0 1 1 0 1 0 1 1

Die Lösung sieht folgendermassen aus:
*Alle Folgen wurden mit Simulationen erzeugt. Bei **A** liegt eine Münze mit* $\mathrm{P}[X=0]=0.3$ *zu Grunde, bei **B** ist eine faire Münze, bei **C** eine symmetrische Markovkette mit* $\alpha=0.7$*, bei **D** eine symmetrische Markovkette mit* $\alpha=0.3$ *und bei **E** wieder eine faire Münze. Eine Markovkette ist ein zufälliger Prozess, bei dem die Wahrscheinlichkeit für den nächsten Zustand abhängt vom jetzigen Zustand, aber nicht von der Vorgeschichte. Markovketten mit endlich vielen Zuständen können durch gerichtete Graphen beschrieben werden. Die Abb. 1.3 zeigt eine binäre symmetrische Markovkette.*

Wie kann man herausfinden, ob es sich um eine faire Münze handelt oder nicht? Zwei mögliche und meist auch effektive Testgrössen sind die Anzahl der Wechsel und die Länge des längsten "Run's". Beide nützen aus, dass der Zufall weniger oft ausgleicht, als man naiverweise vermutet.

Die Anzahl der Wechsel ist natürlich binomial$(n-1,0.5)$*-verteilt, also sind wegen des*

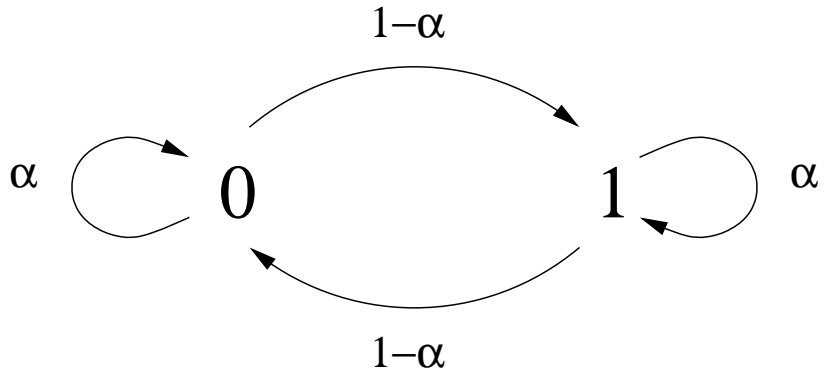

Abbildung 1.3: Symmetrische Markovkette mit den Zuständen 0 und 1

zentralen Grenzwertsatzes mehr als $(n-1)/2 + 0.82 * \sqrt{n-1} + 0.5$ *Wechsel verdächtig.*

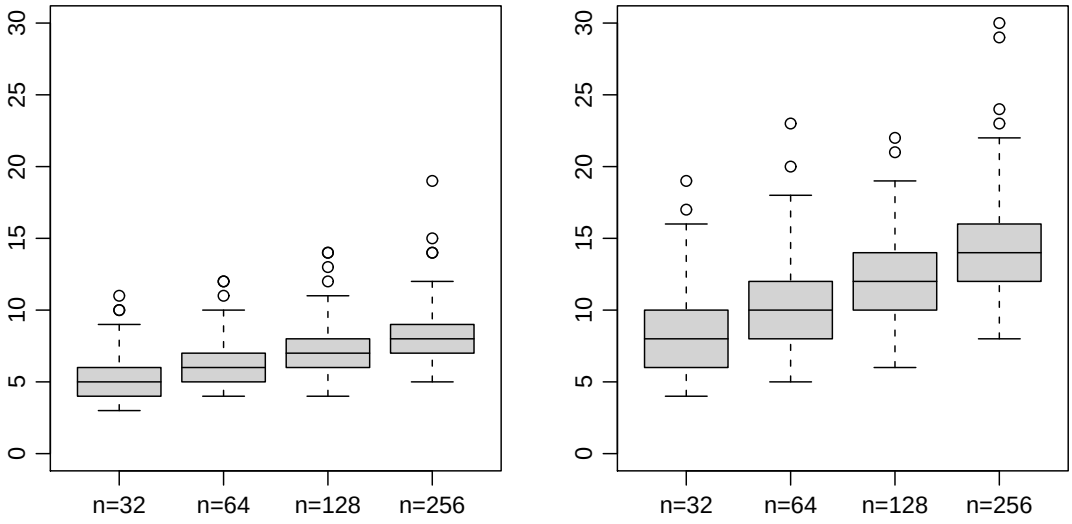

Abbildung 1.4: Im linken Bild sind Boxplots der längsten Run's bei n Münzwürfen mit einer fairen Münze dargestellt, rechts als Vergleich dazu, die Boxplots einer Markovkette mit $\alpha = 0.7$, für $n = 32,\ 64,\ 128,\ 256$.

Die Verteilung der Längen des maximalen Run's ist etwas komplizierter zu bestimmen, siehe z.B. Feller (1968), Kap. XIII.7. Einfacher geht es mit Simulationen. Abbildung 1.4, linkes Bild, zeigt Boxplots der Längen der maximalen Run's bei n Münzwürfen mit einer fairen Münze, kommt folgendes Bild heraus. Man erkennt, dass der Median des längsten Run's ca. $\log_2 n$ *ist. Als Gegensatz dazu zeigt das linke Bild das Verhalten bei einer Markovkette mit* $\alpha = 0.7$*. Man sieht, dass die Mediane deutlich höher liegen und die Streuung gegen oben grösser ist. Man hat damit also ein Kriterium um Zufall zu erkennen.*

Beispiel 1.5 (Irrfahrt und das Gesetz der Grossen Zahlen)**.** *Irrfahrten sind ein sehr gutes Beispiel für den Stochastikunterricht, weil man dort naiverweise die Variabilität des Zu-*

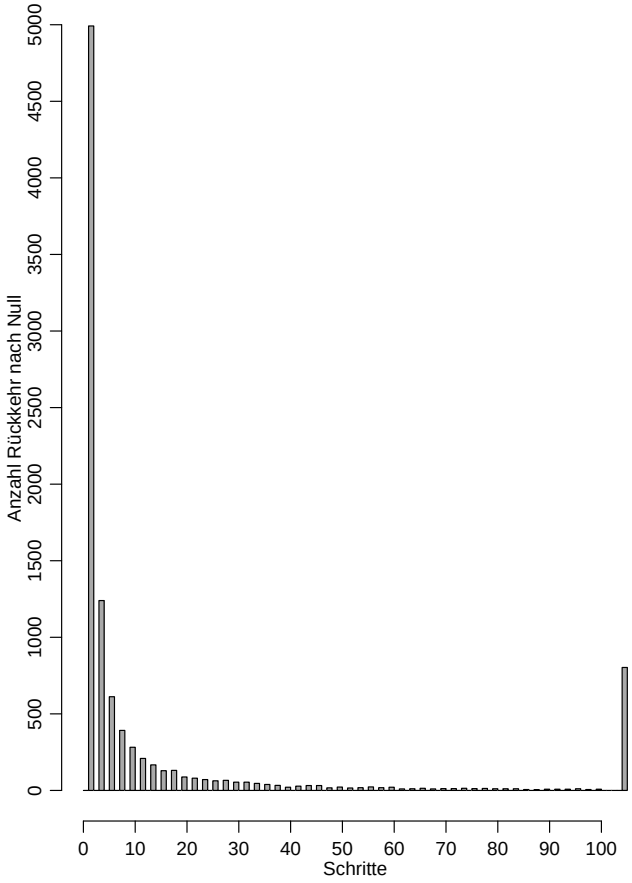

Abbildung 1.5: Histogramm der ersten Rückkehr nach Null bei 10'000 simulierten Irrfahrten der Länge 100. Der letzte Balken im Histogramm gibt die Anzahl der Irrfahrten an, die während der ersten 100 Schritte nie nach Null zurückkehrten.

falls sehr leicht unterschätzt. Eine relativ elementare analytische Behandlung ist möglich, welche nur die Stirling Formel und das Reflektionsprinzip benutzt, siehe die Einführungsvorlesung, bzw. Feller (1968), Band 1, Kapitel III. Alternativ kann man die Phänomene auch sehr gut mit Simulationen illustrieren und so auch auf der Mittelschulstufe behandeln.

Wir betrachten zuerst die erste Rückkehr T_0 *einer Irrfahrt nach dem Startpunkt Null. In der Interpretation einer Irrfahrt als Bilanzentwicklung in einem fairen Spiel ist das der erste Zeitpunkt, wo das Spiel ausgeglichen war. Es ist klar, dass diese Zeit gerade sein muss, und man könnte denken, dass diese Verteilung praktisch nur auf ein paar wenige kleine Werte konzentriert ist. Wir haben 100 Schritte einer Irrfahrt tausend Mal simuliert und die Verteilung von* $\min(T_0, 102)$ *empirisch bestimmt, siehe Abbildung 1.5. Es ist offensichtlich, dass die Wahrscheinlichkeiten sehr langsam abfallen und dass die Wahrscheinlichkeit* $\mathrm{P}[T_0 > 100]$ *bei weitem nicht vernachlässigbar ist. Dies stimmt natürlich sehr gut mit dem analytischen Resultat* $\mathrm{P}[T_0 > 2n] \sim \frac{1}{\sqrt{\pi n}}$ *überein.*

Als zweites betrachten wir den letzten Besuch L *in Null vor* $t = 100$ *(wenn die Irrfahrt in 100 Schritten nie nach Null zurückkehrt, ist definitionsgemäss* $L = 0$*). Bei einem fairen Spiel ist das der Zeitpunkt, von dem an ein Spieler immer führt. Was für eine Form des Histogramms würden wir erwarten ? Eine mögliche Vermutung wäre eine Verteilung, die vor allem auf Werte nahe bei 100 konzentriert ist. Analytisch kann man zeigen, dass das*

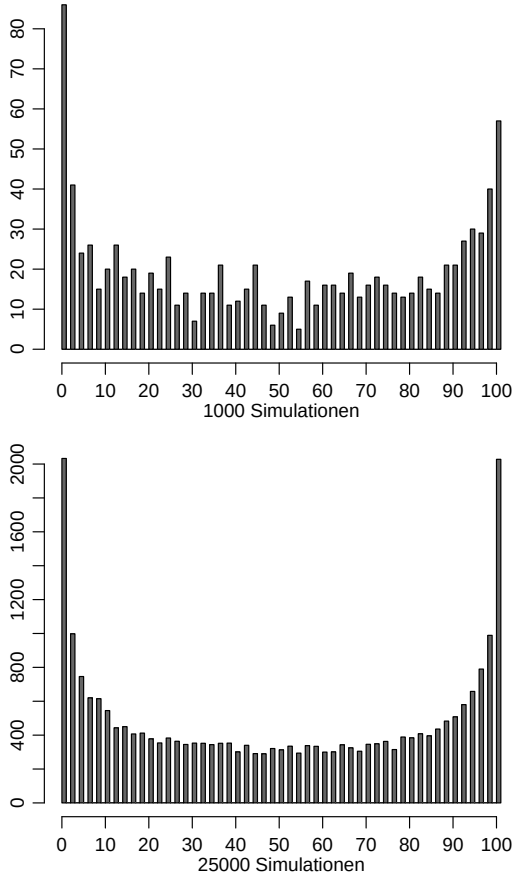

Abbildung 1.6: Im oberen Histogramm wurde der letzte Besuch in Null vor dem hundertsten Schritt von 1000 Random Walks aufgetragen. Hier ist die U-Form schon erkennbar. Im unteren Histogramm ist das gleiche für 25000 Random Walks aufgetragen. Hier sieht man sehr schön die U-Form.

Histogramm eine U-Form haben wird (sogenanntes arcus-sinus Gesetz). Der letzte Besuch in Null findet also entweder am Anfang oder kurz bevor die Irrfahrt gestoppt wird, statt. In Abbildung 1.6 wurden je einmal 1000 und 25000 Random Walks simuliert und der Zeitpunkt ihrer letzte Rückkehr nach Null aufgetragen. In beiden ist tatsächlich die U-Form gut ersichtlich. Im oberen Histogramm ist einfach die Streuung, die von der Variabilität der Simulation her kommt, grösser.

Beispiel 1.6 (Vertrauensintervalle)**.** *Hier simuliert man 100 Konfidenzintervalle für die Binomialverteilung* $\mathrm{Bin}\,(20, 0.38)$*. D.h. wir erzeugen einen Wert* x *nach dieser Verteilung und berechnen das Vertrauensintervall für den Erfolgsparameter. Das Vertrauensintervall* $I(x) = [k_1(x), k_2(x)]$ *ist zufällig, und es gilt* $\mathrm{P}\left[I(x) \ni p\right] \geq 1 - \alpha \quad \forall\, p$*. In unserem Fall ist* $p = 0.38$*. In einer Simulation kann man den wahren Wert sehr leicht variieren. Die 100 Vertrauensintervalle für* $\alpha = 0.05$ *sind in Abb. 1.7 gezeigt. Man sieht, dass es Vertrauensintervalle gibt, die den Wert* $p = 0.38$ *gar nicht enthalten. Die Wahrscheinlichkeit, dass das Vertrauensintervall den Wert* $p = 0.38$ *enthält, ist* $1 - \alpha$*. Im Beispiel stimmt dies ziemlich genau.* 94 *mal ist der Wert* 0.38 *in den Konfidenzintervallen enthalten.*

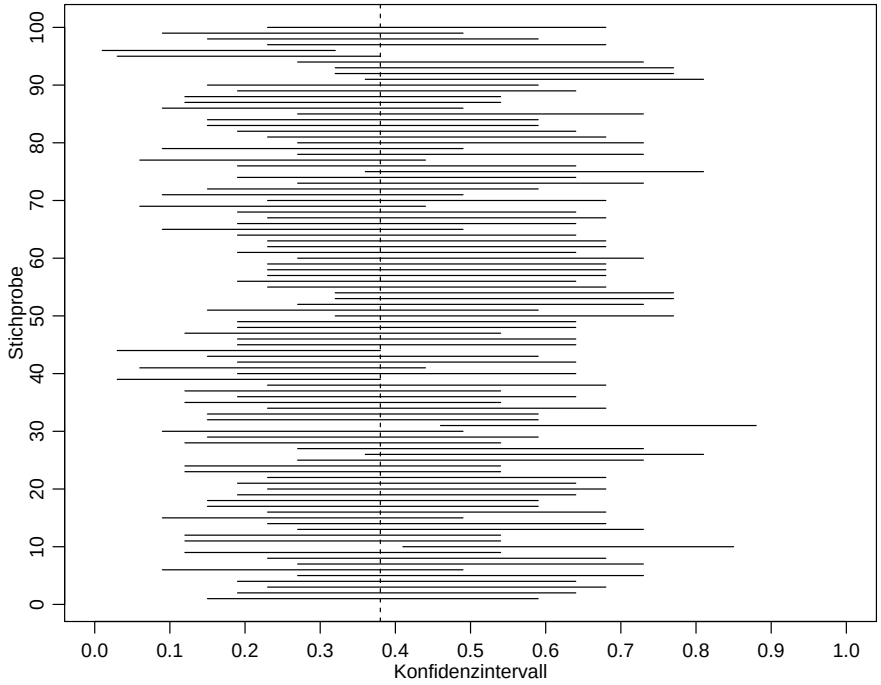

Abbildung 1.7: 95%-Vertrauensintervalle für 100 Zufallszahlen, die nach $\mathrm{Bin}\,(20, 0.38)$ generiert wurden. 94 mal liegt der wahre Wert 0.38 im Intervall.

Hier zeigt sich jedoch auch, dass die verschiedenen Stufen des Zufalls, wie sie bei einer Simulation immer vorhanden sind, verwirren können. Die Vertrauensintervalle sind zufällig und haben eine feste Überdeckungswahrscheinlichkeit $1 - \alpha$*. Die Simulation enthält aber wieder ein Zufallselement, so dass die Anzahl Intervalle, die den wahren Parameter überdecken, ebenfalls variiert. Diese Anzahl ist wieder binomialverteilt, mit Erfolgswahrscheinlichkeit* α *und Anzahl Wiederholungen gleich der Anzahl Replikate der Simulation. Dies könnte man auf einer nächsten Stufe wieder mit einer Simulation überprüfen, etc..*

1.5 Verteilung von Schätzern und Teststatistiken

1.5.1 Genauigkeit des gestutzten Mittels

Van Zwet (1985) beschreibt ein historisches Beispiel einer Simulation, die in der Vor-Computerzeit von Astronomen durchgeführt wurde. Wir zitieren aus diesem Artikel

"In the issue of May 20, 1942, of the Bulletin of the Astronomical Institutes of the Netherlands, E. Hertzsprung, director of the Observatory at Leiden, describes a sampling experiment to determine the variance of the trimmed mean. In connection with the determination of relative proper motions of stars in the Pleiades, Hertzsprung discusses how one should assign weights to the observed values to account for differences in quality of the observations. He writes: "The simplest way to deal with exorbitant observations is to reject them. In order to avoid special rules for onesided rejection the easy way of symmetrical rejection of the largest deviations to each side may be considered. The first question is then: How much is, in the case of Gaussian distribution of errors, the weight of the result diminished by a priori symmetrical rejection of outstanding observations? As the

mathematical treatment of this question appears to be laborious beyond the needs mentioned above I gave preference to an empirical answer. On each of 12534 slips of paper was written with two decimals a deviation from zero in units of the mean error, in such a way that these deviations showed a Gaussian distribution. Thus 50 slips were marked with .00, 50 with +.01, 50 with -.01 etc.. Of these slips somewhat more than 1000 times 24 were picked out arbitrarily. Such 24 slips were in each case arranged according to the size of the deviation and the mean squares of the sums of $24 - x$ deviations calculated after symmetrical rejection of $x = 0, 2, 4, \ldots, 22$ extreme values."

This paragraph should warm a statistician's heart, except that he may feel slightly uneasy about "somewhat more than 1000" replications. And he has reason to feel uneasy: "Of all these samples of 24 exactly 1000 were picked out in such a way that the sum of all 24 deviations ($x = 0$) fairly well showed a Gaussian distribution with a mean square of 24." From a theoretical point of view, this ruins a perfectly good sampling experiment, as Van Dantzig was quick to point out, especially since no further information is supplied. There is no way of assessing the accuracy of the estimated variances any more. On the other hand, if we assume that this data cleaning was done sensibly, there seems to be no reason, a priori, why the estimates should be much worse than they would have been otherwise."

Formulieren wir dieses Problem und die Lösung von Hertzsprung mathematisch: Wir betrachten das gestutzte arithmetische Mittel

$$\bar{X}^{(k)} = \frac{1}{n-2k} \sum_{j=k+1}^{n-k} X_{(j)},$$

wobei die X_i i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ sind und $X_{(j)}$ das j-te Element der geordneten Stichprobe ist. Es geht darum zu untersuchen, um wieviel grösser die Varianz $\sigma(n,k)^2$ dieses gestutzten Mittels (mit $k = 1, 2, \ldots$) ist als die Varianz des gewöhnlichen Mittels $\sigma(n,0)^2 = \sigma^2/n$ im Fall $n = 24$. Offensichtlich kann man sich auf den Fall $\mu = 0$ und $\sigma = 1$ beschränken. Dazu wurde ein Simulationsexperiment durchgeführt, aber im Gegensatz zu heute standen keine Zufallszahlen auf dem Computer zur Verfügung. Daher wurde die reelle Achse diskretisiert in Intervalle der Länge 0.01, so dass nur Werte der Form $x_k = k \cdot 0.01$ vorkommen. Die simulierten Werte entstanden durch Ziehen mit Zurücklegen von Zetteln aus einer Urne.

Wie muss die Urne zusammengesetzt sein, damit die simulierten Werte in guter Näherung normalverteilt sind? Wenn die Urne insgesamt M Zettel enthält und der Wert x_k n_k-mal vorkommt, dann hat x_k die Wahrscheinlichkeit n_k/M. Daher muss

$$n_k \approx M(\Phi(x_k + 0.005) - \Phi(x_k - 0.005)) \approx 0.01 \cdot M\varphi(x_k)$$

sein. Mit der Wahl $n_0 = 50$ erhält man $M = 12533.2$. Vermutlich wurde aufgerundet, um eine perfekt symmetrische Zusammensetzung der Urne zu erreichen.

Für ein Simulationsexperiment erzeugt man $N = 1000$ Replikate einer Messreihe der Länge $n = 24$ (durch Ziehen von Zetteln oder mit dem Computer), d.h. X_{ij} mit $i = 1, \ldots, N = 1000$ und $j = 1, \ldots, n = 24$. Die zeilenweise sortierten Werte bezeichnen wir mit $X_{i,(1)}, \ldots, X_{i,(n)}$ und das gestutzte Mittel mit

$$\bar{X}_i^{(k)} = \frac{1}{n-2k} \sum_{j=k+1}^{n-k} X_{i,(j)}.$$

Schliesslich schätzt man $\sigma(n,k)^2/\sigma(n,0)^2 = n\sigma(n,k)^2$ durch

$$\frac{n}{N}\sum_{i=1}^{N}\left(\bar{X}_i^{(k)}\right)^2.$$

Hertzsprung verwendete jedoch nicht diese Schätzung, sondern

$$\frac{\sum_{i=1}^{N}(\bar{X}_i^{(k)})^2}{\sum_{i=1}^{N}(\bar{X}_i^{(0)})^2},$$

d.h. er ignorierte die Tatsache, dass $\sigma(n,0)^2 = 1/n$ ist. Das scheint auf den ersten Blick keine gute Idee, aber bei näherer Betrachtung sieht man, dass sich dadurch die Genauigkeit der Simulation verbessert, weil Zähler und Nenner stark positiv korreliert sind. Dies werden wir im Abschnitt 3.10.2 genauer besprechen.

Das Vorgehen, einzelne Zeilen von X_{ij} zu eliminieren und so die Übereinstimmung zwischen dem Histogramm der arithmetischen Mittel $\bar{X}_i^{(0)}$ und der theoretischen $\mathcal{N}\left(0,\frac{1}{n}\right)$-Verteilung zu verbessern, ist zweifelhaft. Insbesondere kann man danach die Genauigkeit nur noch sehr schwer beurteilen, und man weiss nie, ob man nicht zuviel korrigiert hat.

Heutzutage würde man dieses Problem mit Asymptotik lösen, d.h. mit einer analytischen Näherung anstelle einer Simulation. Dies war sogar in den vierziger Jahren des letzten Jahrhunderts naheliegend. Van de Hulst, ein Student in Astronomie zu der Zeit, hatte von diesem Experiment gehört und beschloss, stattdessen eine analytische Antwort zu suchen. Das gelang ihm auch im Wesentlichen, was für diese Zeit eine erstaunliche Leistung war. Er korrespondierte darüber mit van Dantzig, dem Pionier der Statistik in Holland, der aber die fehlende mathematische Strenge bemängelte und der Leistung nicht gerecht wurde.

Das relevante Resultat für dieses Problem lautet

Satz 1.1. *Wenn X_i i.i.d. $\sim f(x)dx$ und $f(x)$ symmetrisch ist bezüglich Null, dann gilt für $n \to \infty$ und $\frac{k}{n} \to \alpha$*

$$\mathrm{P}\left[\sqrt{n}\bar{X}_n^{(k)} \le x\right] \to \Phi\left(\frac{x}{\sigma_\alpha}\right)$$

wobei $\sigma_\alpha^2 = \frac{1}{(1-2\alpha)^2}\int \min\left(x^2,a^2\right) f(x)dx$ und $a = F^{-1}(1-\alpha)$.

Beweis. Siehe math. Statistik □

D.h. also anschaulich $\sigma(n,k)^2 \approx \sigma_{k/n}^2/n$.

Mit Simulation, bzw. Asymptotik kann man nicht nur das gestutzte Mittel untersuchen, sondern ganz allgemein die Verteilungen von beliebigen Schätzern bestimmen und so Vergleiche zwischen Schätzern durchführen. Analog kann man bei Tests die kritischen Grenzen bestimmen, bzw. Machtberechnungen durchführen.

Sowohl die Asymptotik als auch die Simulation haben Vorteile. Die Asymptotik zeigt im Allgemeinen die ganze Abhängigkeit von zusätzlichen Parametern, z.B. der Verteilungsfamilie, während die Simulation immer nur ein paar wenige spezifische Situationen studieren kann. Umgekehrt beruht die Asymptotik immer auf einem fiktiven Grenzwert, während eine Simulation ein paar feste Werte des Stichprobenumfangs untersucht. Heutzutage kombiniert man meist Asymptotik mit Simulationen, um so wenigstens ein paar Hinweise dafür zu erhalten, wie gut die Asymptotik schon für endliches n funktioniert. Analytische Fehlerabschätzungen sind nämlich sehr schwierig und kommen mit den heutigen Techniken zu pessimistisch heraus.

1.5.2 Bootstrap

In 1.5.1 wurde die Verteilung der X_i als bekannt vorausgesetzt (z.B. als normalverteilt). Dies ist kein Problem, wenn man entscheiden will, ob man das gestutzte oder das ungestutzte Mittel verwenden soll. Die Normalverteilung stellt dann die Idealsituation dar, und wenn in diesem Fall mit Stutzen die Genauigkeit nur unwesentlich schlechter ist, soll man stutzen, denn man gewinnt dadurch Schutz vor Ausreissern.

Anders sieht es aus, wenn man die Standardabweichung $\sqrt{\mathrm{Var}\,(T_n)}$ einer Schätzung T_n als Genauigkeitsangabe verwenden will, z.B. für ein genähertes Vertrauensintervall $T_n \pm 2\sqrt{\mathrm{Var}\,(T_n)}$. In diesem Fall kennt man die Verteilung der X_i nicht, und die Annahme der Normalverteilung (oder einer andern Verteilung) kann zu falschen Schlüssen führen.

In dieser Situation schätzt man als Ausweg auch die Verteilung der X_i aus den gleichen Daten, mit denen man T_n berechnet hat, und simuliert mit der geschätzten Verteilung. Solche Verfahren heissen *Bootstrap*, ein englischer Ausdruck für "sich an den eigenen Haaren aus dem Sumpf ziehen".

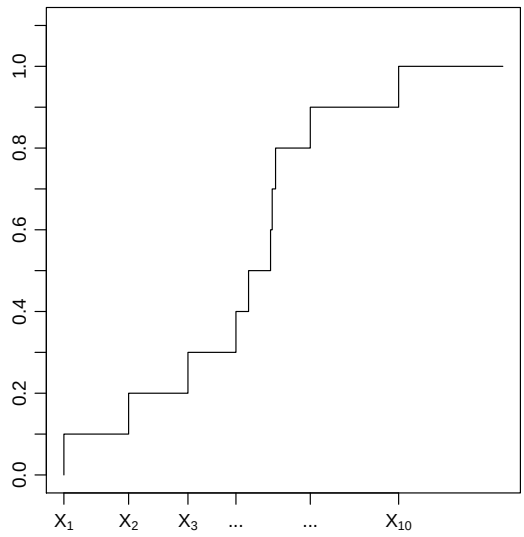

Abbildung 1.8: Empirische Verteilungsfunktion.

In mathematischer Sprache haben wir die folgende Situation

- Die $X_1, \ldots, X_n$ seien i.i.d. p-dimensionale Zufallsvektoren mit der Verteilung F.
- Wir benutzen eine Schätzung $T_n = t_n(X_1, \ldots, X_n)$, $t_n : \mathbb{R}^{np} \to \mathbb{R}$, eines Parameters der zugrundeliegenden Verteilung F, z.B. schätzen wir die Korrelation einer Verteilung durch die empirische Korrelation der Daten.
- Die Genauigkeit von T_n ist gegeben durch den Standardfehler $\sigma_n(F) = \sqrt{\mathrm{Var}_{\mathrm{F}}\,(t_n(X_1, \ldots, X_n))}$.
- Gesucht ist eine Schätzung von $\sigma_n(F)$.

Eine Schätzung von $\sigma_n(F)$ erhalten wir durch das sogenannte Einsetzprinzip: Schätze F durch die empirische Verteilung $\widehat{F}_n$ und schätze dann $\sigma_n(F)$ durch $\sigma_n(\widehat{F}_n)$.

Die empirische Verteilung $\widehat{F}_n$ der Daten ist eine diskrete Verteilung, die jedem beobachteten Wert x_i $(i = 1, \ldots, n)$ die Masse $\frac{1}{n}$ gibt. Wenn X_i univariat ist, ist die zugehörige Verteilungsfunktion eine Treppenfunktion wie in Abbildung 1.8 gezeigt. Dies heisst, wenn $X^* \sim \widehat{F}_n$, dann $\mathrm{P}[X^* = x_j] = \frac{1}{n}$ für $j = 1, \ldots, n$. Für eine Simulation von $\widehat{F}_n$-verteilten Zufallsvariable X^* schreibt man daher einfach die Beobachtungen auf n Zettel, legt diese in eine Urne und zieht zufällig mit Zurücklegen aus dieser Urne.

Mit dem Einsetzprinzip müssen wir die Varianz von $t_n(X_1^*, \ldots, X_n^*)$ berechnen, wobei die $X_1^*, \ldots, X_n^*$ i.i.d. sind mit $X_i^* \sim \widehat{F}_n$. Das kann man praktisch nie geschlossen durchführen, aber man kann analog wie beim gestutzten Mittel simulieren. Die Realisierungen X_i^* erhält man mit dem oben beschriebenen Verfahren.

Zusammenfassend lautet also der Bootstrap-Algorithmus:

Algorithmus 1.1.

1. *Ziehe $n \cdot N$ Werte gemäss $\widehat{F}_n$, d.h. ziehe $n \cdot N$ mal mit Zurücklegen aus den Beobachtungen $(x_1, \ldots x_n)$. Man bekommt dann eine Matrix $(X_{ij}^*;\ 1 \le i \le N,\ 1 \le j \le n)$*
2. *Wende t_n auf jede Zeile an: $T_{n,i}^* = t_n(X_{i1}^*, \ldots, X_{in}^*)$*
3. *$\sigma_n^2(\widehat{F}_n) \approx \frac{1}{N-1} \sum_{i=1}^N (T_{n,i}^* - \bar{T}_n^*)^2$ mit $\bar{T}_n^* = \frac{1}{N} \sum_{k=1}^N T_{n,k}^*$*

Für den Bootstrap ist es nicht zwingend, die Verteilung der X_i durch die empirische Verteilung zu schätzen. Man könnte auch ein parametrisches Modell anpassen, oder eine geglättete Version der empirischen Verteilung. Das entscheidende am Bootstrap ist, dass man die Genauigkeit einer Statistik schätzt, indem man weitere, künstliche Beobachtungen erzeugt aus einer Verteilung, die man mit den echten Beobachtungen geschätzt hat.

1.6 Simulation in der Bayesstatistik

Wir benötigen hier die Konzepte von bedingten Verteilungen im absolut stetigen Fall, die wir zunächst kurz einführen (bzw. repetieren).

1.6.1 Absolut stetige Verteilungen von Zufallsvektoren

Sei $\mathbf{X} = (X_1, X_2)$ ein zweidimensionaler Zufallsvektor, dessen Verteilung eine Dichte f hat. Das bedeutet, dass

$$\mathrm{P}[(X_1, X_2) \in A] = \int_A f(x_1, x_2) dx_1 dx_2$$

gilt für alle (messbaren) Teilmengen A von $\mathbb{R}^2$. Als Dichte kann eine beliebige messbare Funktion f auf $\mathbb{R}^2$ auftreten, welche die beiden Bedingungen

$$f \ge 0, \quad \int_{\mathbb{R}^2} f(x_1, x_2) dx_1 dx_2 = 1$$

erfüllt. Anschaulich kann man die Dichte interpretieren als

$$\mathrm{P}[X_1 \in dx_1, X_2 \in dx_2] = f(x_1, x_2) dx_1 dx_2,$$

d.h. die Dichte ist die Wahrscheinlichkeit, dass $\mathbf{X}$ in einem kleinen Rechteck liegt, dividiert durch die Fläche dieses Rechtecks. Eine zweidimensionale Dichte beschreibt eine Massenverteilung auf der Ebene mit Gesamtmasse 1.

Um Verwechslungen mit der Randdichte (siehe unten) auszuschliessen, nennt man f manchmal auch die *gemeinsame Dichte*.

Randdichten und Unabhängigkeit. Unter den Randverteilungen versteht man die Verteilung von X_1, bzw. X_2 für sich allein, d.h. $\mathrm{P}[X_i \in B]$ für $B \subset \mathbb{R}$. Da $\mathrm{P}[X_1 \in B] = \mathrm{P}[(X_1, X_2) \in B \times \mathbb{R}]$ folgt sofort, dass die Randverteilungen absolut stetig sind mit Dichten

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$$

und

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1.$$

Man erhält also die Randdichten, indem man die Massenverteilung entlang einer der beiden Koordinatenachsen akkumuliert.

Die beiden Komponenten X_1 und X_2 heissen *unabhängig*, falls

$$\mathrm{P}[X_1 \in B_1, X_2 \in B_2] = \mathrm{P}[X_1 \in B_1] \cdot \mathrm{P}[X_2 \in B_2]$$

gilt für alle $B_i \subset \mathbb{R}$. Wenn man für B_i ein infinitesimales Intervall dx_i wählt, ergibt das die Bedingung

$$f(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2),$$

und man kann auch streng beweisen, dass diese beiden Bedingungen äquivalent sind. Meist geht man so vor, dass man die Unabhängigkeit von X_1 und X_2 *postuliert* und so aus den Randdichten die gemeinsame Dichte als das Produkt erhält. Ohne Unabhängigkeit genügt es nicht, die Randdichten zu kennen, um die gemeinsame Dichte festzulegen.

Bedingte Verteilung und Bayesformel. Für stetige Zufallsvariablen ist $\mathrm{P}[X_i = x] = 0$ für jedes feste x. Daher bereitet die Definition von bedingten Wahrscheinlichkeiten gegeben $X_i = x$ Schwierigkeiten. Anschaulich setzt man für ein Ereignis E

$$\mathrm{P}[E \mid X_i = x] = \frac{\mathrm{P}[E, X_i \in dx]}{\mathrm{P}[X_i \in dx]},$$

d.h.man ersetzt $X_i = x$ durch das Ereignis, dass X_i in einem infinitesimalen Intervall ist. Die rechte Seite kann man als Grenzwert von

$$\frac{\mathrm{P}[E, x \leq X_i \leq x + h]}{\mathrm{P}[x \leq X_i \leq x + h]}$$

für h gegen null berechnen, wenn f stetig ist. Insbesondere ergibt sich

$$\mathrm{P}[X_2 \leq b \mid X_1 = x_1] = \int_{-\infty}^{b} f(x_1, x_2) dx_2 / f_{X_1}(x_1).$$

Deshalb definiert man die bedingte Dichte von X_2 gegeben $X_1 = x_1$ als

$$f_{X_2}(x_2 \mid X_1 = x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)}.$$

Dies ist für jedes x_1 eine eindimensionale Dichte. Der Nenner sorgt dafür, dass das Integral bezüglich x_2 über ganz $\mathbb{R}$ gleich 1 ist. Die bedingte Dichte ist also nichts anderes als ein eindimensionaler Schnitt der gemeinsamen Dichte mal eine Normierung auf Gesamtmasse 1. Wenn $f_{X_1}(x_1) = 0$ ist, versagt obige Definition. In dem Falle spielt es keine Rolle, wie man die bedingte Dichte von X_2 gegeben $X_1 = x_1$ definiert (da ein solches x_1 nicht auftritt). Wenn X_1 und X_2 unabhängig sind, ist die bedingte Dichte nichts anderes als die Randdichte, was natürlich intuitiv sofort einleuchtet.

Durch Multiplikation mit $f_{X_1}(x_1)$ in obiger Formel folgt sofort

$$f(x_1, x_2) = f_{X_2}(x_2 \mid X_1 = x_1) f_{X_1}(x_1).$$

Das bedeutet insbesondere, dass man statt der gemeinsamen Dichte auch eine Randdichte und eine bedingte Dichte festlegen kann (und zwar beliebig) und dann daraus die gemeinsame Dichte berechnen kann. Dies wird sehr oft angewandt: Es ist manchmal leichter sich zu überlegen, was plausible Formen für die Rand- und für die bedingte Dichte sind, als direkt eine plausible gemeinsame Verteilung hinzuschreiben.

Die folgenden Formeln folgen nun sehr leicht:

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_2}(x_2 \mid X_1 = x_1) f_{X_1}(x_1) dx_1,$$

$$f_{X_1}(x_1 \mid X_2 = x_2) = \frac{f_{X_2}(x_2 \mid X_1 = x_1) f_{X_1}(x_1)}{\int_{-\infty}^{\infty} f_{X_2}(x_2 \mid X_1 = x_1') f_{X_1}(x_1') dx_1'}.$$

Die erste Formel ist ein stetiges Analog zum Satz von der totalen Wahrscheinlichkeit, und die zweite Formel entspricht dem Satz von Bayes. Die Bedeutung dieser beiden Resultate kann nie genügend stark betont werden. In vielen Fällen ist es nützlich zu beachten, dass der Nenner in der Bayes-Formel nur eine Normierung ist. Man kann also auch schreiben

$$f_{X_1}(x_1 \mid X_2 = x_2) \propto f_{X_2}(x_2 \mid X_1 = x_1) f_{X_1}(x_1).$$

In mehr als zwei Dimensionen geht im Wesentlichen alles analog. Es gibt dann jedoch nicht nur eindimensionale, sondern auch höherdimensionale Randverteilungen, und man kann auf beliebige Untergruppen von Variablen bedingen.

Bedingte Erwartung. Die bedingte Erwartung von X_2 gegeben $X_1 = x_1$ ist definiert als der Erwartungswert bezüglich der bedingten Verteilung von X_2 gegeben $X_1 = x_1$:

$$\mathbf{E}[X_2 \mid X_1 = x_1] = \frac{\int_{-\infty}^{\infty} x_2 f(x_1, x_2) dx_2}{f_{X_1}(x_1)}.$$

Dies ist also eine Funktion von $\mathbb{R}$ nach $\mathbb{R}$. Wenn wir diese Funktion zusammensetzen mit der Zufallsvariable X_1, dann erhalten wir eine neue Zufallsvariable, die wir mit $\mathbf{E}[X_2 \mid X_1]$ bezeichnen. Diese Zufallsvariable nimmt einfach den Wert $\mathbf{E}[X_2 \mid X_1 = x_1]$ an, falls $X_1 = x_1$ ist. Das macht z.B. Sinn, wenn der Wert von X_1 im Moment noch nicht bekannt ist, aber vor dem Wert von X_2 ermittelt werden wird.

Die grösste Bedeutung hat die bedingte Erwartung bei einer schrittweisen Berechnung eines Erwartungswertes. Es gilt per Definition

$$\begin{aligned} \mathbf{E}[\mathbf{E}[X_2 \mid X_1]] &= \int_{-\infty}^{\infty} \mathbf{E}[X_2 \mid X_1 = x_1] f_{X_1}(x_1) dx_1 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 f(x_1, x_2) dx_2 dx_1 = \int_{-\infty}^{\infty} x_2 f_{X_2}(x_2) dx_2 = \mathbf{E}[X_2]. \end{aligned}$$

Analog kann man zeigen, dass für jede (beschränkte) Funktion $g : \mathbb{R} \to \mathbb{R}$ gilt

$$\mathbf{E}\left[g(X_1)\,\mathbf{E}\left[X_2 \mid X_1\right]\right] = \mathbf{E}\left[g(X_1)X_2\right].$$

In einer mathematisch abstrakten Behandlung der bedingten Erwartung wird diese Eigenschaft zur Definition verwendet.

1.6.2 Einführung in die Bayesstatistik

In der Statistik geht man meist von einem Modell für die Beobachtungen aus: $\mathbf{X} = (X_1, \ldots, X_n) \sim p_\theta(\mathbf{x})\mu(d\mathbf{x})$ (μ ist das Bezugsmass, z.B. das Lebesguemass in $\mathbb{R}^n$). Die Parameter sind $\theta \in \Theta$, und man will von den Beobachtungen Rückschlüsse auf die Parameter ziehen. Da die Parameter auch unendlich dimensional sein können, ist das sehr allgemein. Wir betrachten hier aber nur Fälle, wo Θ eine offene Teilmenge im $\mathbb{R}^p$ ist.

In der "üblichen" (frequentistischen) Statistik ist θ unbekannt, aber fest. Replikate von θ machen meist keinen Sinn. Die Prinzipien für Tests und Vertrauensintervalle in diesem Rahmen sollten aus andern Vorlesungen bekannt sein. Die Bayes'sche Auffassung von Wahrscheinlichkeit ist allgemeiner : Sie fasst Wahrscheinlichkeit auf als eine subjektive Einschätzung von Unsicherheit, die auch in nicht reproduzierbaren Situationen Sinn macht. Damit können alle Grössen, deren Wert wir (noch) nicht kennen, insbesondere auch θ, als Zufallsvariablen aufgefasst werden. Die Wahrscheinlichkeit $\mathrm{P}\left[\theta \in A\right]$ ($A \subset \Theta$) drückt dann aus, wie plausibel für jemanden die Behauptung "θ liegt in A" ist.

Die Hauptfrage der Bayes'schen-Statistik lautet dann: Wie ändern sich Einschätzungen über θ aufgrund von Daten ? Wenn jemand a priori, d.h. bevor Daten vorliegen, eine bestimmte Wahrscheinlichkeitsverteilung für θ hat, wie sollte er dann seine Einschätzungen modifizieren im Lichte der Daten ? Es stellt sich heraus, dass es bei dieser Modifikation keine Freiheit oder Subjektivität mehr gibt, sofern man Widersprüche bei der Einschätzung von Unsicherheiten vermeiden will. Der Übergang von der a-priori zur a-posteriori Verteilung muss gemäss der Bayes'schen Formel erfolgen.

In mathematischer Formulierung bedeutet dies das Folgende: Die Dichte $p_\theta(\mathbf{x})$ wird als bedingte Dichte der Beobachtung X gegeben den Wert von θ aufgefasst und die a priori Dichte $\alpha(\theta)$ als die Randdichte von θ. Das Paar $(\theta, \mathbf{X})$ ist also ein Zufallsvektor auf $\mathbb{R}^{n+p}$ mit gemeinsamer Dichte

$$\alpha(\theta)p_\theta(\mathbf{x}).$$

Gemäss der Bayes-Formel im vorangegangenen Abschnitt ist die bedingte Dichte von θ gegeben $\mathbf{X} = \mathbf{x}$ daher gleich

$$\alpha(\theta|\mathbf{x}) = \frac{\alpha(\theta)p_\theta(\mathbf{x})}{\int_\Theta \alpha(\theta')p_{\theta'}(\mathbf{x})d\theta'}.$$

Beispiel 1.7. *Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}\left(\theta, \sigma^2\right)$, wobei σ^2 bekannt ist und θ unbekannt. Damit ist also*

$$p_\theta(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right).$$

Als a-priori Verteilung für θ wählen wir eine $\mathcal{N}\left(\xi, \kappa^2\right)$-Verteilung, das heisst

$$\alpha(\theta) = \frac{1}{\sqrt{2\pi}\kappa}\exp\left(-\frac{1}{(2\kappa^2)}(\theta - \xi)^2\right)$$

Wie man in praktischen Anwendungen die "Hyperparameter" ξ und κ bestimmen kann, ist eines der Hauptprobleme der Bayes-Statistik, das wir hier nicht näher diskutieren.

Damit ist die gemeinsame Dichte

$$\alpha(\theta, \mathbf{x}) = \frac{1}{(2\pi)^{\frac{n+1}{2}}\sigma^n\kappa} \exp\left(-\frac{1}{2\kappa^2}(\theta-\xi)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)$$

Die bedingte Dichte von θ gegeben $\mathbf{X} = \mathbf{x}$ ist proportional zu $\alpha(\theta, \mathbf{x})$, und wir können alle Terme ignorieren, die θ nicht enthalten. Also folgt mit quadratischem Ergänzen, dass

$$\begin{aligned}\alpha(\theta|\mathbf{x}) &\propto \exp\left(-\frac{1}{2\kappa^2}(\theta-\xi)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\\ &\propto \exp\left(-\frac{1}{2\nu^2}\left(\theta - \frac{\nu^2}{\kappa^2}\xi - \frac{\nu^2}{\sigma^2}n\bar{x}\right)^2\right)\end{aligned}$$

Dabei ist ν^2 definiert durch

$$\frac{1}{\nu^2} = \frac{1}{\kappa^2} + \frac{n}{\sigma^2}, \quad \textit{d.h. } \nu^2 = \frac{\kappa^2\ \sigma^2}{n\kappa^2 + \sigma^2}.$$

Der letzte Ausdruck in der vorangegangenen Formel ist bis auf eine Konstante die Dichte einer Normalverteilung. Das heisst, die a-posteriori Verteilung ist wieder eine Normalverteilung, und zwar mit Erwartungswert

$$\mu(\mathbf{x}) = \frac{\nu^2}{\kappa^2}\xi + \frac{\nu^2 n}{\sigma^2}\bar{x} = \frac{\sigma^2}{\sigma^2+n\kappa^2}\cdot\underbrace{\xi}_{\textit{a-priori EW}} + \frac{n\kappa^2}{\sigma^2+n\kappa^2}\cdot\underbrace{\bar{x}}_{\textit{MLE von }\theta}$$

und Standardabweichung ν. Der a-posteriori Erwartungswert ist also eine konvexe Kombination des a-priori Erwartungswerts und des Maximum-Likelihood Schätzers von θ, d.h. ein Kompromiss zwischen diesen beiden Grössen. Die Gewichtung bedeutet, dass wir auf die a-priori Verteilung vertrauen und den Daten nur wenig Gewicht geben, falls κ^2 klein ist verglichen mit σ^2/n. Falls aber κ^2 gross ist (d.h. die a-priori Unsicherheit ist gross), gewichten wir die Daten stärker. Die a-posteriori Standardabweichung ν ist kleiner als die Standardabweichung $\sigma/\sqrt{n}$ des arithmetischen Mittels, was bedeutet, dass die Information aus der a-priori Verteilung ebenfalls zu einer Reduktion der a-posteriori Unsicherheit beiträgt.

Was kann man nun mit der a-posteriori Dichte anfangen ? Wenn man an einer Punktschätzung für θ interessiert ist, wird man ein Lagemass wie Erwartungswert oder Median der a-posteriori Verteilung verwenden. Für ein $(1-\alpha)$-Vertrauensintervall wird ein Bayesianer einfach ein Intervall mit der a-posteriori Wahrscheinlichkeit $1-\alpha$ angeben. In obigem Beispiel gilt etwa

$$\mathrm{P}\left[\theta \in \mu(\mathbf{x}) \pm \Phi^{-1}(1-\alpha/2)\nu \mid \mathbf{X} = \mathbf{x}\right] = 1-\alpha,$$

das heisst, im Licht der Daten sind wir 95%-sicher, dass θ im obigen Intervall liegt. Im Unterschied zum frequentistischen Intervall ist hier also θ zufällig und die Beobachtung $\mathbf{x}$ fest.

Die a-posteriori Verteilung erlaubt es auch, die Unsicherheit betrefffend den Wert von θ in Prognosen für neue Beobachtungen $\mathbf{Y}$ zu integrieren. Wenn $\mathbf{Y}$ bei bekanntem θ von

$\mathbf{X}$ unabhängig ist mit Dichte $g_\theta(\mathbf{y})$, dann ist die Wahrscheinlichkeit, dass $\mathbf{Y}$ in B liegt, gestützt auf die Beobachtungen $\mathbf{X} = \mathbf{x}$ gleich

$$\mathrm{P}\left[\mathbf{Y} \in B \mid \mathbf{X} = \mathbf{x}\right] = \int \mathrm{P}\left[\mathbf{Y} \in B \mid \theta\right] \alpha(\theta \mid \mathbf{x}) d\theta = \int_B \int_\Theta g_\theta(\mathbf{y}) \alpha(\theta \mid \mathbf{x}) d\theta d\mathbf{y}.$$

Dieses Vorhersageintervall berücksichtigt die Unsicherheit über θ, indem nicht einfach ein Schätzwert für θ eingesetzt wird, sondern über die möglichen Werte von θ gemäss der a-posteriori Verteilung gemittelt wird. Wenn in obigem Beispiel Y eine zukünftige Beobachtung ist, welche ebenfalls $\mathcal{N}\left(\theta, \sigma^2\right)$-verteilt ist, dann gilt also

$$\mathrm{P}\left[Y \leq b \mid \mathbf{X} = \mathbf{x}\right] = \int_{-\infty}^{b} \int_{\mathbb{R}} \frac{1}{\sigma^2} \phi\left(\frac{y-\theta}{\sigma}\right) \frac{1}{\nu} \phi\left(\frac{\theta - \mu(\mathbf{x})}{\nu}\right) d\theta dy.$$

Da die Faltung zweier Normalverteilungsdichten wieder normal ist, folgt also, dass

$$\mathrm{P}\left[Y \leq b \mid \mathbf{X} = \mathbf{x}\right] = \Phi\left(\frac{b - \mu(\mathbf{x})}{\sqrt{\sigma^2 + \nu^2}}\right).$$

Das obige Beispiel ist untypisch in zwei Aspekten: Erstens ist die a-posteriori Verteilung eine Standardverteilung und wir können deren Kennzahlen explizit berechnen, und zweitens hat der unbekannte Parameter θ nur eine Komponente.

Wenn wir etwa in obigem Beispiel als a-priori Verteilung eine Cauchy-Verteilung wählen, gehört die a-posteriori Verteilung nicht mehr zu einer bekannten Verteilungsfamilie und wir können keine Kennzahlen mehr berechnen. Wir können aber immer noch die a-posteriori Dichte bis auf eine Konstante plotten und so die wesentlichen Eigenschaften ersehen. Sobald jedoch der Parameter drei oder mehr Komponenten hat, wird die Sache schwieriger. Die Berechnung von Kennzahlen, von Randdichten und Vertrauensintervallen einzelner Komponenten oder von Vorhersageintervallen zukünftiger Beobachtungen erfordert Integration. Dazu bietet sich Simulation als Alternative zu numerischen Verfahren oder asymptotischen Näherungen an. Die Bayes'sche Statistik hat die Entwicklung und Untersuchung von Methoden zur Simulation gemäss hochdimensionalen Verteilungen, die nicht zu einer Standardfamilie gehören, in den letzten 10 Jahren sehr stark vorangetrieben. Einen kleinen Einblick werden wir im Kapitel 4 geben.

1.7 Simulation in der statistischen Mechanik

Wir betrachten ein sogenanntes Spinsystem. Das sind magnetische Teilchen, welche auf einem Gitter $L = \{1, 2, \ldots, n\}^d$ angeordnet sind und einen Spin ± 1 haben. Die möglichen Zustände (Konfigurationen) des Systems sind also $\mathbf{x} \in \{\pm 1\}^L$. Aus physikalischen Gründen nimmt man eine sogenannte *Gibbsverteilung* für $\mathbf{x}$ an:

$$p(\mathbf{x}) \;=\; \frac{1}{\text{Normierung}} \exp\left(-\frac{1}{T} \sum_{i \neq k} J_{ik} x_i x_k\right).$$

Dabei ist T die absolute Temperatur und $J_{ik} = J_{ki}$ die Interaktion zwischen den Teilchen an den Plätzen i und k. Wenn $x_i = x_k$, dann erhalten wir einen Faktor $\exp(-J_{ik})$, und wenn $x_i \neq x_k$ einen Faktor $\exp(J_{ik})$. Wenn also $J_{ik} < 0$, werden gleiche Spins an den Plätzen i und k bevorzugt, sonst entgegengesetzte. Wenn alle $J_{ik} \leq 0$ sind, spricht man

von *ferromagnetischer Interaktion*. Offensichtlich ändert sich die Verteilung nicht, wenn alle Spins umgekehrt werden, d.h. $p(\mathbf{x}) = p(-\mathbf{x})$. Wenn alle $J_{ik} = 0$, dann haben wir einfach die Gleichverteilung.

In der Physik ist $\sum_{i \neq k} J_{ik} x_i x_k$ die Energie der Konfiguration (Konfigurationen hoher Energie haben niedrige Wahrscheinlichkeit). Die Temperatur T im Nenner macht die Verteilung flacher für grosses T, d.h. bei hoher Temperatur kann das System leichter einen Zustand hoher Energie annehmen. Umgekehrt ist im Grenzwert $T \to 0$ die Verteilung konzentriert auf die Zustände niedrigster Energie.

Der einfachste Spezialfall, bei dem bereits interessante Phänomene auftreten, ist das sogenannte Ising-Modell, bei dem

$$J_{ik} = \begin{cases} 0 & \text{falls } \|i-k\| \neq 1 \\ -1 & \text{falls } \|i-k\| = 1. \end{cases}$$

Man interessiert sich zum Beispiel dafür, wie die Verteilung der mittleren Magnetisierung

$$M_n = \frac{1}{n^d} \sum_{i \in L} X_i$$

aussieht für grosses n. Dies ist sehr schwierig analytisch zu beantworten. Direkte Berechnung ist ausgeschlossen, da man über 2^{n^d} Terme summieren müsste. Für das Ising-Modell kann man zeigen, dass in einer Dimension $\sqrt{n} M_n$ für alle T's asymptotisch normalverteilt ist mit Erwartungswert 0. Für $d = 2$ existiert eine kritische Temperatur T_c, so dass für $T > T_c$ nM_n ebenfalls normalverteilt ist mit Erwartungswert 0 (die Skalierung n statt $\sqrt{n}$ kommt daher, dass wir jetzt n^2 und nicht mehr n Teilchen haben), während für $T < T_c$ M_n gegen $\frac{1}{2}(\delta_{\mu^+} + \delta_{\mu^-})$ konvergiert. Dabei ist δ_{μ^+} das Diracmass im Punkt $\mu^+ > 0$ und $\mu^- = -\mu^+$. Das heisst, dass es unterhalb einer kritischen Temperatur T_c zu einer spontanen Magnetisierung kommt, bei der entweder die positiven oder die negativen Spins überwiegen. Der Wert von T_c lässt sich ebenfalls angeben. Sobald die Dimension grösser oder die Interaktion komplizierter ist, lassen sich viele Dinge nicht mehr analytisch berechnen, weshalb die Physiker gerne auf Simulationen zurückgreifen.

Wie man gemäss einer Gibbsverteilung simuliert, ist nicht auf den ersten Blick klar. Der Schlüssel dazu ist die Tatsache, dass die bedingten Verteilungen eines Spins X_i gegeben der Rest eine einfache Gestalt haben:

$$\begin{aligned} &\mathrm{P}\left[X_i = +1 | X_k,\, k \neq i\right] \\ &= \frac{\exp(-\frac{1}{T}(\sum_{k \neq i} J_{ik} x_k + \sum_{\ell \neq k \neq i} J_{\ell k} x_\ell x_k))}{\exp(-\frac{1}{T}(\sum_{k \neq i} \cdots + \sum_{\ell \neq k \neq i} \cdots)) + \exp(+\frac{1}{T}(\sum_{k \neq i} \cdots - \sum_{\ell \neq k \neq i} \cdots))} \\ &= \frac{\exp(-\frac{1}{T} \sum_{k \neq i} J_{ik} x_k)}{\exp(-\frac{1}{T} \sum_{k \neq i} J_{ik} x_k) + \exp(+\frac{1}{T} \sum_{k \neq i} J_{ik} x_k)} \end{aligned}$$

(der letzte Term in den Exponentialfunktionen kürzt sich weg). Diese bedingten Wahrscheinlichkeiten lassen sich also einfach berechnen solange die meisten $J_{ik} = 0$ sind.

Aus der Definition der bedingten Wahrscheinlichkeiten folgt ferner: Wenn $\mathbf{X}$ die Verteilung $p(\mathbf{x})$ hat, dann gilt das auch für die Konfiguration $\mathbf{X}'$, die aus $\mathbf{X}$ entsteht, indem man alle X_k, $k \neq i$, festhält und X_i neu auswürfelt gemäss obiger Verteilung. Der *Gibbs-sampler* beginnt mit irgendeiner Konfiguration und führt diese Operation des Neu-Auswürfelns einer Komponente wiederholt aus, wobei alle Komponenten gemäss einem bestimmten

Plan an die Reihe kommen. Es ist plausibel und man kann auch beweisen, dass dieser Algorithmus im Limes Konfigurationen gemäss $p(\mathbf{x})$ liefert, siehe Kapitel 4. Der Gibbssampler ist auch in der Bayes-Statistik anwendbar, weil es häufig möglich ist, gemäss der bedingten Dichte $\alpha(\theta_i \mid \theta_j, j \neq i, \mathbf{x})$ zu simulieren.

1.8 Simulation im Operations Research

Wir geben zwei Beispiele, "Lagerhaltung einer Firma" und "Warteschlangen".

Beispiel 1.8. *Das betrachtete System sei die Lagerhaltung einer Firma mit verschiedenen Produkten. Diese hängt einerseits ab von der Nachfrage (zu welchen Zeitpunkten wird ein Produkt in welcher Menge verlangt) und andrerseits von der Lagerhaltungspolitik (wie gross ist das Lager, und wann wird nachbestellt) und den Lieferfristen (wie lange dauert es von der Bestellung zur Auslieferung ans Lager). Nachfrage und Lieferfristen werden meist stochastisch modelliert. Das Ziel ist die Optimierung der Lagerhaltungspolitik im Hinblick auf Kosten und Wahrscheinlichkeit, dass ein verlangter Artikel am Lager ist.*

Beispiel 1.9. *Bei einer Warteschlange hat man einen oder mehrere Schalter, an denen Kunden warten, bis sie bedient werden. Das Verhalten einer Warteschlange hängt davon ab, was die Ankunftszeiten und die Bedienungszeiten der Kunden sind, in welcher Reihenfolge die Schalter besucht werden und wie die Bedienungsstrategie an den einzelnen Schaltern ist. Ankunfts- und Bedienungszeiten werden praktisch immer als stochastisch modelliert. In Abbildung 1.9 ist eine Warteschlange mit einem Schalter und der "first in, first out" Bedienungsstrategie dargestellt.*

Bei diesen Beispielen enthält das Modell eine Zeitkomponente: der Zustand des Systems verändert sich zufällig mit der Zeit, wir haben einen stochastischen Prozess. Der Zustand des Systems ändert sich jedoch nicht kontinuierlich, sondern zu diskreten (zufälligen) Zeitpunkten. Man spricht auch von "discrete event simulation". Wegen dieser diskreten Natur hat man nur abzählbar viele Input-Grössen.

Die Zielgrössen wie Lieferfrist oder die Länge der Schlange kann man jedoch für beliebigen Zeitpunkt betrachten. Typischerweise betrachtet man aber Modelle, wo sich wenigstens asymptotisch ein stationärer Zustand einstellt, und man will dann die Verteilung der interessierenden Grössen in diesem stationären Zustand bestimmen. Diese Verteilung kann man in der Regel so bestimmen, dass man eine einzige Realisierung des stochastischen Prozesses (einen ganzen zeitlichen Verlauf des Systems) erzeugt, und dann über die Zeit mittelt. Die Terme bei der Mittelung sind dann aber nicht unabhängig, was bei der Abschätzung der Genauigkeit berücksichtigt werden muss. Im Kapitel 4 werden wir ebenfalls Simulationen mit abhängigen Realisierungen antreffen, allerdings aus andern Gründen. Die Methoden zur Abschätzung der Genauigkeit, die wir in Abschnitt 4.3 besprechen, sind jedoch die Gleichen.

Im Gegensatz zu den sprunghaften Zustandsänderungen in obigen Beispielen gibt es auch Fälle, wo sich der Zustand kontinuierlich ändert. In deterministischen Systemen ist das sogar die Regel, da diese fast ausschliesslich als gewöhnliche oder partielle Differentialgleichungen formuliert werden. Wenn man gewöhnliche Differentialgleichungen durch Rauschen stört, kommt man auf stochastische Differentialgleichungen, bei denen Simulation ebenfalls eine weit verbreitete Methode zur Lösung ist. Wir geben in Abschnitt 3.8.1 eine ganz kurze Einführung in dieses Thema.

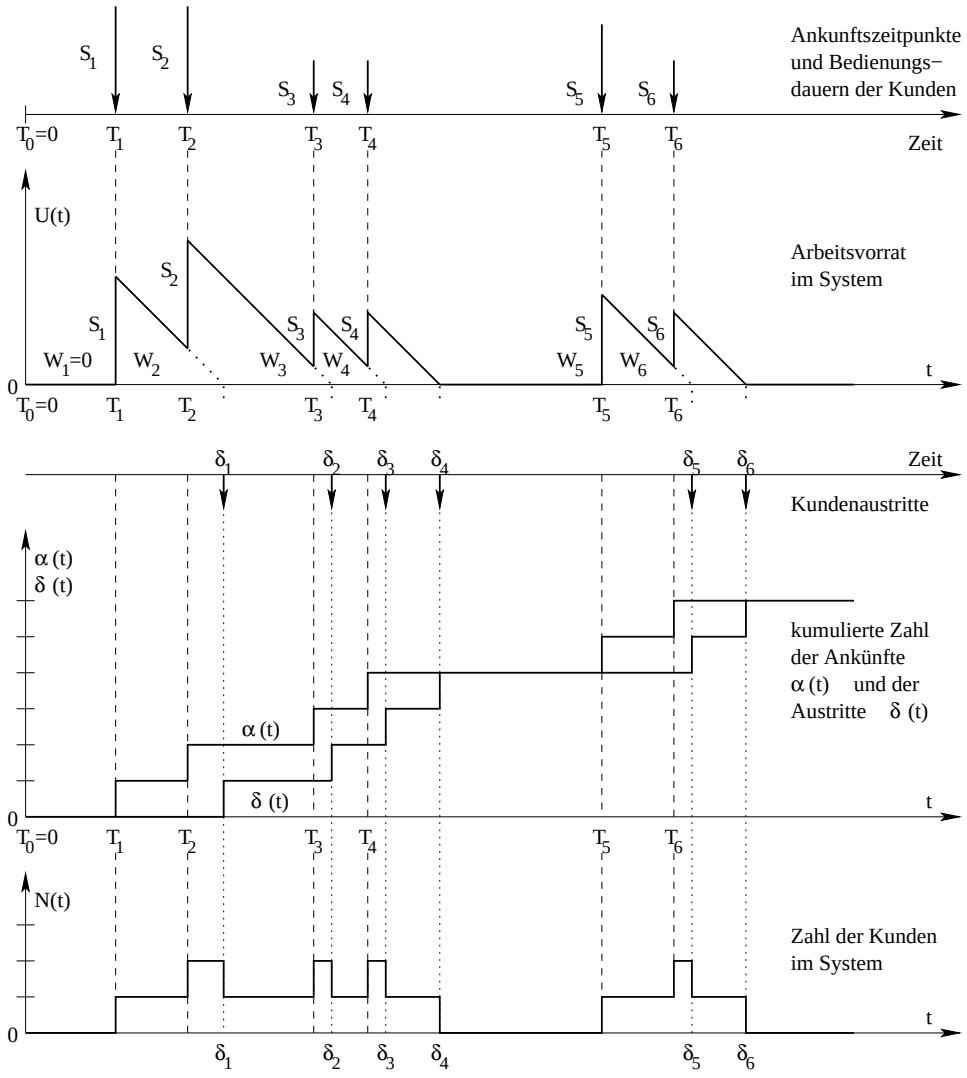

Abbildung 1.9: Graphische Simulation einer Warteschlange mit einem Schalter und der "first in, first out" Bedienungsstrategie (aus Warteschlangen-Modelle, WS 1997/98, K. Hazeghi).

1.9 Simulation in der Finanzmathematik

Simulation ist heutzutage eine sehr weit verbreitete Methode in der Finanzmathematik. Wir betrachten ein einfaches Modell für das Risiko bei einem Portfolio bestehend aus N

Krediten: Der gesamte Verlust L (in einem vorgegebenen Zeitraum) ist

$$L = \sum_{i=1}^{N} Y_i \ell_i,$$

wobei Y_i der Indikator dafür ist, dass der i-te Kredit im vorgegebenen Zeitraum platzt und ℓ_i der Verlust beim Platzen des i-ten Kredits bezeichnet. Wir nehmen an, dass ℓ_i deterministisch ist und mit Werten in $\{1, 2, \ldots, m\}$ (in geeigneten Einheiten). Die Indikatoren Y_i sind stochastisch, und ihre gemeinsame Verteilung ist von folgendem Typ: Es gibt nicht beobachtbare Variablen $W = (W_1, \ldots, W_p) \sim F$ derart, dass die Y_i gegeben W bedingt unabhängig sind mit

$$\mathrm{P}\left[Y_i = 1 \mid W\right] = f_i(W).$$

Die Variablen W stellen die ökonomische Situation in verschiedenen Ländern oder Branchen dar, welche das Risiko mehrerer Kredite gleichzeitig beeinflussen. Die Verteilung F und die Funktionen f_i sind ebenfalls als bekannt vorausgesetzt, zum Beispiel nimmt man die W_j's als unabhängig und Gamma-verteilt und

$$f_i(W) = \min(1, \sum_{j=1}^{p} a_{ij} W_j)$$

mit geeigneten nicht-negativen Gewichten a_{ij}.

Das Ziel ist es, die Verteilung von L zu bestimmen. Wegen der Annahmen an ℓ_i ist L ebenfalls diskret, d.h. man braucht $\mathrm{P}\left[L = n\right]$ für $n = 0, 1, 2, \ldots, \sum \ell_i$. Im Prinzip kann man direkt Realisierungen von L erzeugen, indem man zuerst $W_1, \ldots, W_p$ erzeugt, und dann die bedingte Unabhängigkeit der Y_i's benutzt. Weil N sehr gross ist, wird dies aber ziemlich ungenau. Es ist daher besser, die Beziehung

$$\mathrm{P}\left[L = n\right] = \mathbf{E}\left[\mathrm{P}\left[\sum_{i=1}^{N} Y_i \ell_i = n \mid W\right]\right]$$

zu benutzen. Die bedingten Wahrscheinlichkeiten im Innern werden dann analytisch approximiert und nur der äussere Erwartungswert mit Simulation bestimmt. Genaueres findet man in S. Merino und M. Nyfeler, "Calculating portfolio loss", RISK, August 2002, 82–86.

Kapitel 2

Erzeugung uniformer Zufallszahlen

Praktisch alle Simulationen verwenden "Zufallszahlen", die nicht wirklich zufällig sind, sondern von einem deterministischen Algorithmus erzeugt werden (Pseudo-Zufallszahlen). Dies scheint zunächst ein Widerspruch zu sein, aber von einem praktischen Standpunkt aus kommt es ja nur darauf an, ob sich die erzeugten Zahlen ähnlich verhalten wie Realisationen von $(U_1, U_2, \ldots)$ wobei U_i i.i.d. $\sim$ Uniform$(0, 1)$. Es gab zwar auch Versuche, die Zufälligkeit gewisser physikalischer Vorgänge auszunützen, aber diese Zahlen hatten stets schlechtere Eigenschaften als die deterministisch erzeugten. Ausserdem ist bei Simulationen auch die Reproduzierbarkeit wichtig, was mit einem deterministischen Algorithmus leichter zu erreichen ist.

Die Frage, was ähnliches Verhalten wie uniforme Zufallszahlen heisst, führt auf tiefe mathematische Theorien, die von Kolmogorov und Martin-Löf entwickelt wurden. Wir verfolgen hier einen pragmatischen Ansatz, der auf visueller Überprüfung und einigen ausgewählten statistischen Tests angewandt auf d-Tupel basiert.

Alle verwendeten Verfahren zur Konstruktion von Pseudo-Zufallszahlen (u_n) sind von folgender Bauart:

$$z_{n+1} = f(z_n), u_n = h(z_n). \tag{2.1}$$

Das heisst, man hat eine (i. A. nicht umkehrbare) Funktion der Elemente einer rekursiven Folge. Wenn die Anzahl M der möglichen Werte für z_n endlich ist, dann sind (z_n) und (u_n) offensichtlich periodisch bis auf ein eventuelles transientes Teilstück am Anfang, und die Periodenlänge ist höchstens gleich M (Betrachte das kleinste n so dass z_n gleich einem Element aus $\{z_0, z_1, \ldots z_{n-1}\}$ ist).

Bei einem guten Generator sollte die Periode wesentlich grösser sein als die Anzahl benötigter Zufallszahlen in einem gegebenen Problem. Eine lange Periode ist aber nur notwendig und nicht hinreichend für einen guten Generator. Wichtig ist vor allem, dass aufeinanderfolgende d-Tupel den d-dimensionalen Einheitswürfel gut ausfüllen. Für Kontrollzwecke sollte es ausserdem möglich sein, mehrere lange Teilstücke zu erzeugen, die möglichst unabhängig sind.

Wir besprechen zunächst ein paar einfache Generatoren, welche später kombiniert werden, um höheren Ansprüchen zu genügen.

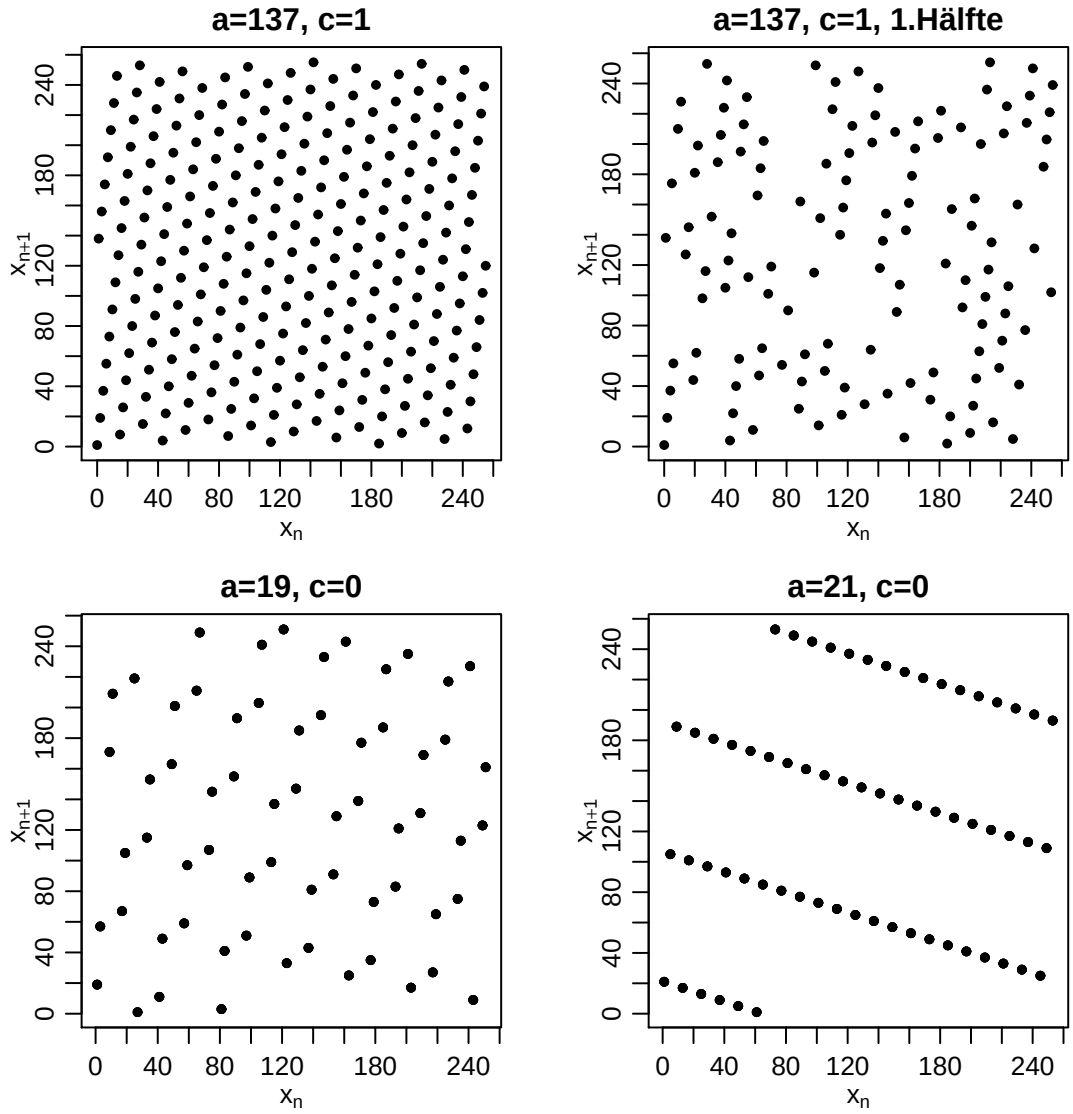

Abbildung 2.1: Paare von aufeinanderfolgenden Werten (x_n, x_{n+1}) von linearen Kongruenzgeneratoren für $M = 256$. Die beiden unteren Figuren illustrieren den Unterschied zwischen $a \mod 8 = 3$ und $a \mod 8 = 5$ im Fall $c = 0$ (vgl. Theorem 2.1).

2.1 Lineare Kongruenzgeneratoren

Ein linearer Kongruenzgenerator ist von der Form $u_n = x_n/M$, wobei (x_n) der Rekursion

$$x_{n+1} \;=\; (ax_n \;+\; c) \mod M$$

genügt mit $x_0, a, c, M \in \mathbb{N}$. In den Abbildungen 2.1 und 2.2 werden Beispiele mit $M = 256$, und in 2.3 Beispiele mit $M = 2048$ gezeigt.

Die Frage nach der Periodenlänge dieser Generatoren wird von folgendem Theorem beantwortet:

Satz 2.1. *Es gilt*

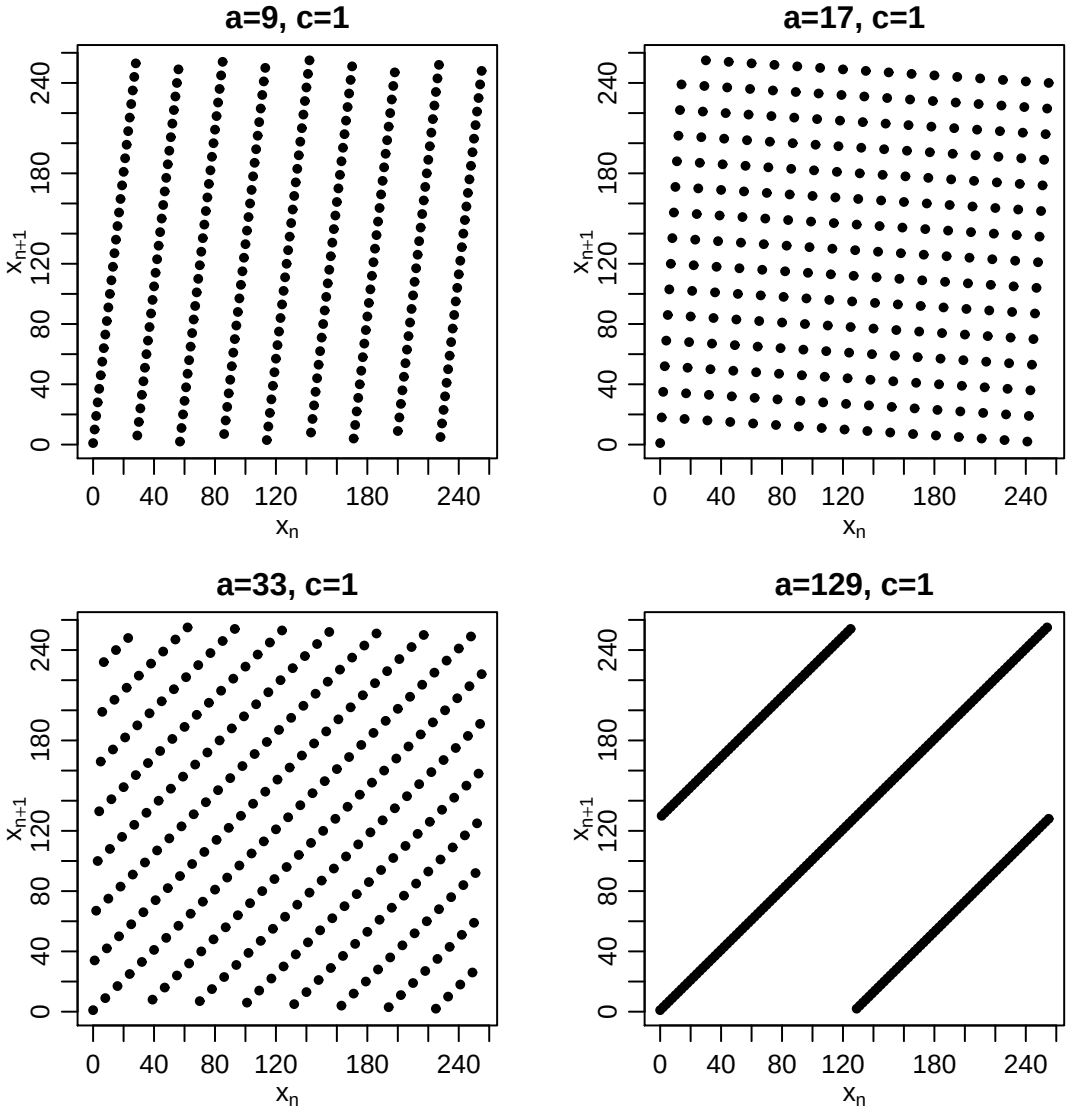

Abbildung 2.2: Paare von aufeinanderfolgenden Werten (x_n, x_{n+1}) von linearen Kongruenzgeneratoren für $M = 256$. Oben rechts ist ein Beispiel eines guten Generators, unten rechts ein krasses Beispiel eines unbefriedigenden Generators.

1. *Falls* $c \neq 0$, *dann ist die Periode* $= M$ *für alle* x_0 *genau dann wenn* c *und* M *teilerfremd sind und* $a \equiv 1 \mod p$ *für alle Teiler* p *von* M, *welche prim oder* $= 4$ *sind.*

2. *Falls* $c = 0$, *dann ist die Periode* $= M - 1$ *für alle* $x_0 \neq 0$ *genau dann, wenn* M *prim ist und* $a^{(M-1)/p} \neq 1 \mod M$ *für alle Primfaktoren* p *von* $M - 1$.

3. *Falls* $c = 0$ *und* $M = 2^k \geq 16$, *dann ist die Periode* $= \frac{M}{4}$ *genau dann wenn* x_0 *ungerade und* $a \mod 8 = 5$ *oder* 3.

4. *Wenn* $c = 0$, $M = 2^k \geq 16$ *und* $a \mod 8 = 5$, *dann ist* $x_n \mod 4$ *konstant* $=: b$, *und wenn* $b \in \{1, 3\}$, *dann ist* $\frac{1}{4}(x_n - b)$ *gerade das Resultat des Generators mit* $a' = a$, $c' = b\frac{a-1}{4}$, $M' = \frac{M}{4}$. *(Man soll also einfach die letzten zwei Bits ignorieren).*

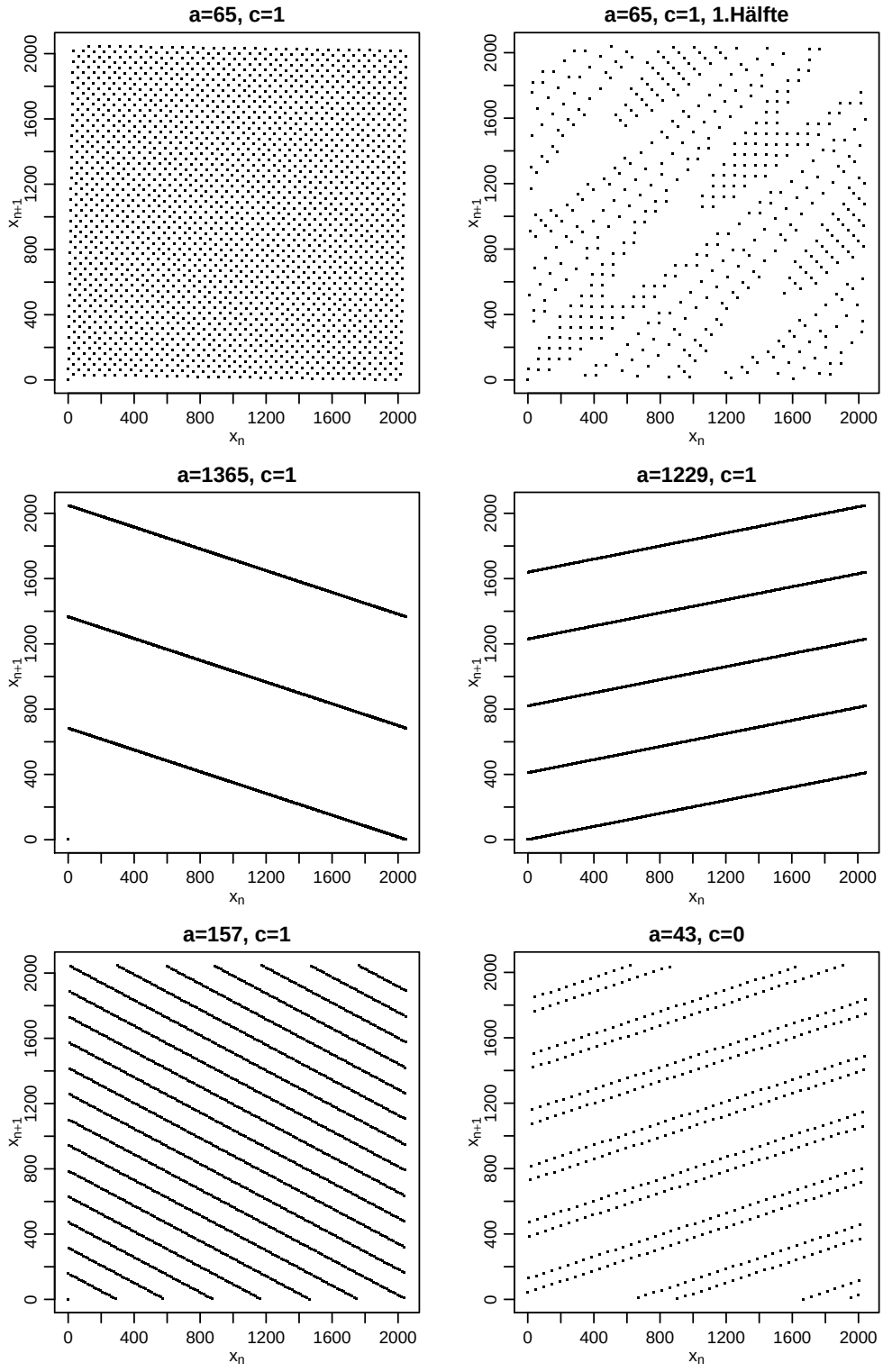

Abbildung 2.3: Weitere Beispiele zu linearen Generatoren für $M = 2048$. Die Beispiele stammen aus Ripley (1997).

Beweis. Der Beweis benutzt Resultate der Zahlentheorie (kleiner Satz von Fermat), siehe Ripley (1987) Abschnitt 2.7, oder Knuth (1998) Theoreme A und C in Abschnitt 3.2. □

Der Fall $c \neq 0$ gibt also das einfachste Kriterium, dafür hat der Generator den Nachteil, dass der Wert Null auftritt. Für $M = 2^k$ ist die modulo Operation besonders einfach auf dem Computer, dafür haben die niedrigen bits eine kurze Periode. Für $M = 2^k - r$ mit r klein, ist die modulo Operation ebenfalls relativ einfach, aber die Bedingung für a (Fall 2) ist aufwändiger. In jedem Fall gibt es aber für festes M sehr viele Wahlen von a (und ev. c) mit langer Periode. Wie die Figuren zeigen, unterscheiden sich diese Wahlen aber in anderer Hinsicht.

Wir schauen uns daher die Verteilung von d-Tupeln an. Sei Λ_d, die Menge aller d-Tupel, welche der Generator produziert. Im Fall $c = 0$ fügen wir noch den Ursprung dazu:

$$\Lambda_d = \{(x_n, x_{n+1}, \ldots, x_{n+d-1}),\ 0 \leq n < M\} \cup \underbrace{\{(0, \ldots, 0)\}}_{\text{falls c=0}}.$$

Wenn wir annehmen, dass die Periode maximal ist, hat diese Menge M Punkte in $\{0, \ldots, M-1\}^d$. Bei einem guten Generator sollte jeder Punkt nach Skalierung mit $\frac{1}{M}$ im Mittel einen Würfel mit Volumen $\frac{1}{M}$, d.h. Seitenlänge $M^{-1/d}$, abdecken.

Die Figuren legen die Vermutung nahe, dass die Punkte eines linearen Kongruenzgenerators auf einem Gitter liegen. Ein Gitter $L \subset \mathbb{R}^d$ besteht aus allen Punkten der Form

$$\{x = t_1 g_1 + \ldots t_d g_d;\ t_i \in \mathbb{Z}\}$$

mit festen, linear unabhängigen Vektoren g_i. Die g_i's heissen erzeugende Vektoren. (Man beachte, dass die erzeugenden Vektoren nicht eindeutig bestimmt sind).

Satz 2.2. *Betrachte das Gitter L_d mit erzeugenden Vektoren*

$$\begin{aligned} g_1 &= (1, a, a^2, \ldots, a^{d-1})^T, \\ g_j &= (0, \ldots, \underbrace{M}_{j}, \ldots, 0)^T \quad (j = 2, 3, \ldots, d). \end{aligned}$$

Falls $c > 0$ und die Periode $= M$ ist oder $c = 0$ und die Periode $= M - 1$, dann ist

$$\Lambda_d = \Big(c(0, 1, 1 + a, \ldots, (1 + a + \cdots + a^{d-2}))^T + L_d\Big) \cap \{0, \ldots, M-1\}^d.$$

Beweis. Man zeigt mit Induktion, dass

$$x_{n+j} = (a^j x_n + c(a^{j-1} + \cdots + 1) + M \cdot \mathbb{Z}) \cap \{0, 1, \ldots, M-1\}.$$

Damit gehört

$$(x_n, \ldots, x_{n+d-1})^T - c(0, 1, \ldots, (1 + \cdots + a^{d-2}))^T$$

offensichtlich zu L_d. Weil die Periode als maximal angenommen wurde, kommt jeder Wert für x_n einmal vor (ausser 0 falls $c = 0$). Ferner gehört zu gegebenem t_1 genau ein Punkt von L_d zu $\{0, \ldots, M-1\}^d$. Daraus folgt die behauptete Gleichheit der Mengen. □

Insbesondere sieht man, dass die Qualität eines Generators nicht von c abhängt, weil dadurch nur das Gitter verschoben wird.

Die Punkte eines Gitters L liegen auf parallelen, gleichabständigen Hyperebenen, wie es in den Figuren klar zu erkennen ist. Die möglichen Scharen solcher Hyperebenen sind gegeben durch das sogenannte duale (oder reziproke) Gitter $L^\perp$:

$$\{v \in \mathbb{R}^d;\ v \cdot x \in \mathbb{Z}\ \ \forall x \in L\}.$$

(Man kann sich davon überzeugen, dass dies wieder ein Gitter ist).

Aus der speziellen Form des Gitters L_d im vorigen Satz folgt sofort, dass jedes $v \in L_d^\perp$ die Form $v = \frac{1}{M}(t_1, \ldots, t_d)$ hat mit $t_i \in \mathbb{Z}$ und

$$t_1 + at_2 + a^2 t_3 + \cdots + a^{d-1} t_d = 0 \mod M. \tag{2.2}$$

Der Abstand zwischen zwei benachbarten Hyperebenen der gleichen Schar ist gleich $\frac{1}{\|v\|}$, und bei einem guten Generator sind die Abstände klein für alle möglichen v. Damit hat man ein Mass ν_d für die Qualität eines Generators

$$\nu_d := M \min\{\|t\| \,;\, \|t\| \neq 0,\ t \in \mathbb{Z}^d,\ \text{und } t \text{ erfüllt (2.2)}\}$$

Je grösser also ν_d, desto besser der Generator.

Leider existiert keine Formel, die ν_d als Funktion von a und M ausdrückt, aber es existieren Algorithmen zur Berechnung von ν_d, siehe Ripley (1987) oder Knuth (1998). Typischerweise wählt man die Periode M gross und so, dass der Generator schnell berechnet werden kann, und sucht dann nach dem Prinzip von Versuch und Irrtum ein a mit maximaler Periode, welches möglichst grosse Werte von ν_d liefert für $d \leq 10$ oder $d \leq 20$.

Die obigen Betrachtungen gelten für überlappende d-Tupel. Wenn d und M teilerfremd sind, erhält man die gleichen Punkte, wenn man die Periode d Mal durchläuft. Ansonsten erhält man ein Untergitter.

2.2 Andere Generatoren

Wegen der Gitterstruktur und der für manche Anwendungen zu kurzen Periode gibt man sich nicht mit linearen Kongruenzgeneratoren zufrieden. Es gibt eine ganze Vielfalt von weiteren Vorschlägen, die alle in den allgemeinen Formalismus passen, wenn man die Grösse z_n geeignet definiert.

Nichtlineare Kongruenzgeneratoren Eine offensichtliche Variante ist $x_i = g(x_{i-1})$ mit $g : \{0, 1, \ldots M-1\} \to \{0, 1, \ldots M-1\}$ nichtlinear, z.B. $g(x) = (ax^2 + bx + c) \mod M$ oder $x \cdot g(x) = 1 \mod M$ mit M prim. Diese Art von Generatoren vermeiden zwar die Gitterstruktur der d-Tupel, aber der Rechenaufwand wird gross. Zudem hat man immer noch höchstens M verschiedene d-Tupel. Daher haben sich diese Art von Generatoren nicht wirklich durchgesetzt.

Schieberegister-Generatoren Hier geht man aus von einer binären Folge $b_n \in \{0, 1\}$, welche die Rekursion

$$b_n = b_{n-p} + b_{n-q} \mod 2$$

erfüllt. Teilstücke der Länge L im Abstand t stellen dann die Elemente der Folge u_n als Dualbruch dar:

$$u_n = \sum_{j=1}^{L} b_{nt+j} 2^{-j}.$$

(siehe Abb. 2.4). Wenn man Überlappungen vermeiden will, wählt man einfach $t = L$. Die verwendeten bits müssen auch nicht nebeneinander liegen, d.h. man kann auch

$$u_n = \sum_{j=1}^{L} b_{n+d_j} 2^{-j}$$

verwenden.

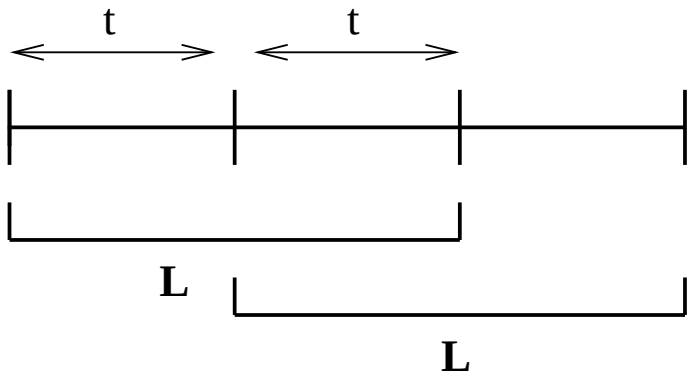

Abbildung 2.4: Schieberegister-Generator.

Die maximale Periode von (b_n) ist $2^p - 1$, und man kann p und q so wählen, dass diese Periode erreicht wird.

Lagged Fibonacci Hier verwendet man eine Rekursion der Form

$$x_i = F(x_{i-p}, x_{i-q}).$$

Dies ist analog wie beim Schieberegister-Generator, aber die x_i müssen nicht mehr binär sein, und F ist beliebig. Zum Beispiel wählt man:

$$F(X_{i-p}, X_{i-q}) = (X_{i-p} + X_{i-q}) \mod M$$

oder wenn die X_i Zahlen im Zweiersystem sind, die logische Operation "Ausschliessendes Oder" für jede Ziffer.

Multiplikation mit Übertrag Sei $z_i = (x_i, c_i)$ mit $x_i = (ax_{i-1} + c_{i-1}) \mod M$ und $c_i = \left\lfloor \frac{aX_{i-1}+c_{i-1}}{M} \right\rfloor$ (ganzzahlige Division). Man schaut also nicht nur den Rest an, sondern auch noch, wieviel mal man teilen kann. Bezüglich der Periode gilt: Wenn $M = 2^k$ und

a so, dass sowohl $aM-1$ als auch $(aM-2)/2$ prim sind, dann ist die Periode gleich $(aM-2)/2$.

Die Implementation ist sehr einfach für $M=2^k$, $a<M$ und $c_0 \le a$. Dann ist $c_n \le a$ für alle n und $aX_{n-1}+c_{n-1} \le aM \le M^2$. Also wenn X_i im Binärsystem geschrieben wird, hat $aX_{n-1}+c_{n-1}$ höchstens $2k$ Stellen, die letzten k Stellen von $aX_{n-1}+c_{n-1}$ sind gerade x_n und die ersten k Stellen von $aX_{n-1}+c_{n-1}$ ergeben c_n.

2.3 Kombination von Generatoren

Alle raffinierteren Generatoren, die verwendet werden, kombinieren mehrere Grundgeneratoren. Damit vergrössert man nicht nur die Periode, sondern man erhält typischerweise auch eine gleichmässigere Verteilung der d-Tupel. Es gibt mindestens zwei Vorschläge zur Kombination von Generatoren:

Die erste Variante kombiniert elementweise mit einer binären Operation. Der erste Generator erzeuge (x'_i) und der zweite (x''_i). Der kombinierte Generator ist dann von der Form $x_i = F(x'_i, x''_i)$ für gegebenes F, z.B. Addition modulo M wenn x'_i und x''_i Werte in $\{0,1,2,\ldots,M-1\}$ haben, oder bitweise Addition modulo 2 wenn x'_i und x''_i im Zweiersystem geschrieben werden.

Dann ist die Periode von (x_i) $\le$ kgV der beiden Perioden. Das folgende Resultat erklärt, warum die Kombination mindestens keine Verschlechterung der Eigenschaften bringt.

Lemma 2.1. *Seien X und Y zwei unabhängige Zufallsvariablen mit Werten in $\{0,1,2,\ldots,M-1\}$, und sei F eine Abbildung von $\{0,1,2,\ldots,M-1\}^2$ nach $\{0,1,2,\ldots,M-1\}$, so dass sowohl $F(x,a)=b$ als auch $F(a,y)=b$ eine Lösung haben für beliebiges a,b. Dann gilt*

$$\sum_b |\,\mathrm{P}\,[F(X,Y)=b] - \frac{1}{M}| \;\le\; (1-\min_b \mathrm{P}\,[Y=b]) \sum_b |\,\mathrm{P}\,[X=b]-\frac{1}{M}|,$$
$$\sum_b |\,\mathrm{P}\,[F(X,Y)=b] - \frac{1}{M}| \;\le\; (1-\min_b \mathrm{P}\,[X=b]) \sum_b |\,\mathrm{P}\,[Y=b]-\frac{1}{M}|).$$

Beweis. Die Annahme impliziert, dass die Lösung von $F(a,y)=b$ eindeutig ist (surjektive Abbildungen einer endlichen Menge in sich sind auch injektiv). Wir bezeichnen die Lösung mit $y(a,b)$. Wenn wir $Z=F(X,Y)$ setzen, dann ist

$$\mathrm{P}\,[Z=b \mid X=a] = \mathrm{P}\,[F(a,Y)=b] = \mathrm{P}\,[Y=y(a,b)] \ge \min_b \mathrm{P}\,[Y=b]$$

. Weil auch $F(x,a)=b$ für beliebiges a und b eine Lösung hat, gilt ferner

$$\sum_a \mathrm{P}\,[Z=b \mid X=a]\,\frac{1}{M} = \sum_a \mathrm{P}\,[Y=y(a,b)]\,\frac{1}{M} = \sum_a \mathrm{P}\,[Y=a]\,\frac{1}{M} = \frac{1}{M},$$

das heisst, wenn X gleichverteilt ist, dann ist auch Z gleichverteilt. Damit folgt das Lemma aus einem allgemeineren Lemma über Markovketten, siehe Abschnitt 4.3.1.

□

Die Verteilung von $F(X,Y)$ ist also näher bei der Gleichverteilung als die Verteilungen von X bzw. von Y. Dies lässt sich auch anwenden auf die Verteilung der d-Tupel.

Die zweite Kombinationsmöglichkeit ist Mischen (Shuffling) der Folge (x'_i), die vom ersten Generator erzeugt wurde, mit Hilfe der Folge (x''_i). Man beginnt mit dem Vektor $t = (x'_1, x'_2, \ldots, x'_k)$. Im n-ten Schritt erzeugt man mit x''_n einen zufälligen Index $i_n \in \{1, \ldots k\}$, setzt $x_n = t_{i_n}$ und ersetzt dann t_{i_n} durch x'_{n+k}.

2.4 Testen von Zufallszahlen

Jeder Test auf Gleichverteilung auf $[0,1]^d$ kann zum Testen von Generatoren verwendet werden. Der Fall $d = 1$ ist meist nicht interessant, da alle Generatoren diese Tests bestehen.

Insbesondere bietet sich der Chiquadrat-Test an. Die Schwierigkeit ist die Wahl der Klassen. Wenn man für jede Komponente k Klassen bildet, ergibt das bei den d-Tupeln insgesamt k^d Klassen, was sehr rasch wächst. Wenn man mit überlappenden d-Tupeln arbeitet, hat man ausserdem Abhängigkeit, die man bei der Wahl der kritischen Grenze beachten muss.

Um mit diesen Problemen zurechtzukommen, wurde vorgeschlagen, statt der Häufigkeiten aller Klassen einfach die Anzahl Klassen W , in die keine d-Tupel fallen, zu zählen. Das bietet wenig Probleme für den Speicher, und die Berechnung von Erwartungswert und Varianz von W ist dann ein kombinatorisches Problem, das man in einfacheren Fällen lösen kann. Für $d = 2$, $k = 1024$ und eine Folge der Länge $n = 2^{21}$ ist z.B. der Erwartungswert von W 141'909 und die Standardabweichung 290. Solche Tests heissen "Affen-Tests", weil man zählt, wie viele Wörter der Länge d ein Affe, der zufällig n mal auf eine Schreibmaschine mit k Tasten drückt, nicht produziert (Gemäss dem Lemma von Borel-Cantelli produziert ein Affe, der unendlich lange schreibt, sogar die gesammelten Werke von Shakespeare unendlich oft).

Kapitel 3

Direkte Erzeugung von Zufallsvariablen

Wir beschäftigen uns hier mit dem folgenden Problem: Gegeben ist eine Verteilung π auf einem beliebigen Raum $\mathbb{X}$ mit einer σ-Algebra $\mathcal{F}$. Ferner stehen uns uniforme Zufallsvariablen $U_1, U_2, \ldots$ i.i.d. $\sim \text{Uniform}(0,1)$ zur Verfügung. Gesucht ist ein Algorithmus, der uns Variablen $X_1, X_2, \ldots$ liefert, welche i.i.d. und gemäss π verteilt sind. Das Problem, dass bei der Implementation auf dem Computer $U_1, U_2, \ldots$ nur Pseudo-Zufallsvariabeln sind, ignorieren wir dabei.

Wir werden sehen, dass viele Verfahren existieren, wenn π eine Verteilung auf $\mathbb{R}$ ist. In diesem Fall identifizieren wir π meist mit der kumulativen Verteilungsfunktion, die wir F nennen. Die zugehörige Dichte bezeichnen wir mit f. In höheren Dimensionen ist es meist schwierig, direkt gemäss einer vorgegebenen Verteilung zu simulieren. Wie man dann vorgehen kann, wird im nächsten Kapitel besprochen.

Die letzten Abschnitte in diesem Kapitel behandeln die Frage, wie genau die Approximation eines Erwartungswerts

$$\mathbf{E}\left[h(X_i)\right] = \int_{\mathbb{X}} h(x)\pi(dx)$$

durch das arithmetische Mittel

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i)$$

ist. Wir werden auch sehen, wie man mit geeigneten Tricks diese Genauigkeit verbessern kann.

3.1 Quantiltransformation

Sei F eine kumulative Verteilungsfunktion auf $\mathbb{R}$.

Definition 3.1. $F^{-1}(u) = \inf\{x | F(x) \geq u\}$ *mit* $u \in (0,1)$ *heisst Quantilfunktion.*

Satz 3.1. *Falls* $U \sim$ *Uniform*$(0,1)$*, dann gilt* $X = F^{-1}(U) \sim F$*.*

Beweis. Siehe Einführungsvorlesung. □

Beispiel 3.1 (Exponentialverteilung)**.** *Sei $F(x) = 1 - e^{-\lambda x}$. Da F strikt monoton und stetig ist, ist $F^{-1}(u)$ die übliche Umkehrfunktion, also $F^{-1}(u) = -\frac{1}{\lambda}\log(1-u)$. Wenn U uniform verteilt ist, dann auch $1-U$ und somit ist $-\frac{1}{\lambda}\log(U)$ exponential verteilt.*

Beispiel 3.2 (Cauchyverteilung)**.** *Die Dichte und die Verteilungsfunktion lauten*

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x).$$

Also ist $F^{-1}(u) = \tan(\pi(u-0.5))$, wir können eine Cauchy-Variable erzeugen mit dem Tangens einer auf $(-\pi/2, \pi/2)$ uniformen Variablen.

Die Quantiltransformation lässt sich nur anwenden, wenn man die kumulative Verteilungsfunktion explizit berechnen und umkehren kann. Weitere Beispiele sind die Dreiecksverteilung oder die Dichte $f(x) = \alpha x^{\alpha-1}$ $(0 < x < 1)$ (sogenannte Beta$(\alpha, 1)$-Verteilung). Für die Normal- oder die allgemeine Gamma- oder Betaverteilung ist die Methode nicht geeignet (ausser man begnügt sich mit numerischen Näherungen der Quantiltransformation).

Die Methode setzt nicht voraus, dass die Verteilungsfunktion stetig ist.

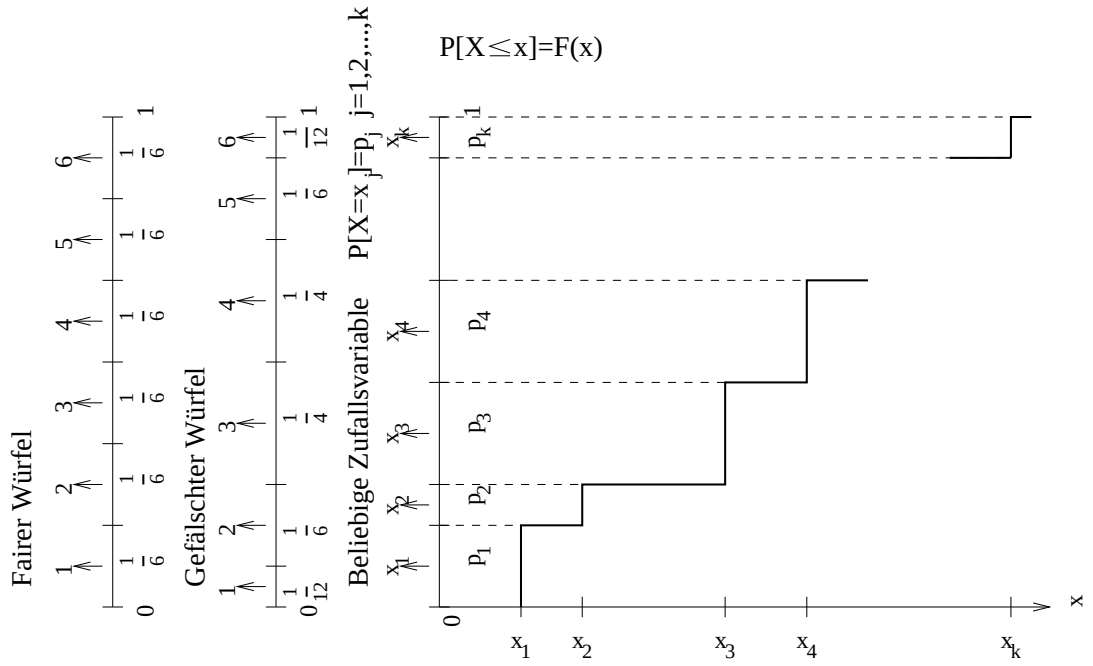

Abbildung 3.1: Illustration von Beispiel 3.3 mit einem fairen und einem gefälschten Würfel.

Beispiel 3.3 (Diskrete Verteilungen)**.** *Die Zufallsvariable X sei diskret und nehme die Werte $x_1 < x_2 < \ldots$ mit den Wahrscheinlichkeiten $p_1, p_2, \ldots$ an. Dann sind F und F^{-1} Treppenfunktionen:*

$$F(x) = \sum_{x_k \le x} p_k, \quad F^{-1}(u) = x_k \text{ für } p_1 + p_2 + \cdots + p_{k-1} < u \le p_1 + p_2 + \cdots + p_k.$$

Wie Figur 3.1 zeigt, ist die Quantiltransformation hier nichts anderes als eine direkte Verallgemeinerung des offensichtlichen Verfahrens zur Simulation eines diskreten Experiments.

Noch eine Bemerkung zum Rechenaufwand: Falls die kumulativen Summen $p_1+p_2+\cdots+p_k$ gespeichert sind, dann muss man nur noch Vergleiche durchführen. Die erwartete Anzahl Vergleiche ist gleich $\sum_k kp_k$, was für $x_k = k$ nichts anderes ist als $\mathbf{E}[X]$. Wir können diesen Aufwand reduzieren durch eine Umnummerierung, so dass $p_1 \geq p_2 \geq p_3 \geq \cdots$. Bei der Poissonverteilung gilt zum Beispiel die folgende Rekursion für die Wahrscheinlichkeiten

$$p_k \;=\; \frac{\lambda}{k} p_{k-1}.$$

Die Wahrscheinlichkeiten sind also maximal für $k \approx \lambda$ und man nimmt am Besten die Anordnung $[\lambda]$, $[\lambda]\pm 1$, $[\lambda]\pm 2, \ldots$. Noch schneller geht es, wenn man den Wertebereich sukzessive halbiert, so dass die Teilbereiche jeweils etwa gleiche Wahrscheinlichkeiten haben, und dann die Vergleiche entsprechend organisiert.

3.2 Verwerfungsmethode

Die Grundidee ist, dass man zuerst nicht nach der vorgegebenen Verteilung π simuliert, sondern nach einer andern Verteilung τ und danach korrigiert. Dies funktioniert im Prinzip auf beliebigen Räumen. Wir nehmen ohne Beschränkung der Allgemeinheit an, dass π und τ Dichten f bzw. g bezüglich eines Bezugsmasses μ haben. In den meisten Anwendungen ist $\mathbb{X}$ entweder diskret oder gleich $\mathbb{R}^d$. Im ersten Fall ist μ dann das Zählmass und die Dichten sind einfach die gewöhnlichen Wahrscheinlichkeiten, im zweiten Fall ist μ das Lebesguemass und f und g sind dann die üblichen Dichten.

Satz 3.2. *Seien π und τ zwei Verteilungen auf einem beliebigen Raum $(\mathbb{X}, \mathcal{F})$ mit Dichten f bzw. g (im Sinne der Masstheorie, bezüglich einem Bezugsmass μ). Ferner sei der Quotient $f(x)/g(x)$ beschränkt durch eine Konstante $M < \infty$, so dass*

$$a(x) := \frac{f(x)}{Mg(x)} \leq 1.$$

Wenn X und U unabhängige Zufallsvariablen sind mit $X \sim \tau$ und $U \sim$ Uniform$(0,1)$, dann ist X gegeben $a(X) \geq U$ π-verteilt, d.h.

$$\mathrm{P}[X \in A \mid U \leq a(X)] \;=\; \pi(A) \quad \forall A \in \mathcal{F}.$$

Beweis. Gemäss der Definition der bedingten Wahrscheinlichkeit gilt

$$\mathrm{P}[X \in A \mid U \leq a(X)] = \frac{\mathrm{P}[\{X \in A\} \cap \{U \leq a(X)\}]}{\mathrm{P}[U \leq a(X)]}$$

Der Zähler ist mit dem Satz der totalen Wahrscheinlichkeit

$$\begin{aligned} &= \mathbf{E}[\mathrm{P}[\{X \in A\} \cap \{U \leq a(X)\} \mid X]] = \int_A \mathrm{P}[U \leq a(x)]\, \tau(dx) \\ &= \int_A a(x)\tau(dx) = \frac{1}{M}\int_A \frac{f(x)}{g(x)} g(x)\mu(dx) = \frac{1}{M}\int_A f(x)\mu(dx) = \frac{1}{M}\int_A \pi(dx). \end{aligned}$$

Analog folgt, dass der Nenner gleich $1/M$ ist. □

Dies führt zu folgendem Algorithmus:

Algorithmus 3.1.

1. *Erzeuge X, U unabhängig mit $X \sim \tau$ und $U \sim$ Uniform$(0,1)$.*
2. *Falls $U \leq f(X)/(Mg(X))$, nehme X als Ergebnis, sonst gehe zurück zu 1.*

Für dieses Verfahren genügt es, die Dichte f nur bis auf eine Konstante zu kennen. Wenn $f(x) = cf_u(x)$ und $f_u(x)/g(x) \leq M_u$ (u steht für unnormiert), dann ist offensichtlich $M = cM_u$ eine obere Schranke für $f(x)/g(x)$ und es gilt

$$\frac{f(x)}{Mg(x)} = \frac{f_u(x)}{g(x)M_u}.$$

Dies ist in vielen Anwendungen wichtig.

Beispiel 3.4 (Gleichverteilung auf einer beschränkten Teilmenge von $\mathbb{R}^p$)**.** *Sei $\mathbb{X} = \mathbb{R}^p$ und $A \subset \mathbb{R}^p$ eine offene, beschränkte Menge. Die Gleichverteilung auf A hat die Dichte*

$$f(x) = \begin{cases} const. & x \in A \\ 0 & x \notin A \end{cases}$$

Als Vorschlagsverteilung wählen wir die Gleichverteilung auf einem Rechteck R mit $A \subset R$. Dann sind nämlich die Koordinaten unabhängig und uniform, also ist die Simulation einfach. Die Akzeptierungsfunktion a ist dann gerade der Indikator von A, d.h. wir akzpetieren, wenn $X \in A$, und verwerfen, wenn $X \notin A$.

Beispiel 3.5 (Beta-Verteilung)**.** *Die Dichte lautet*

$$f(x) \;\propto\; f_u(x) = x^{\alpha-1}(1-x)^{\beta-1} \quad (0 < x < 1)$$

Wenn $\alpha \geq 1$ und $\beta \geq 1$, ist f_u beschränkt, also kann man $g(x) = 1$ wählen. Man erhält

$$\sup_x f_u(x) = \frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}}$$

Wenn $\alpha < 1$ und $\beta \geq 1$, können wir $g(x) = \alpha x^{\alpha-1}$ wählen. Dann ist

$$\frac{f_u(x)}{g(x)} = \frac{(1-x)^{\beta-1}}{\alpha} \;\leq\; \frac{1}{\alpha}.$$

Den Fall, wo $\alpha < 1$ und $\beta < 1$ sind, diskutieren wir unten.

Häufig findet man eine "Vorschlagsdichte" g durch "Aufteilung": Betrachte eine disjunkte Zerlegung von $\mathbb{X}$

$$\mathbb{X} = \bigcup B_i, \quad B_i \cap B_j = \emptyset \; (i \neq j).$$

Auf jedem B_i sei eine Wahrscheinlichkeitsdichte g_i gegeben, so dass $f_u(x) \leq M_i g_i(x)$ auf B_i ist. Wenn wir setzen

$$g(x) = \sum_{i=1}^{k} \frac{M_i}{M_1 + \cdots + M_k} g_i(x) I_{B_i}(x), \tag{3.1}$$

dann gilt nach Voraussetzung für $x \in B_i$

$$\frac{f_u(x)}{g(x)} = \frac{f_u(x)}{\frac{M_i}{M_1+\cdots+M_k} g_i(x)} \leq (M_1 + \cdots + M_k),$$

und daher

$$a(x) = \frac{f_u(x)}{M_i g_i(x)} \ (x \in B_i).$$

Die Simulation gemäss g von (3.1) erfolgt in zwei Schritten: Wähle zuerst I mit

$$\mathrm{P}[I = i] = \frac{M_i}{M_1 + \ldots + M_k}$$

und falls $I = i$, dann $X \sim g_i(x)dx$.

Beispiel 3.6 (Beta-Verteilung mit $\alpha < 1, \ \beta < 1$)**.** *Hier kann man die Zerlegung* $B_1 = (0, 0.5)$ *und* $B_2 = (0.5, 1)$ *und die Dichten*

$$g_1(x) = 2^\alpha \alpha x^{\alpha-1} \ (x \in B_1), \quad g_2(x) = 2^\beta \beta (1-x)^{\beta-1} \ (x \in B_2).$$

wählen. Die Berechnung von M_1 *und* M_2 *bietet keine Probleme.*

Beispiel 3.7 (Gammaverteilung mit $\gamma < 1$)**.** *Die Dichte lautet*

$$f_u(x) = x^{\gamma-1} \exp(-x) \quad (x \geq 0).$$

(Der zweite Parameter der Gamma-Verteilung ist ein Skalenparameter, der ohne Einschränkung als eins angenommen werden kann). Für $\gamma < 1$ *verwenden wir* $B_1 = [0, 1)$, $B_2 = [1, \infty)$ *und die Dichten*

$$g_1(x) = \gamma x^{\gamma-1}, \quad g_2(x) = \exp(-(x-1)).$$

Dann ist $M_1 = 1/\gamma$, $M_2 = 1/e$.

Die Idee der Aufteilung liefert gute Vorschläge für beliebige log-konkave Dichten f auf $\mathbb{R}$: Wenn $\log f$ konkav ist, dann sind alle Tangenten an $\log f$ Majoranten von $\log f$. Wir können also $c_1 < \ldots < c_k$ wählen, und die Tangenten $t_i(x) = a_i x + b_i$ an den Punkten c_i berechnen. Die Schnittpunkte zweier benachbarter Tangenten ergeben die Grenzen der Intervalle B_i, und die Vorschlagsdichte auf g_i auf B_i ist proportional zu $\exp(a_i x)$ (Dies ist integrierbar auf B_i, wenn wir die Punkte c_1 und c_k so wählen, dass $a_1 > 0$ und $a_k < 0$). Es genügt, wenn man zu Beginn $k = 2$ nimmt, und dann nach jedem vorgeschlagenen Wert die Tangente an diesem Wert hinzunimmt und so die Akzeptierungswahrscheinlichkeit vergrössert.

Betrachten wir zum Schluss noch den Rechenaufwand für den Verwerfungsalgorithmus. Zunächst können wir uns fragen, wieviele Variablen X wir erzeugen müssen, bis die erste akzeptiert wird. Offensichtlich hat diese Anzahl eine geometrische Verteilung, und der Erfolgsparameter ist gemäss dem Beweis von Satz 3.2 gleich

$$\mathrm{P}[X \text{ wird akzeptiert}] = \int a(x) G(dx) = \frac{1}{M} = \frac{\int f_u(x)dx}{M_u}.$$

Ideal ist also $a(x)$ möglichst nahe bei 1, bzw. ein kleines M. Dies bedeutet, dass g möglichst ähnlich wie f sein soll.

Neben der Anzahl erzeugter Variablen spielt jedoch auch der Aufwand zur Berechnung von $a(x) = \frac{f(x)}{Mg(x)}$ eine Rolle. Falls dieser gross ist, kann die Idee eines Prätests (auch squeezing genannt) nützlich sein: Finde zwei Funktionen h_1 und h_2, deren Berechnung einfach ist, und so dass $h_1(x) \leq a(x) \leq h_2(x)$. Dann prüft man zuerst $U \leq h_1(X)$. Falls ja, weiss man, dass X akzeptiert wird. Falls nein, prüft man als nächstes $U \geq h_2(X)$. Falls ja, weiss man, dass man ein neues Paar (X, U) erzeugen muss, sonst prüft man $U \leq a(X)$.

3.3 Quotienten von uniformen Zufallsvariabeln

Diese Methode ist wieder auf den eindimensionalen Fall beschränkt. Wir brauchen dafür das Beispiel 3.4, weshalb wir zuerst die Verwerfungsmethode behandeln mussten.

Wir haben oben gesehen, dass man eine Cauchy-Variable als Tangens einer uniform verteilten Variable erzeugen kann. Wenn man die Berechnung des Tangens vermeiden will, kann stattdessen zwei uniform verteilte Variablen (U, V) auf einem Halbkreis erzeugen und dann den Quotienten V/U bilden. In Polarkoordinaten ist nämlich $V/U = \tan(\phi)$, vgl. auch Lemma 3.2 unten. Wir zeigen hier, dass man sehr viele Verteilungen als Quotienten von Zufallsvariablen (U, V) erhalten kann, welche uniform sind auf einer geeigneten Menge G.

Sei $f \propto f_u$ eine beliebige Dichte auf $\mathbb{R}$ (man muss die Verteilung nur bis auf eine Normierungskonstante kennen).

Satz 3.3. *Sei (U, V) Uniform auf $G = \{(u, v);\ 0 < u < \sqrt{f_u(v/u)}\}$. Dann hat $\frac{V}{U}$ die Dichte f.*

Anschaulich sieht man diesen Satz wie folgt ein. Die Dichte von V/U an der Stelle x ist gleich

$$\frac{\mathrm{P}\left[Ux \leq V \leq U(x + dx)\right]}{dx},$$

und der Zähler ist proportional zur Fläche von

$$\{(u, v);\ 0 < u < \sqrt{f_u(v/u)},\ ux \leq v \leq u(x + dx)\}.$$

Diese Fläche ist bis auf Terme höherer Ordnung gleich der Fläche des Dreiecks mit den Ecken $(0, 0)$, $(\sqrt{f_u(x)}, x\sqrt{f_u(x)})$ und $(\sqrt{f_u(x)}, (x + dx)\sqrt{f_u(x)})$, also gleich $\frac{1}{2} f_u(x) dx$.

Für einen strengen Beweis brauchen wir einen Satz über die Transformation von mehrdimensionalen Dichten unter umkehrbaren differenzierbaren Abbildungen.

Satz 3.4. *Sei $g : G \subseteq \mathbb{R}^2 \to G' \subseteq \mathbb{R}^2$ eine stetig differenzierbare, umkehrbare Abbildung, deren Funktionaldeterminante $D(\mathbf{u}) = \det(\frac{\partial g_i}{\partial u_j})(\mathbf{u})$ nirgends verschwindet in G. Ferner sei $\mathbf{U}$ ein Zufallsvektor mit Werten in G, dessen Verteilung die Dichte $f_{\mathbf{U}}$ hat. Dann hat auch $\mathbf{X} = g(\mathbf{U})$ eine Dichte, nämlich*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{f_{\mathbf{U}}(g^{-1}(\mathbf{x}))}{|D(g^{-1}(\mathbf{x}))|}.$$

Beweis von Satz 3.4. Sei $h : \mathbb{R}^2 \to \mathbb{R}$ stetig und beschränkt. Dann folgt mit einer Substitution $\mathbf{x} = g(\mathbf{u})$

$$\mathbf{E}\left[h(\mathbf{X})\right] = \mathbf{E}\left[h(g(\mathbf{U}))\right] = \int_G h(g(\mathbf{u})) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} = \int_{G'} \frac{h(\mathbf{x}) f_{\mathbf{U}}(g^{-1}(\mathbf{x}))}{|D(g^{-1}(\mathbf{x}))|} d\mathbf{x}.$$

Durch solche Erwartungswerte ist die Dichte festgelegt, also folgt die Behauptung. □

Beweis von Satz 3.3. $g : (u,v) \to (x,y) = (u, \frac{v}{u})$ ist eine Bijektion von $\mathbb{R}^+ \times \mathbb{R}$ in sich selbst mit Funktionaldeterminante

$$D(u,v) = \det\left(\frac{\partial g}{\partial(u,v)}\right) = \det\begin{pmatrix} 1 & 0 \\ -\frac{v}{u^2} & \frac{1}{u} \end{pmatrix} = \frac{1}{u}.$$

Damit ist $f_{X,Y}(x,y) \propto x \cdot \mathbf{1}_{[0<x<\sqrt{f_u(y)}]}$, also ist die Randdichte von Y

$$f_Y(y) = \int f_{X,Y}(x,y)dx \propto \int_0^{\sqrt{f_u(y)}} x dx = \frac{f_u(y)}{2}.$$

□

Um (U,V) uniform auf G mit der Verwerfungsmethode zu erzeugen, muss man G in ein Rechteck R einschliessen. Das folgende Lemma gibt ein solches Rechteck R an.

Lemma 3.1. $G \subset [0, \sup_x \sqrt{f_u(x)}] \times [\inf_x x\sqrt{f_u(x)}, \sup_x x\sqrt{f_u(x)}]$

Beweis. Es ist klar, dass $u \le \sqrt{f_u(v/u)}$ impliziert $u \le \sup_x \sqrt{f_u(x)}$. Ausserdem

$$u^2 \le f_u(\frac{v}{u}) \Leftrightarrow v^2 \le \frac{v^2}{u^2} f_u(\frac{v}{u}),$$

also auch $\inf_x x\sqrt{f_u(x)} \le v \le \sup_x x\sqrt{f_u(x)}$. □

Beispiel 3.8 (Standardnormalverteilung)**.** *Da* $f_u(x) = \exp(-x^2/2)$*, folgt* $\sup_x \sqrt{f_u(x)} = 1$ *und* $\sup_x x\sqrt{f_u(x)} = \sqrt{2/e}$*. Die Bedingung* $u < \sqrt{f_u(v/u)}$ *ist äquivalent zu* $\log u \le -v^2/(4u^2)$*, bzw.* $v^2 \le -4u^2 \log u$*.*

Analog geht man für Gamma$(\gamma, 1)$-Verteilung mit $\gamma > 1$ und die Cauchy-Verteilung vor.

3.4 Beziehungen zwischen Verteilungen

Beziehungen zwischen Verteilungen kann man ausnützen zur Simulation. Zum Beispiel ist die t-Verteilung mit k Freiheitsgraden die Verteilung von X/Y, wobei X und Y unabhängig sind, X standard normalverteilt und kY^2 chiquadrat-verteilt mit k Freiheitsgraden. Ferner ist die Chiquadrat-Verteilung mit k Freiheitsgraden die Verteilung der Summe $\sum Z_i^2$ von k unabhängigen standard-normalverteilten Variablen, bzw. für gerades k die Verteilung der Summe von $k/2$ unabhängigen exp(1/2)-verteilten Variablen.

3.4.1 Anwendung für die Normalverteilung

Sei (X,Y) eine zweidimensionale Zufallsvariable. Betrachte die Polarkoordinaten:

$$R = \sqrt{X^2+Y^2}, \quad \Phi = \arctan\left(\frac{Y}{X}\right).$$

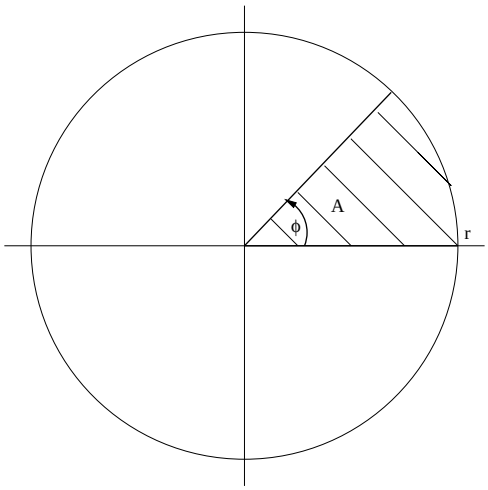

Abbildung 3.2: Das Ereignis $\{R \leq r, \Phi \leq \phi\}$

Lemma 3.2. *1. Seien X, Y i.i.d. $\sim \mathcal{N}(0,1)$ verteilt. Dann sind R und Φ unabhängig voneinander und Φ ist uniform auf $[0, 2\pi]$ und R hat die Verteilungsfunktion $1 - e^{-\frac{1}{2}r^2}$.*

2. Sei (X, Y) uniform auf $\{x^2 + y^2 \leq 1\}$. Dann sind R und Φ unabhängig mit $\Phi \sim$ Uniform$(0, 2\pi)$ und $R^2 \sim$ Uniform$(0, 1)$.

Beweis. Sei $A \subset \mathbb{R}^2$ die Menge $\{R \leq r, \Phi \leq \phi\}$, vgl. Abbildung 3.2. Die erste Behauptung folgt aus

$$\begin{aligned} \mathrm{P}\left[R \leq r, \Phi \leq \phi\right] &= \frac{1}{2\pi} \int\int_A e^{-\frac{1}{2}(x^2+y^2)} dxdy \\ &= \frac{1}{2\pi} \int_0^\phi \int_0^r r e^{-\frac{r^2}{2}} dr d\phi \\ &= -\frac{\phi}{2\pi} e^{-\frac{r^2}{2}}\Big|_0^r = \frac{\phi}{2\pi}\left(1 - e^{-\frac{r^2}{2}}\right) \end{aligned}$$

Analog folgt die zweite Behauptung:

$$\mathrm{P}\left[R^2 \leq r^2, \Phi \leq \phi\right] = \frac{\text{Fläche von } A}{\pi} = r^2 \cdot \frac{\phi}{2\pi}.$$

□

Dies gibt zwei Möglichkeiten, Paare von unabhängigen standard-normalverteilten Zufallsvariablen zu erzeugen. Beide gehen aus von U, V i.i.d. $\sim$ Uniform$(0, 1)$. Für die erste Variante benutzt man, dass $2\pi V$ und $\sqrt{-2\log(U)}$ unabhängig sind und die gleiche Verteilung haben wie Φ bzw. R (dies folgt aus der Quantiltransformation). Also sind die

Zufallsvariablen

$$(X, Y) = \sqrt{-2\log(U)}(\cos(2\pi V), \sin(2\pi V))$$

i.i.d. $\sim \mathcal{N}(0,1)$ verteilt (Box-Muller-Algorithmus).

Für die zweite Variante erzeugen wir mit dem Verwerfungsalgorithmus zunächst (U, V) uniform auf $\{x^2 + y^2 \leq 1\}$ und bilden dann

$$(X, Y) = \sqrt{\frac{-2\log(U^2+V^2)}{U^2+V^2}}(U, V)$$

Diese Variante braucht die trigonometrischen Funktionen nicht.

Wenn wir einmal univariate normalverteilte Zufallsvariablen haben, dann können wir damit auch gemäss einer multivariaten Normalverteilung simulieren. Die multivariate Normalverteilung $\mathcal{N}_{\mathrm{p}}(\mu, \Sigma)$ hat die Dichte

$$(2\pi)^{-p/2}(\det\Sigma)^{-1/2}\exp(-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)),$$

wobei $\mu = \mathbf{E}[X]$ der Vektor der Erwartungswerte von X ist und Σ die Matrix der Varianzen und Kovarianzen von X. Zur Simulation benützen wir, dass sich X darstellen lässt in der Form

$$X = \mu + AY, \quad \text{mit } Y_1, \ldots, Y_p \text{ i.i.d. } \sim \mathcal{N}(0,1).$$

Wir verzichten auf die Herleitung dieses Resultats, geben jedoch an, wie man die Matrix A wählen muss. Aus den Rechenregeln für den Erwartungswert folgt

$$\Sigma = \mathbf{E}\left[(X-\mu)(X-\mu)^T\right] = A\,\mathbf{E}\left[\mathbf{Y}\mathbf{Y}^T\right]A^T = AA^T.$$

Diese Gleichung hat viele Lösungen. Am einfachsten ist es, zusätzlich zu verlangen, dass A eine untere Dreiecksmatrix ist. Dann kann man A mit der Cholesky-Zerlegung schnell und numerisch stabil berechnen.

Falls Σ^{-1} gegeben ist anstatt Σ, zerlegt man Σ^{-1} als BB^T, wobei B eine untere Dreiecksmatrix ist. Es folgt dann $\Sigma = (BB^T)^{-1} = B^{-T}B^{-1}$ d.h. $A = B^{-T}$. Zur Berechnung von $\mathbf{X}$ aus $\mathbf{Y}$ löst man $B^T\mathbf{X} = \mathbf{Y}$ mit Rückwärtseinsetzen, d.h. man muss keine Matrizen invertieren.

Dieses Vorgehen ist praktikabel für Dimension $p \leq 1000$. Grössere p's treten auf bei der Simulation von Gauss'schen stochastischen Prozessen. Dort gilt jedoch meistens $\Sigma_{ij} = R(i-j)$, d.h. $\boldsymbol{\Sigma}$ ist eine sogenannte Toeplitz-Matrix. Dafür existieren spezielle Algorithmen, welche auf der Fouriertransformation beruhen.

3.4.2 Anwendung für die Poissonverteilung

Lemma 3.3. *Sei* (X_i) *i.i.d.* $\sim \mathrm{Exp}(1)$ *und* $S_n = \sum_{i=1}^n X_i$ *mit* $S_0 = 0$. *Dann ist* $S_n \sim$ *Gamma*$(n,1)$*-verteilt und*

$$\mathrm{P}[S_n \leq t < S_{n+1}] = e^{-t}\frac{t^n}{n!}.$$

Beweis. Die erste Behauptung folgt mit der Faltungsformel und Induktion nach n. Für die zweite Behauptung beachtet man

$$\begin{aligned}\mathrm{P}[S_n \leq t < S_{n+1}] &= \mathrm{P}[S_n \leq t] - \mathrm{P}[S_{n+1} \leq t] \\ &= \int_0^t e^{-x}\left(\frac{x^{n-1}}{(n-1)!} - \frac{x^n}{n!}\right)dx = e^{-x}\frac{x^n}{n!}\Big|_0^t\end{aligned}$$

□

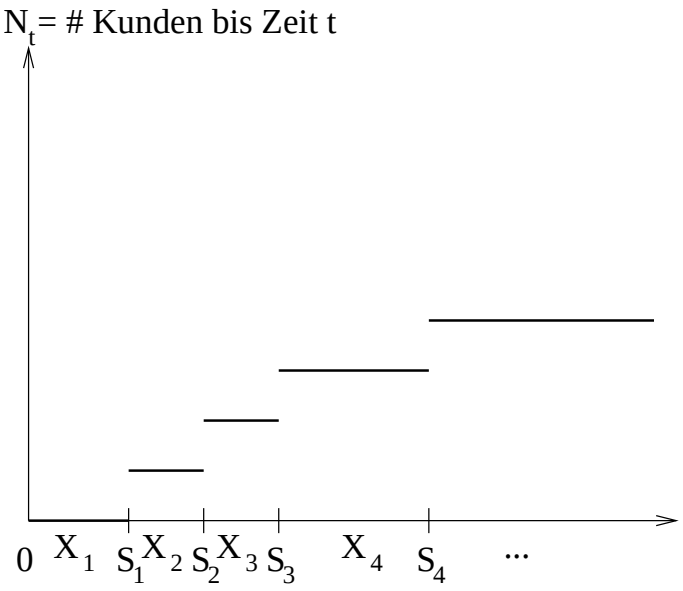

Abbildung 3.3: Anzahl Kunden in einem Bedienungssystem

Interpretation: Wir betrachten die X_i als die Zeiten zwischen den Ankünften von Kunden in einem Bedienungssystem, vgl. die Abbildung 3.3. Dann ist S_n die Ankunftszeit des n-ten Kunden, und $S_n \leq t < S_{n+1}$ bedeutet, dass die Anzahl Kunden, die bis zur Zeit t angekommen sind, gleich n ist. Diese Anzahl ist also poissonverteilt.

Anwendung auf die Simulation: Wenn U_i uniform ist, dann hat $X_i = -\log(U_i)$ eine Exp (1)-Verteilung, und gemäss obigem Resultat ist daher

$$Y = \min\{n \mid \sum_{i=1}^{n} (-\log(U_i)) > t\} - 1 = \min\{n \mid U_1 \cdot U_2 \cdots U_n < e^{-t}\} - 1$$

Poisson(t)-verteilt.

3.5 Zusammenfassung: Simulation der wichtigsten Verteilungen

Normalverteilung. Verwende den Quotient von Uniformen Z.V. $X = \frac{V}{U}$ mit (U, V) uniform auf

$$\{v^2 \leq -4u^2 \log(u)\} \subset [0, 1] \times [-\sqrt{2/e}, \sqrt{2/e}],$$

oder die Darstellung von unabhängigen normalverteilten Paaren in Polarkoordinaten:

$$(X, Y) = \sqrt{-2\log(U)}\ (\sin(2\pi V), \cos(2\pi V)),$$

mit U und V unabhängig und uniform$(0,1)$, bzw.

$$(X,Y) = \sqrt{\frac{-2\log(U^2+V^2)}{U^2+V^2}}\,(U,V),$$

mit (U,V) uniform auf $\{u^2+v^2 \le 1\}$.

Binomialverteilung. Benütze für kleines n die Darstellung als Summe von unabhängigen binären Variablen, bzw. für grösseres n die Quantiltransformation mit einer Aufzählung der möglichen Werte, die bei $[np]$ anfängt (vgl. Beispiel 3.3).

Poissonverteilung. Verwende die Quantiltransformation (für grosses λ mit einer Aufzählung beginnend bei $[\lambda]$) oder den Zusammenhang mit i.i.d. exponentialverteilten Ankunftszeiten:

$$X = \min\{n \ge 1; U_1 U_2 \cdots U_n < \exp(-\lambda)\} - 1$$

mit (U_i) i.i.d. Uniform$(0,1)$.

Cauchy-Verteilung. Verwende die Quantiltransformation $X = \tan(\pi(U-0.5))$ mit U Uniform$(0,1)$, oder den Quotienten von uniformen Zufallsvariablen $X = \frac{V}{U}$ mit (U,V) uniform auf $\{u^2+v^2 \le 1\}$.

Gamma$(\gamma,1)$-Verteilung. Für $\gamma < 1$ verwende die Verwerfungsmethode mit Vorschlagsdichte

$$g(x) = \frac{e}{e+\gamma}\gamma x^{\gamma-1}\mathbf{1}_{[0,1]}(x) + \frac{\gamma}{e+\gamma}\exp(-(x-1))\mathbf{1}_{(1,\infty)}(x).$$

Für $\gamma = 1$ (Exponentialverteilung) verwende die Quantiltransformation $X = -\log(U)$. Für $\gamma > 1$ verwende den Quotient von Uniformen Z.V. $X = \frac{V}{U}$ mit (U,V) uniform auf

$$\{2\log(u) < (\gamma-1)\log(v/u) - v/u\} \subset [0,a]\times[0,b]$$

mit $a^2 = ((\gamma-1)/e)^{\gamma-1}$ und $b^2 = ((\gamma+1)/e)^{\gamma+1}$. Für γ gross schreibe $\gamma = k+\gamma_1$ mit k ganzzahlig und $1 < \gamma < 2$. Dann verwendet man die Darstellung als Summe von k exp(1)-verteilten und einer $\Gamma(\gamma_1,1)$-verteilten Zufallsvariable.

Beta(α,β)-Verteilung. Man verwendet entweder die Darstellung

$$X = \frac{X_1}{X_1+X_2},$$

wobei X_1 und X_2 unabhängig und Gamma$(\alpha,1)$- bzw. Gamma$(\beta,1)$-verteilt sind, oder die Verwerfungsmethode mit Vorschlagsdichte

$$\begin{aligned}
g(x) &= 1 \quad (\text{wenn } \alpha>1, \beta>1),\\
g(x) &= \alpha x^{\alpha-1} \quad (\text{wenn } \alpha<1, \beta>1),\\
g(x) &= \beta(1-x)^{\beta-1} \quad (\text{wenn } \alpha>1, \beta<1),\\
g(x) &= \Big(\frac{\beta}{\alpha+\beta}2^{\alpha}\alpha x^{\alpha-1}\mathbf{1}_{[0,0.5]}(x) + \frac{\alpha}{\alpha+\beta}2^{\beta}\beta(1-x)^{\beta-1}\mathbf{1}_{[0.5,1]}(x)\Big)\\
&\quad (\text{wenn } \alpha<1, \beta<1).
\end{aligned}$$

***t*-Verteilung.** Die Anzahl Freiheitsgrade sei $\nu > 0$. Man verwendet die Darstellung

$$X = \frac{X_1}{\sqrt{2X_2/\nu}},$$

wobei X_1 und X_2 unabhängig und normal-, bzw. Gamma$(\nu/2, 1)$-verteilt sind.

3.6 Zufallsstichproben und Zufallspermutationen

Wir nehmen an, wir möchten eine Stichprobe $S = (i_1, i_2, \ldots, i_n)$ ohne Zurücklegen aus einer durchnummerierten Population $\{1, 2, \ldots, N\}$ ziehen. Die entsprechenden Wahrscheinlichkeiten sind

$$\frac{1}{N(N-1)\cdots(N-n+1)},$$

wenn die Reihenfolge innerhalb der Stichprobe wichtig ist, bzw.

$$\frac{1}{\binom{N}{n}},$$

wenn die Reihenfolge unwesentlich ist. Bei Berücksichtigung der Reihenfolge haben wir für $N = n$ eine Permutation.

Wir haben die folgenden Algorithmen zur Auswahl:

Algorithmus 3.2 (ohne Berücksichtigung der Reihenfolge)**.**

1. *Setze $S = (1, 2, \ldots, n)$, $k = n$.*
2. *Falls $k = N$, ist S das Resultat, sonst setze $k = k + 1$.*
3. *Wähle $U \sim$ Uniform$(0, 1)$. Wenn $U < \frac{n}{k}$, wähle I gleichverteilt auf $\{1, 2, \ldots, n\}$ und ersetze das I-te Element S_I von S durch k, andernfalls ändert sich S nicht.*
4. *zurück zu 2.*

Für diesen Algorithmus braucht man N nicht im Voraus zu kennen, man stoppt, sobald man am Ende der Liste ist.

Algorithmus 3.3 (mit Berücksichtigung der Reihenfolge)**.**

1. *Erzeuge $U_1, \ldots, U_N$ i.i.d. $\sim$ Uniform$(0, 1)$*
2. *Bestimme $R_i =$ Rang(U_i) durch sortieren.*
3. *$S = \{$Rang$(U_1), \ldots,$ Rang$(U_n)\}$.*

Dieser Algorithmus ist schnell zu programmieren, aber das Sortieren braucht $N \log N$ Operationen, ist also langsam für N sehr gross.

Algorithmus 3.4 (mit Berücksichtigung der Reihenfolge)**.**

1. *Setze $M = (1, 2, \ldots, N)$ und $k = 1$.*
2. *Wähle I gleichverteilt auf $\{k, k + 1, \ldots, N\}$ und vertausche M_k und M_I.*
3. *Wenn $k = n$, ist $S = \{M_1, \ldots, M_n\}$ das Resultat, sonst setze $k = k + 1$ und gehe zurück zu 2.*

Der Beweis der Korrektheit der Algorithmen ist eine Übungsaufgabe.

3.7 Importance sampling

Zur Erinnerung: Der Verwerfungsalgorithmus simuliert gemäss einer Verteilung π, indem er zuerst eine Variable X gemäss einer falschen Verteilung τ erzeugt und dieses X mit Wahrscheinlichkeit $a(X) = f(X)/(Mg(X)$ akzeptiert. Dabei sind f, bzw. g die Dichten von π, bzw. τ und M ist eine obere Schranke für den Quotienten $w = f/g$.

Importance sampling beruht auf einer ähnlichen Idee, nur erfolgt die Korrektur nicht bei der Erzeugung der Variablen, sondern durch eine Gewichtung bei der Mittelung. Es gilt

$$\mathbf{E}_\pi [h(X)] = \int h(x)\pi(dx) = \int h(x)\frac{f(x)}{g(x)}g(x)\mu(dx) = \int h(x)w(x)\tau(dx) = \mathbf{E}_\tau [h(X)w(X)] .$$

Wenn wir also Variablen X_i haben, welche i.i.d und gemäss τ verteilt sind, dann können wir die Schätzung

$$\widetilde{\theta} = \frac{1}{N}\sum_{i=1}^{N} h(X_i)w(X_i)$$

verwenden. Im Fall, wo f nur bis auf eine Normierungskonstante gegeben ist (d.h. $f \propto f_u$), verwenden wir analog

$$\frac{\sum_{i=1}^N h(X_i)w_u(X_i)}{\sum_{i=1}^N w_u(X_i)}$$

mit $w_u = f_u/g$. Im Unterschied zum Verwerfungsalgorithmus brauchen wir hier keine obere Schranke für den Quotienten w. Offensichtlich ist

$$\mathrm{Var}(\widetilde{\theta}) = \frac{1}{N}\mathrm{Var}(h(\mathbf{X}_i)w(X_i)) \leq \frac{1}{N}\int h(x)^2 w(x)^2 \tau(dx) = \frac{1}{N}\int h(x)^2 \frac{f(x)^2}{g(x)}\mu(dx).$$

Damit dies endlich ist für alle beschränkten Funktionen h, muss $\int f(x)^2/g(x)\mu(dx)$ endlich sein. Dies ist eine schwächere Bedingung als f/g beschränkt. Die Dichte g sollte aber längerschwänzig sein als f. Damit die Varianz nicht riesig wird, muss ausserdem g ähnlich sein wie f. In hohen Dimensionen ist dies schwierig zu erreichen, weil sich dort die meisten Verteilungen tendenziell stark unterscheiden und weil man dort nur wenige Kandidaten als Vorschlagsverteilung g zur Verfügung hat. Daher ist Importance Sampling nur beschränkt nützlich in hohen Dimensionen.

3.8 Markovketten und Markovprozesse

Im Prinzip kann man von einer p-dimensionalen Verteilung auch rekursiv simulieren: Sei π_1 die Randverteilung von X_1 und $\pi_{j|j-1,\ldots,1}$ die bedingte Verteilung von X_j gegeben $X_1 = x_1, \ldots, X_{j-1} = x_{j-1}$. Dann kann man zuerst X_1 gemäss π_1 erzeugen und dann iterativ X_j gemäss $\pi_{j|j-1,\ldots,1}$ für $j = 2, \ldots, p$. Dies geht jedoch nur, wenn man π_1 und $\pi_{j|j-1,\ldots,1}$ explizit berechnen kann. Im Allgemeinen muss man dazu eine Reihe von Integralen berechnen, was oft nicht möglich ist. Mit der Simulation will man gerade numerische Integration vermeiden.

Bei der mehrdimensionalen Normalverteilung sind alle bedingten Verteilungen wieder normal, und mit der Choleski-Zerlegung berechnet man gerade die bedingten Erwartungswerte und Varianzen.

Ein anderes wichtiges Beispiel, wo dieses Vorgehen praktikabel ist, sind *Markovketten*. Das sind stochastische Prozesse in diskreter Zeit, bei denen die bedingte Verteilung von X_j gegeben $X_1 = x_1, \ldots, X_{j-1} = x_{j-1}$ nur von x_{j-1} abhängt und explizit durch einen sogenannten Übergangskern gegeben ist, siehe den Abschnitt 4.1 unten.

Markovprozesse in stetiger Zeit, bei denen auch der Wertebereich kontinuierlich ist, sind schwieriger zu simulieren. Solche Prozesse werden durch stochastische Differentialgleichungen erzeugt, auf die wir im Folgenden kurz eingehen.

3.8.1 Simulation stochastischer Differentialgleichungen

Eine stochastische Differentialgleichung entsteht aus einer gewöhnlichen Differentialgleichung durch Addition eines stochastischen Rauschens N_t:

$$\frac{dX_t}{dt} = f(X_t) + \sigma(X_t)N_t.$$

Das Rauschen soll die Eigenschaft haben, dass N_t und N_s stochastisch unabhängig sind für $t \neq s$ (sogenanntes weisses Rauschen). Dies ist jedoch ein pathologisches Objekt, denn dann muss gelten

$$\mathrm{Var}\left(\int_0^t N_s ds\right) = \text{const } t.$$

Damit ist aber $\int_0^t N_s ds$ von der Grössenordnung $\sqrt{t}$, und damit existiert (N_t) gar nicht.

Der Ausweg besteht darin, der obigen Differentialgleichung eine Interpretation zu geben, die N_t nicht enthält. Man beginnt mit der Brown'schen Bewegung (B_t), welche definiert ist durch die beiden folgenden Eigenschaften

1. $B_0 = 0$ f.s..
2. Für alle $t_0 = 0 < t_1 < t_2 < \cdots < t_n$ sind die Zuwächse $B_{t_i} - B_{t_{i-1}}$ $(i = 1, \ldots n)$ unabhängig und $\mathcal{N}(0, t_i - t_{i-1})$-verteilt.

Wiener hat gezeigt, dass ein solcher Prozess existiert und sogar so gewählt werden kann, dass die Pfade fast sicher stetig, aber nirgends differenzierbar sind. Daher heisst die Brown'sche Bewegung oft auch ein Wiener Prozess. Formal ist N_t die Ableitung von B_t, und daher kann die obige stochastische Differentialgleichung geschrieben werden als

$$dX_t = f(X_t)dt + \sigma(X_t)\, dB_t,$$

bzw. in integrierter Form als

$$X_t = X_0 + \int_0^t f(X_s)ds + \int_0^t \sigma(X_s)dB_s.$$

Der entscheidende Punkt dabei ist, dass man das "stochastische Integral" $\int Z_s dB_s$ mathematisch exakt definieren kann für Prozesse (Z_t), welche endliches zweites Moment haben und bei denen Z_t nur abhängt von B_s für $s \leq t$. Dies geht jedoch nicht als ein Lebesgue- oder Stieltjes-Integral, weil (B_t) nicht nur nicht differenzierbar ist, sondern auch unendliche totale Variation hat. Man benutzt vielmehr eine Riemann-Approximation der Form

$$\sum_{j=1}^{n} Z_{t_{j-1}}(B_{t_j} - B_{t_{j-1}})$$

und zeigt, dass diese in L_2 konvergiert, wenn die Partition feiner wird. Dabei ist es wesentlich, dass man als Stützstelle den linken Rand des Intervalls und nicht irgend einen andern Punkt nimmt.

Damit kann man also genau definieren, was man unter einer Lösung versteht. Als nächstes muss man zeigen, dass Lösungen wirklich existieren. Dies geht analog wie bei gewöhnlichen Differentialgleichungen mit der iterativen Approximation

$$X_t^{(m)} = X_0 + \int_0^t f(X_s^{(m-1)})ds + \int_0^t \sigma(X_s^{(m-1)})dB_s.$$

Die Simulation von Lösungen solcher stochastischer Differentialgleichungen ist in den letzten 20 Jahren auf sehr viel Interesse gestossen. Das einfachste Verfahren erzeugt eine approximative Lösung an den Zeitpunkten $k\Delta$ gemäss dem Euler-Schema

$$X_{(k+1)\Delta} = f(X_{k\Delta}) + \sigma(X_{k\Delta})(B_{(k+1)\Delta} - B_{k\Delta}).$$

Weil die Zuwächse von B normalverteilt sind, ist die Implementation davon trivial. Man kann zeigen, dass dies gegen die Lösung konvergiert für $\Delta \to 0$, allerdings langsam. In Analogie zur Numerik gewöhnlicher Differentialgleichungen kommen einem sofort Approximationsschemata höherer Ordnung in den Sinn. Es zeigt sich aber, dass man damit die Konvergenz nicht verbessern kann. Es gibt Verfahren, die eine schnellere Konvergenzrate haben, aber diese sind kompliziert. Für mehr Details verweise ich auf die Literatur. Neuere Arbeiten von Beskos, Papaspiliopoulos und Roberts zeigen, dass man unter gewissen Annahmen an f und σ mit der Verwerfungsmethode exakt simulieren kann.

3.9 Genauigkeit der Monte Carlo Schätzung

Seien $X_1, X_2, \ldots$ i.i.d. mit Verteilung π.

Die uns interessierende Grösse ist $\theta = \int h(x)\pi(dx)$, die wir approximieren durch $\widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^N h(X_i)$. Folgender Satz gibt uns an, wie gross der Approximationsfehler ist:

Satz 3.5. *Wenn $\int h(x)^2\pi(dx) < \infty$ ist, dann gilt*

1. $\mathrm{P}\left[\sqrt{N}(\widehat{\theta}_N - \theta) \le \sigma t\right] \to \Phi(t)$ *für alle* $t \in \mathbb{R}$, *wobei* $\sigma^2 = \sigma^2(h) = \int (h(x) - \theta)^2\pi(dx)$ *ist.*

2. *Das Intervall*
$$I_N = \widehat{\theta}_N \pm \Phi^{-1}(1 - \frac{\alpha}{2})\frac{S_N}{\sqrt{N}},$$
wobei $S_N^2 = \frac{1}{N}\sum_{i=1}^N (h(X_i) - \widehat{\theta}_N)^2$ *die Stichprobenvarianz ist, enthält den unbekannten wahren Wert* θ *mit einer Wahrscheinlichkeit, die für* $N \to \infty$ *gegen* $1 - \alpha$ *konvergiert.*

Bemerkungen:

1. Die Genauigkeitsangabe ist a-posteriori und mit Unsicherheit α behaftet.
2. Die Geschwindigkeit $\frac{1}{\sqrt{N}}$ ist langsam!

3. Weil N immer gross ist, spielt es keine Rolle, ob N oder $N-1$ im Nenner von S_N^2 verwendet wird.

Beweis. Die erste Aussage ist einfach der Zentrale Grenzwertsatz. Die zweite Aussage ist ein Teil des Satzes von Slutsky, siehe Mathematische Statistik. Wir skizzieren die Idee und fixieren dazu zunächst ein $\delta > 0$. Dann gilt

$$\begin{aligned}
\mathrm{P}\left[I_N \ni \theta\right] &= \mathrm{P}\left[\frac{\sqrt{N}|\widehat{\theta}_N - \theta|}{S_N} \leq \Phi^{-1}(1-\frac{\alpha}{2})\right] \\
&\geq \mathrm{P}\left[\frac{\sqrt{N}|\widehat{\theta}_N - \theta|}{S_N} \leq \Phi^{-1}(1-\frac{\alpha}{2}), 1-\delta \leq \frac{S_N}{\sigma} \leq 1+\delta\right] \\
&\geq \mathrm{P}\left[\sqrt{N}|\widehat{\theta}_N - \theta| \leq (1-\delta)\sigma\Phi^{-1}(1-\frac{\alpha}{2})\right] \\
&\to \Phi\Big((1-\delta)\Phi^{-1}(1-\frac{\alpha}{2})\Big) - \Phi\Big(-(1-\delta)\Phi^{-1}(1-\frac{\alpha}{2})\Big)
\end{aligned}$$

für $N \to \infty$. Für $\delta \to 0$ konvergiert der letzte Ausdruck gegen $1-\alpha$. Asymptotisch sind die Fehler bei den Ungleichungen vernachlässigbar, weil $\mathrm{P}\left[1-\delta \leq \frac{S_N}{\sigma} \leq 1+\delta\right] \to 1$ für alle δ. □

Im Spezialfall, wo man eine Wahrscheinlichkeit $\theta = \pi(A)$ approximieren will (d.h. $h(x) = 1_A(x)$), gibt es auch eine a priori Abschätzung für die benötigte Anzahl Replikate: Die Abweichung soll z. B. mit Wahrscheinlichkeit 0.95 höchstens $0.1 \cdot \pi(A)$ sein, d.h

$$1.96 \cdot \sqrt{\frac{\pi(A)(1-\pi(A))}{N}} \leq 0.1 \cdot \pi(A).$$

Dies ergibt

$$N \geq 385 \cdot \frac{(1-\pi(A))}{\pi(A)}.$$

Wenn $\pi(A)$ sehr klein ist, dann muss also N sehr gross sein und dies ist unerfreulich. Verschiedene Methoden zur Erhöhung der Genauigkeit werden im Abschnitt 3.10 besprochen.

Analog kann man Genauigkeitsangaben erhalten für kompliziertere Funktionale von π als einfach Erwartungswerte. Wir betrachten noch speziell die Quantile und verweisen im Übrigen auf die Mathematische Statistik.

Sei $h(X_i) =: Y_i$ und q_α das α-Quantil von Y_i, $\inf\{y; \mathrm{P}\left[Y_i \leq y\right] \geq \alpha\}$. Die empirischen Quantile sind: $\hat{q}_\alpha = Y_{([(N+1)\alpha])}$, wobei $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(N)}$ die geordneten Beobachtungen bezeichnen und $[x]$ den ganzzahligen Teil der Zahl x. Es gilt dann:

Satz 3.6. *Wenn* $\mathrm{P}\left[Y_i \leq q_\alpha\right] = \mathrm{P}\left[Y_i < q_\alpha\right] = \alpha$*, dann ist* $[Y_{(k_1)}, Y_{(k_2)}]$ *ein genähertes Vertrauensintervall zum Niveau* $1-\gamma$ *für* q_α*, wenn*

$$\begin{aligned}
k_1 &= [N\alpha + 0.5 - \sqrt{N\alpha(1-\alpha)}\Phi^{-1}(1-\frac{\gamma}{2})] \\
k_2 &= [N\alpha + 0.5 + \sqrt{N\alpha(1-\alpha)}\Phi^{-1}(1-\frac{\gamma}{2})] + 1
\end{aligned}$$

Beweis. Dies folgt aus dem Zentralen Grenzwersatz für binomialverteilte Zufallsvariabeln. Es gilt

$$\mathrm{P}\left[q_\alpha \notin [Y_{(k_1)}, Y_{(k_2)}]\right] = \mathrm{P}\left[Y_{(k_1)} > q_\alpha\right] + \mathrm{P}\left[Y_{(k_2)} < q_\alpha\right].$$

Das Ereignis $Y_{(k_1)} > q_\alpha$ ist gleichbedeutend damit, dass es höchstens $k_1 - 1$ Beobachtungen $\leq q_\alpha$ gibt. Also folgt

$$\mathrm{P}\left[Y_{(k_1)} > q_\alpha\right] = \sum_{j=0}^{k_1-1} \binom{N}{j} \alpha^j (1-\alpha)^{N-j} \to \Phi\left(\frac{k_1 - 1 - N\alpha + 0.5}{\sqrt{N\alpha(1-\alpha)}}\right) = \frac{\gamma}{2}$$

für $N \to \infty$ (die 0.5 ist eine Stetigkeitskorrektur).

Für den zweiten Term geht man analog vor. □

Für ein konkretes Beispiel sei $\alpha = 0.9$, $N = 1000$ und $\gamma = 0.05$. Dann bekommt man $k_1 = 881$ und $k_2 = 920$.

3.10 Reduktion der Varianz

3.10.1 Antithetische Variabeln

Die Varianz des arithmetischen Mittels von abhängigen Zufallsvariablen hängt von den Kovarianzen ab. Man sieht sofort, dass sich die Varianz gegenüber dem unabhängigen Fall verkleinert, wenn alle Kovarianzen $\mathrm{Cov}(X_i, X_j)$ für $i \neq j$ negativ sind. Antithetische Variablen stellen eine Möglichkeit dar, solche negativen Korrelationen einzuführen.

Wir betrachten folgende Situation:

$$\theta = \int_0^1 h(x)dx = \mathbf{E}\left[h(U)\right], \qquad U \sim \mathrm{Uniform}(0,1).$$

Anstatt $\widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^N h(U_i)$ verwenden wir nun $\widetilde{\theta}_N = \frac{1}{2N}\sum_{i=1}^N \left(h(U_i) + h(1-U_i)\right)$. Dann ist

$$\begin{aligned}\mathrm{Var}\left(\widetilde{\theta}_N\right) &= \frac{N}{4N^2}\mathrm{Var}\left(h(U_i) + h(1-U_i)\right) \\ &= \frac{1}{2N}\left(\mathrm{Var}\left(h(U_i)\right) + \mathrm{Cov}\left(h(U_i), h(1-U_i)\right)\right).\end{aligned}$$

Falls $\mathrm{Cov}\left(h(U_i), h(1-U_i)\right) < 0$, ist die Varianz $\mathrm{Var}\left(\widetilde{\theta}_N\right)$ kleiner als $\mathrm{Var}\left(\widehat{\theta}_{2N}\right)$. Eine Reihe von Beispielen ist durch folgendes Lemma abgedeckt:

Lemma 3.4. *Ist die Funktion h monoton, dann ist* $\mathrm{Cov}\left(h(U), h(1-U)\right) < 0$, *ausser wenn h konstant ist auf* $(0,1)$.

Beweis. Seien U_1 und U_2 unabhängig und Uniform$(0,1)$ verteilt. Dann haben wir

$$\mathrm{Cov}\left(h(U), h(1-U)\right) = \frac{1}{2}\mathbf{E}\left[\left(h(U_1) - h(U_2)\right)\cdot\left(h(1-U_1) - h(1-U_2)\right)\right].$$

Wir nehmen an, dass h z.B. monoton wachsend ist. Wenn $U_1 < U_2$, dann ist der 1. Faktor ≤ 0 und der 2. Faktor ≥ 0. Wenn aber $U_1 > U_2$, dann ist der 1. Faktor ≥ 0 und der 2. Faktor ≤ 0. Damit ist der ganze Integrand ≤ 0.

Um nachzuprüfen, dass die Kovarianz strikt negativ ist, untersuchen wir, wann der Integrand $= 0$ ist. Dazu muss einer der Faktoren $= 0$ sein, das heisst fast sicher muss entweder $h(U_1) = h(U_2)$ oder $h(1-U_1) = h(1-U_2)$ sein. Wegen der Monotonie ist das nur möglich, wenn h konstant ist. □

Dies lässt sich insbesondere anwenden zur Approximation von $\int h(x)F(dx)$, wenn h monoton ist und wir gemäss F mit der Quantiltransformation simulieren können (F^{-1} ist monoton).

3.10.2 Kontrollvariabeln

Wir nehmen an, es existiere eine Funktion r so dass $\mathbf{E}[r(X_i)]$ bekannt ist. O.B.d.A. sei $\mathbf{E}[r(X_i)] = 0$.

Wir betrachten nun

$$\widetilde{\theta}_{N,c} = \frac{1}{N}\sum_{i=1}^{N}(h(X_i) - c\,r(X_i)).$$

(Dies ist ein erwartungstreuer Schätzer für beliebiges c). Seine Varianz ist:

$$\begin{aligned}\operatorname{Var}\left(\widetilde{\theta}_{N,c}\right) &= \frac{1}{N}\operatorname{Var}\left(h(X_i) - c\,r(X_i)\right)\\ &= \frac{1}{N}\left[\operatorname{Var}\left(h(X_i)\right) - 2c\operatorname{Cov}\left(h(X_i), r(X_i)\right) + c^2\operatorname{Var}\left(r(X_i)\right)\right]\end{aligned}$$

Das optimale c_{opt}, das diese Varianz minimiert, ist somit

$$c_{opt} = \frac{\operatorname{Cov}\left(h(X_i), r(X_i)\right)}{\operatorname{Var}\left(r(X_i)\right)}$$

Einsetzen ergibt

$$\operatorname{Var}\left(\widetilde{\theta}_{N,c_{opt}}\right) = \frac{1}{N}\operatorname{Var}\left(h(X_i)\right)\left(1 - \operatorname{Corr}\left(h(X_i), r(X_i)\right)^2\right) \le \frac{1}{N}\operatorname{Var}\left(h(X_i)\right).$$

Im Allgemeinen sind die Grössen, von denen c_{opt} abhängt, nicht bekannt. Wir können aber c_{opt} schätzen durch:

$$\widehat{c}_{opt} = \frac{\sum_{i=1}^{N}(h(X_i) - \widehat{\theta}_N)r(X_i)}{\sum_{i=1}^{N} r(X_i)^2}.$$

Dies ist konsistent, und man erhält asymptotisch die gleiche Varianz, wie wenn c_{opt} bekannt ist.

Für $r(X_i) > 0$ und $\mathbf{E}[r(X_i)] = 1$ verwendet man meist eine multiplikative Korrektur:

$$\widetilde{\theta}_N = \frac{\frac{1}{N}\sum_{i=1}^{N} h(X_i)}{\frac{1}{N}\sum_{i=1}^{N} r(X_i)}$$

Hier kann man Erwartungswert und Varianz von $\tilde{\theta}_N$ nicht exakt berechnen. Für $N \to \infty$ knovergiert aber $\tilde{\theta}_N$ fast sicher gegen θ, und ausserdem ist $\sqrt{N}(\widetilde{\theta}_N - \theta)$ asymptotisch $\mathcal{N}\left(0, \operatorname{Var}(h) - 2\theta\operatorname{Cov}(h,r) + \theta^2\operatorname{Var}(r)\right)$-verteilt, weil

$$\tilde{\theta}_N - \theta = \frac{\frac{1}{N}\sum_{i=1}^{N}[h(X_i) - \theta r(X_i)]}{\frac{1}{N}\sum_{i=1}^{N} r(X_i)}.$$

Für den Zähler gilt der ZGS, und der Nenner konvergiert gegen 1 und hat daher asymptotisch keinen Einfluss, vgl. den Beweis von Satz 3.5. Multiplikative Korrektur bringt also eine Verbesserung, wenn $h(X_i)$ und $r(X_i)$ stark korreliert sind.

Beispiel 3.9 (Varianz des gestutzten Mittels)**.** *Wir betrachten die Schätzung der Varianz des gestutzten Mittels von n standard-normalverteilten Zufallsvariablen, vgl. Beispiel 1.5.1. Als Kontrollvariable bietet sich n mal das ungestutzte Mittel im Quadrat an.*

3.10.3 Importance Sampling und Varianzreduktion

Wir haben importance sampling bereits im Abschnitt 3.7 besprochen als eine Alternative zum Verwerfungsalgorithmus. Die Idee besteht darin $\theta = \int h(x)\pi(dx)$ zu schätzen mit Hilfe von Werten, die gemäss der "falschen" Verteilung τ erzeugt wurden. Zur Korrektur betrachtet man dann das gewichtete Mittel

$$\widetilde{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} h(Y_i)w(Y_i), \quad Y_i \text{ i.i.d. } \sim \tau,$$

wobei

$$w(x) := \frac{f(x)}{g(x)}$$

und f, bzw. g die Dichten von π, bzw. τ sind (bezüglich einem Bezugsmass μ).

Diese Methode wird aber nicht nur angewandt, wenn es unmöglich oder schwierig ist, Zufallsvariablen mit Verteilung π zu erzeugen. Es gibt auch Fälle, wo bei geeigneter Wahl von τ die Schätzung $\widetilde{\theta}_N$ genauer ist als die direkte Approximation

$$\widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} h(X_i), \quad X_i \text{ i.i.d. } \sim \pi.$$

Man sieht leicht ein, dass

$$\begin{aligned}
\mathbf{E}\left[\widehat{\theta}_N\right] &= \mathbf{E}\left[\widetilde{\theta}_N\right] = \theta \\
N\,\mathrm{Var}\left(\widehat{\theta}_N\right) &= \int h(x)^2\pi(dx) - \theta^2 \\
N\,\mathrm{Var}\left(\widetilde{\theta}_N\right) &= \int h(x)^2 w(x)^2\tau(dx) - \theta^2 \\
&= \int h(x)^2 w(x)\pi(dx) - \theta^2.
\end{aligned}$$

Das folgende Lemma zeigt, wie man τ wählen muss, damit $\mathrm{Var}\left(\widetilde{\theta}_N\right)$ minimal wird.

Lemma 3.5. *Es gilt stets*

$$\int h(x)^2 w(x)^2\tau(dx) \geq \left(\int |h(x)|\pi(dx)\right)^2,$$

und man hat genau dann Gleichheit, wenn $|h(x)|w(x)$ konstant ist.

Beweis. Mit Cauchy-Schwarz:

$$\begin{aligned}
\left(\int |h(x)|\pi(dx)\right)^2 &= \left(\int |h(x)|w(x)\tau(dx)\right)^2 \\
&\leq \int h(x)^2 w(x)^2\tau(dx)\cdot 1
\end{aligned}$$

□

Das optimale g hängt also von h ab, und wir sollten möglichst $g(x)$ proportional zu $|h(x)|f(x)$ wählen. Dies kann man selten exakt erreichen, aber es gibt einem doch Hinweise, wie ein günstiges g aussehen sollte.

Beispiel 3.10. *Sei $h(x) = \mathbf{1}_A(x)$, mit einem bezüglich π seltenen Ereignis A. Dann benötigt $\widehat{\theta}_N$ sehr viele Beobachtungen, um brauchbar zu sein, vgl. den Abschnitt 3.9. Ideal sollte dann τ einfach π eingeschränkt auf A sein, aber das lässt sich nicht durchführen. Es genügt jedoch, g gross auf A und klein ausserhalb A zu wählen.*

"Did Mendel's Facts Fit His Model ?" heisst ein Kapitel in Freedman, Pisani, Purves und Adhikari (1991). Die Antwort ist, dass die Übereinstimmung zwischen Theorie und Daten von Mendel zu gut ist, um noch glaubhaft zu sein. Mit andern Worten, es ist praktisch sicher, dass die Daten von Mendel systematisch geschönt wurden. Die Grundlage für dieses Urteil ist das folgende: Wenn man für jedes Experiment von Mendel die Chiquadrat-Teststatistik berechnet und diese Werte addiert, erhält man den Wert 42, und bei zufälligen Abweichungen hat diese Summe eine Chiquadrat-Verteilung mit 84 Freiheitsgraden. Wie gross ist die Wahrscheinlichkeit, mit dieser Verteilung einen Wert kleiner gleich 42 zu beobachten ? Das findet man in keiner Tabelle.

Die Dichte von χ^2_{84}, d.h. Gamma$(42, \frac{1}{2})$ ist

$$f(x) = \frac{\left(\frac{1}{2}\right)^{42}}{\Gamma(42)}\ x^{41}\ e^{-(\frac{x}{2})}.$$

Wir wählen für G die Gamma$(42, 1)$-Verteilung, welche den Erwartungswert 42 hat. Dann gilt

$$w(x) = \frac{f(x)}{g(x)} = \underbrace{\left(\frac{1}{2}\right)^{42} e^{21}}_{3\cdot 10^{-4}} e^{\frac{(x-42)}{2}}$$

also

$$\mathrm{P}\left[X \leq 42\right] \approx 3 \cdot 10^{-4} \frac{1}{N} \sum_{i=1}^{N} e^{\frac{1}{2}(Y_i - 42)} \mathbf{1}_{[Y_i \leq 42]}.$$

Eine Simulation mit $N = 1000$ ergibt die Näherung $3.6 \cdot 10^{-5}$. Der genaue Wert aus einer (sehr viel aufwändigeren) numerischen Approximation ist $3.54 \cdot 10^{-5}$.

Wenn wir zu h eine Konstante addieren, dann addiert sich beim importance sampling nicht einfach die gleiche Konstante zur Approximation des Erwartungswerts von h: $\widetilde{\theta}_N(h + const) \neq \widetilde{\theta}_N(h) + const$. Dies kann man korrigieren, indem man die folgende Variante betrachtet:

$$\frac{\frac{1}{N}\sum_{i=1}^{N} h(\mathbf{Y}_i) w(Y_i)}{\frac{1}{N}\sum_{i=1}^{N} w(Y_i)}.$$

Dies ist ein Beispiel einer multiplikativen Kontrollvariable, da $\mathbf{E}\left[w(Y_i)\right] = \int w(x)\tau(dx) = 1$. Ob dies besser ist vom Standpunkt der Varianz, hängt von h ab. Wenn f nur bis auf eine Normierungskonstante bekannt ist, muss man auf jeden Fall diese Variante verwenden.

Kapitel 4

Markovketten Monte Carlo (MCMC)

In vielen Fällen, vor allem in hohen Dimensionen, gibt es keine guten Methoden, um gemäss einer beliebigen Verteilung zu simulieren. Der Verwerfungsalgorithmus versagt, weil man praktisch immer verwirft (die Schranke für den Quotienten der Dichten ist zu gross). Importance sampling versagt, weil die Varianz der Gewichte zu gross ist.

In diesem Kapitel besprechen wir die heutige Standardmethode zur Simulation von Verteilungen in hohen Dimensionen. Die Grundidee ist, eine Folge $\mathbf{X}_t$ rekursiv so zu erzeugen, dass $\mathbf{X}_t$ für grosse t ungefähr die richtige Verteilung π hat.

Rekursiv bedeutet, dass $\mathbf{X}_{t+1}$ von $\mathbf{X}_t$ und neuen (uniformen) Zufallsvariablen abhängt, d.h. die erzeugten Variablen bilden eine Markovkette. Die Übergänge der Markovkette werden so gewählt, dass falls $\mathbf{X}_r$ die gewünschte Verteilung π hat, auch alle weiteren Variablen $\mathbf{X}_{r+1}, \mathbf{X}_{r+2}, \ldots$ die Verteilung π haben. Ein solches π nennt man eine invariante Verteilung der Markovkette. Dies rechtfertigt die Approximation

$$\mathbf{E}_\pi [h(X)] \approx \frac{1}{N-r+1} \sum_{t=r}^{N} h(X_t). \tag{4.1}$$

Allerdings sind die Summanden auf der rechten Seite abhängig, was sich auf die Genauigkeit der Approximation auswirkt.

Wir werden die folgenden Punkte diskutieren

- Wie konstruiert man zu gegebenen π eine Übergangsvorschrift, so dass π invariant ist ?
- Konvergiert die Verteilung von $\mathbf{X}_r$ gegen π bei beliebiger Startverteilung ?
- Wie gross sollte r sein, d.h. wie schnell konvergiert die Verteilung von $\mathbf{X}_r$ gegen π ?
- Wie genau ist obige Schätzung des Erwartungswertes $\mathbf{E}_\pi [h(\mathbf{X})]$?

Die erste Frage ist lösbar, und wir werden sehen, dass es sogar viele geeignete Übergangsvorschriften gibt. Die zweite Frage hat ebenfalls eine relativ einfache Antwort. Fragen 3 und 4 sind schwierig explizit zu beantworten, und wir werden sehen, dass sie zusammenhängen. Zuerst stellen wir jedoch ein paar Resultate über Markovketten zusammen.

4.1 Grundbegriffe über Markovketten

Sei $\mathbb{X}$ ein beliebiger Raum mit einer σ-Algebra $\mathcal{F}$. Eine Markovkette beschreibt eine diskrete Zeitentwicklung auf $\mathbb{X}$ mit einfacher Abhängigkeit: Der nächste Zustand hängt nur vom jetzigen Zustand ab, aber nicht von der Vergangenheit. Die bedingte Verteilung des nächsten Zustands gegeben der jetzige Zustand wird durch einen sogenannten *Kern* beschrieben.

Definition 4.1. *Ein Kern P auf $(\mathbb{X}, \mathcal{F})$ in sich ist eine Abbildung von $\mathbb{X} \times \mathcal{F}$ nach $[0, 1]$ derart, dass*

- *$P(x, .)$ ist eine Wahrscheinlichkeit auf $(\mathbb{X}, \mathcal{F})$ für jedes $x \in \mathbb{X}$.*
- *$P(., A)$ ist eine messbare Funktion für jedes $A \in \mathcal{F}$.*

Ein Kern definiert eine Abbildung der Menge der messbaren und beschränkten Funktionen auf $(\mathbb{X}, \mathcal{F})$ in sich mittels

$$Pf(x) = \int P(x, dy) f(y).$$

Ebenso definiert ein Kern eine Abbildung der Menge der Wahrscheinlichkeiten auf $(\mathbb{X}, \mathcal{F})$ in sich mittels

$$\nu P(A) = \int \nu(dx) P(x, A).$$

Ferner kann man zwei Kerne zu einem neuen Kern zusammensetzen mittels der Vorschrift

$$PQ(x, A) = \int P(x, dy) Q(y, A).$$

Das Nachprüfen dieser Behauptungen ist eine Übungsaufgabe in Masstheorie.

Im Fall, wo $\mathbb{X}$ endlich ist, ist ein Kern einfach eine Matrix $(P(i, j))$ mit nichtnegativen Elementen und Zeilensummen gleich 1. Dann entspricht Pf der Multiplikation von P mit einem Spaltenvektor von rechts, νP entspricht der Multiplikation von P mit einem Zeilenvektor von links, und PQ entspricht der Matrixmultiplikation:

$$\begin{array}{rcccl} Pf(i) & = & \sum_{j=1}^{n} P(i,j) f(j) & = & (Pf)(i) \\ \nu P[\{i\}] & = & \sum_{k=1}^{n} \nu(\{k\}) P(k,i) & = & (\nu^T P)(i) \\ PQ(i,j) & = & \sum_{k=1}^{n} P(i,k) Q(k,j) & = & (PQ)(i,j) \end{array}$$

Definition 4.2. *Eine Markovkette auf $(\mathbb{X}, \mathcal{F})$ mit Startverteilung ν_0 und Übergangskern P ist eine Folge $(X_0, X_1, X_2, \ldots)$ von Zufallsvariablen mit Werten in $\mathbb{X}$ derart dass gilt*

$$\mathrm{P}\left[X_0 \in A\right] = \nu_0(A),$$

und

$$\mathrm{P}\left[X_{t+1} \in A \mid X_t = x_t, \ldots, X_0 = x_0\right] = \mathrm{P}\left[X_{t+1} \in A \mid X_t = x_t\right] = P(x_t, A).$$

Eine erste einfache Folgerungen daraus ist

$$\mathrm{P}\left[X_{t+k} \in A \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0\right] = P^k(x_t, A)$$

(P^k ist die k-fache Zusammensetzung des Kerns P mit sich gemäss obiger Definition). Dies beweist man mit Induktion nach k. Im Induktionsschritt bedingt man auf X_{t+k-1} und verwendet das Gesetz der totalen Wahrscheinlichkeit. Analog zeigt man

$$\mathbf{E}\left[f(X_{t+k}) \mid X_t = x_t\right] = P^k f(x_t),$$

und

$$\mathrm{P}\left[X_t \in A\right] = \nu_0 P^t(A).$$

Die gemeinsame Verteilung von $(X_0, X_1, \ldots, X_t)$ ist

$$\nu_0(dx_0) \prod_{s=1}^{t} P(x_{s-1}, dx_s).$$

Definition 4.3. *Eine Wahrscheinlichkeitsverteilung π auf $(\mathbb{X}, \mathcal{F})$ heisst* invariant *oder* stationär *für einen Übergangskern P, falls gilt*

$$\pi P = \pi.$$

Die Bedeutung ist klar: Wenn man π als Startverteilung wählt, dann haben alle X_t die Verteilung π.

Definition 4.4. *Eine Wahrscheinlichkeitsverteilung π auf $(\mathbb{X}, \mathcal{F})$ heisst* reversibel *für einen Übergangskern P, falls gilt*

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx).$$

Dies bedeutet, dass bei Startverteilung π (X_0, X_1) und (X_1, X_0) die gleiche Verteilung haben. Man kann leicht daraus folgern, dass sogar $(X_0, X_1, \ldots X_t)$ und $(X_t, X_{t-1}, \ldots X_0)$ die gleiche Verteilung haben für jedes t, d.h. die Richtung der Zeit spielt keine Rolle. Durch Integration über $\mathbb{X} \times A$ folgt sofort, dass eine reversible Wahrscheinlichkeitsverteilung immer invariant ist. Die Umkehrung gilt im Allgemeinen nicht.

Definition 4.5. *Ein Übergangskern heisst* irreduzibel*, wenn eine Wahrscheinlichkeitsverteilung ψ auf $(\mathbb{X}, \mathcal{F})$ existiert derart, dass $\sum_{k=1}^{\infty} P^k(x, A) > 0$ ist für alle $A \in \mathcal{F}$ mit $\psi(A) > 0$ und alle $x \in \mathbb{X}$.*

Irreduzibilität heisst anschaulich, dass man bei beliebiger Startverteilung mit positiver Wahrscheinlichkeit überall hingelangen kann. Oft kann man reduzible Kerne P_i $(i = 1, \ldots, k)$ so kombinieren, dass der resultierende Kern irreduzibel wird. Die Kombination kann entweder das Hintereinanderausführen in der Reihenfolge $(i(1), i(2), \ldots i(k))$ sein

$$P = P_{i(1)} P_{i(2)} \cdots P_{i(k)}$$

oder die zufällige Auswahl unter den k möglichen Übergängen:

$$P = \frac{1}{k}(P_1 + P_2 + \cdots + P_k).$$

Wenn π invariant ist für alle P_i's, dann ist π auch invariant für P bei beiden Varianten. Hingegen bleibt Reversibilität nur erhalten bei der zweiten Variante.

Irreduzibilität bewirkt, dass eine invariante Verteilung eindeutig ist und dass das Gesetz der grossen Zahlen gilt.

Satz 4.1. *Sei P ein irreduzibler Übergangskern mit einer invarianten Verteilung π. Dann ist π die einzige invariante Verteilung, und es gilt*

- $\mathrm{P}\left[X_t \in A \text{ unendlich oft } \mid X_0 = x\right] > 0$ *für alle* $x \in \mathbb{X}$ *und alle* $A \in \mathcal{F}$ *mit* $\pi(A) > 0$.
- $\mathrm{P}\left[X_t \in A \text{ unendlich oft } \mid X_0 = x\right] = 1$ *für π-fast alle* $x \in \mathbb{X}$ *und alle* $A \in \mathcal{F}$ *mit* $\pi(A) > 0$.
- $\mathrm{P}\left[\frac{1}{n+1}\sum_{t=0}^{n} f(x_t) \to \int f(x)\pi(dx) | X_0 = x\right] = 1$ *für π-fast alle* $x \in \mathbb{X}$ *und alle* f *mit* $\int |f(x)|\pi(dx) < \infty$.

Die Ausnahmemenge für x in den beiden letzten Aussagen ist unerwünscht. Es gibt hinreichende Bedingungen, unter denen die beiden letzten Aussagen sogar für alle x gelten. Eine solche Bedingung ist z. B. dass ein k existiert, so dass $P^k(x, .)$ für alle x eine bezüglich π absolut stetige Komponente hat. Für Beweise verweise ich auf die Literatur.

Unser Ziel ist es $\int h(x)\pi(dx)$ gemäss (4.1) zu approximieren. Dazu müssen wir einen Übergangskern wählen, der folgende drei Bedingungen erfüllt:

1. P ist irreduzibel.
2. π ist invariant, bzw. reversibel für P.
3. Die Simulation gemäss $P(x, .)$ soll einfach sein für alle x.

Solche Kerne konstruieren wir uns im folgenden in schrittweise immer komplexeren Situationen.

4.2 Der Metropolis-Hastings Algorithmus

Wir betrachten zuerst den diskreten Fall. Die Bedingung für Reversibilität lautet dann

$$\pi(i)P(i,j) \;=\; \pi(j)P(j,i). \tag{4.2}$$

Für jedes Paar $i < j$ kann man also entweder $P(i,j)$ oder $P(j,i)$ wählen, die andere Wahrscheinlichkeit ist dann durch (4.2) bestimmt. Allerdings ist dann $\sum_j P(i,j) = 1$ im Allgemeinen nicht erfüllt. Ist die Summe kleiner als Eins, können wir das korrigieren, indem wir $P(i,i)$ modifizieren. Wenn die Summe jedoch grösser ist als Eins, haben wir ein Problem.

Mit etwas Probieren findet man, dass die folgende Konstruktion zum Ziel führt. Man beginnt mit einer beliebigen Übergangsmatrix $Q(i,j)$ derart, dass

$$Q(i,j) > 0 \;\Leftrightarrow\; Q(j,i) > 0. \tag{4.3}$$

Wir können dann für jedes Paar $i < j$ entweder $P(i,j) = Q(i,j)$ oder $P(j,i) = Q(j,i)$ setzen und den andern Wert aus (4.2) bestimmen. Wenn wir dafür sorgen, dass sowohl $P(i,j) \le Q(i,j)$ als auch $P(j,i) \le Q(j,i)$ erfüllt sind, dann ist offensichtlich

$$\sum_{j \neq i} P(i,j) \le \sum_{j \neq i} Q(i,j) \le 1,$$

und wir erhalten also eine Übergangsmatrix, wenn wir noch

$$P(i,i) = 1 - \sum_{j \neq i} P(i,j)$$

setzen.

Wählen wir $P(i,j) = Q(i,j)$, dann ist $P(j,i) = \pi(i)Q(i,j)/\pi(j)$ und dies ist kleiner oder gleich $Q(j,i)$ genau dann, wenn $\pi(i)Q(i,j) \leq \pi(j)Q(j,i)$. Ist dies nicht erfüllt, dann führt die andere der beiden Möglichkeiten zum Ziel. Dies können wir kompakt schreiben, wenn wir für beliebiges $i \neq j$ setzen

$$P(i,j) = \min\left(Q(i,j), \frac{\pi(j)}{\pi(i)}Q(j,i)\right) = Q(i,j)a(i,j)$$

wobei

$$a(i,j) = \min\left(1, \frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}\right) \leq 1.$$

Simulation gemäss dieser Übergangsmatrix $P(i,.)$ ist nicht schwierig. Wir haben folgenden Algorithmus:

Algorithmus 4.1. *1. Wähle $Y \sim Q(i,.)$ und $U \sim$ Uniform$(0,1)$.*

2. Wenn $U \leq a(i,Y)$, dann setze $X = Y$, sonst $X = i$.

Das ist ähnlich wie bei der Verwerfungsmethode, nur machen wir im Fall von Verwerfung keinen neuen Versuch, sondern behalten einfach den momentanen Wert, was der Festlegung $P(i,i) = 1 - \sum_{j \neq i} P(i,j)$ entspricht.

Q heisst *Vorschlagsverteilung (proposal distribution)* und a heisst *Akzeptierungswahrscheinlichkeit*. Man beachte, dass immer eine der beiden Akzeptierungswahrscheinlichkeiten $a(i,j)$ oder $a(j,i)$ gleich eins ist, d.h. man akzeptiert mit grösst möglicher Wahrscheinlichkeit.

Im stetigen Fall muss man sich überlegen, wie man die Akzeptierungswahrscheinlichkeiten definiert und was die Entsprechung zur Bedingung (4.3) ist. Das allgemeine Resultat lautet wie folgt:

Satz 4.2 (Metropolis-Hastings)**.** *Sei π eine Wahrscheinlichkeit auf $(\mathbb{X}, \mathcal{F})$ und Q ein Kern auf dem gleichen Raum so, dass die beiden Wahrscheinlichkeiten $\pi(dx)Q(x,dy)$ und $\pi(dy)Q(y,dx)$ auf $(\mathbb{X}, \mathcal{F}) \times (\mathbb{X}, \mathcal{F})$ äquivalent sind im Sinne der Masstheorie, d.h. die beiden Wahrscheinlichkeiten sollen die gleichen Nullmengen haben. Dann existiert die Radon-Nikodym Dichte von $\pi(dy)Q(y,dx)$ bezüglich $\pi(dx)Q(x,dy)$, welche wir als $r(y,x)$ bezeichnen. Ferner sei $a(x,y) = \min(1, r(y,x))$. Dann ist der folgende Kern reversibel bezüglich π:*

$$P(x,A) = \int_A a(x,y)Q(x,dy) + \mathbf{1}_A(x) \cdot \left(1 - \int_{\mathbb{X}} a(x,y)Q(x,dy)\right) \tag{4.4}$$

Der erste Term in (4.4) ist die Wahrscheinlichkeit, dass man mit dem Kern $Q(x,.)$ einen Wert in A vorschlägt und dass dieser Wert akzeptiert wird. Der zweite Term ist die Wahrscheinlichkeit, dass der Prozess stehenbleibt, weil der Vorschlag nicht akzeptiert wird.

Beweis. Wir müssen zeigen, dass für eine beliebige beschränkte Funktion h gilt

$$\iint h(x,y)\pi(dx)P(x,dy) \;=\; \iint h(x,y)\pi(dy)P(y,dx).$$

Die linke Seite ist gemäss der Definition von P gleich

$$\iint h(x,y)a(x,y)\pi(dx)Q(x,dy) + \int h(x,x)\Big(1-\int_{\mathbb{X}} a(x,y)Q(x,dy)\Big)\pi(dx),$$

und die rechte Seite ist

$$\iint h(x,y)a(y,x)\pi(dy)Q(y,dx) + \int h(y,y)\Big(1-\int_{\mathbb{X}} a(y,x)Q(y,dx)\Big)\pi(dy).$$

Die beiden zweiten Summanden sind offensichtlich gleich (wie man die Variablen bezeichnet, spielt keine Rolle).

Nach Definition von r gilt $r(y,x)r(x,y)=1$. Wenn $r(y,x)\le 1$, dann ist $a(x,y)=r(y,x)$ und $r(x,y)\ge 1$ und somit noch $a(y,x)=1$. Zusammen mit dem analogen Argument im Fall $r(x,y)\ge 1$ folgt daher

$$a(x,y) = r(y,x)a(y,x).$$

Dies impliziert

$$\begin{aligned}\iint h(x,y)a(y,x)\pi(dy)Q(y,dx) &= \iint h(x,y)a(y,x)r(y,x)\pi(dx)Q(x,dy)\\ &= \iint h(x,y)a(x,y)\pi(dx)Q(x,dy).\end{aligned}$$

□

Wie prüft man die Voraussetzung des obigen Satzes nach, und wie berechnet man die Radon-Nikodym Dichte $r(y,x)$? Das folgende Lemma, das eine einfache Übung in Masstheorie ist, deckt die wichtigen Fälle ab:

Lemma 4.1. (i) *Wenn P_1 und P_2 Dichten p_1 bzw. p_2 haben bezüglich einem σ-endlichen Bezugsmass μ, dann ist P_1 absolut stetig bezüglich P_2 genau dann, wenn $\{x \mid p_2(x) = 0, p_1(x) > 0\}$ eine Nullmenge bezüglich μ ist. Die Radon-Nikodym-Dichte von P_1 bezüglich P_2 ist dann $p_1(x)/p_2(x)$, unabhängig von der Wahl von μ.*

(ii) *Wenn P_1 die Dichte r bezüglich P_2 hat und ϕ eine messbare injektive Abbildung ist, dann hat die Verteilung von $y=\phi(x)$ unter P_1 die Dichte $r(\phi^{-1}(y))$ bezüglich der Verteilung von y unter P_2.*

In den einfachsten Beispielen haben $\pi(dx)$ und die Vorschlagsverteilungen $Q(x,dy)$ Dichten $\pi(x)$, bzw. $q(x,y)$ bezüglich dem Lebesguemass im stetigen oder dem Zählmass im diskreten Fall. Dann sind $\pi(dx)Q(x,dy)$ und $\pi(dy)Q(y,dx)$ äquivalent, falls für alle Paare (x,y) gilt

$$\pi(x)q(x,y) > 0 \iff \pi(y)q(y,x) > 0.$$

Dies heisst $q(x,y)=0$ falls $\pi(x)>0$ und $\pi(y)=0$, sowie $q(x,y)>0 \Leftrightarrow q(y,x)>0$. Ferner ist

$$r(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \;\Rightarrow\; a(x,y) = \min\Big(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\Big).$$

Weil nur die Verhältnisse $\pi(y)/\pi(x)$ vorkommen, genügt es, π bis auf eine Normierungskonstante zu kennen.

Ob der Metropolis-Hastings Übergang irreduzibel ist, hängt von Q ab. Irreduzibilität von Q überträgt sich auf P. Insbesondere ist $q(x,y) > 0$ für alle x, y hinreichend (aber nicht notwendig).

Zwei einfache Beispiele sind

Beispiel 4.1 (Independence sampler)**.** *Sei $q(x,y) = q(y)$ für alle x, d.h. das vorgeschlagene Y ist unabhängig von $X_t = x$ und wird akzeptiert mit Wahrscheinlichkeit*

$$a(x,y) = \min\left(1, \frac{\pi(Y)q(x)}{q(Y)\pi(x)}\right).$$

Dies ist ähnlich wie beim Verwerfungsalgorithmus und beim importance sampling. Wenn wir auch X_0 mit Verteilung Q wählen, dann ist

$$\frac{1}{N+1}\sum_{t=0}^{N} h(X_t) = \sum_{i=1}^{n} w_i h(Y_i)$$

wobei jedes Y_i gemäss Q erzeugt wurde und w_i die relative Häufigkeit von Y_i in $(X_0, X_1, \ldots X_N)$ angibt. Im Unterschied zum importance sampling sind die Y_i jedoch abhängig.

Beispiel 4.2 (Random walk Metropolis)**.** *Sei $\mathbb{X} = \mathbb{R}^p$ und $q(x,y) = q(y-x)$ mit $q(x) = q(-x)$. Das heisst also, wenn $X_t = x$, dann ist $Y = x + \epsilon$ mit $\epsilon \sim q(z)dz$ unabhängig von x; es handelt sich um eine Irrfahrt im $\mathbb{R}^p$. Es ist dann*

$$a(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right),$$

d.h. wenn die Wahrscheinlichkeit des vorgeschlagenen Werts grösser ist als die des aktuellen, akzeptiert man den vorgeschlagenen Wert sicher, sonst nur mit einer gewissen Wahrscheinlichkeit < 1.

4.2.1 Komponentenweise Modifikation

In vielen Fällen ist es ungünstig, eine Vorschlagsverteilung $Q(x, dy)$ mit einer Dichte zu wählen. Das bedeutet nämlich, dass der vorgeschlagene Wert irgendwo im Raum liegen kann. Wenn der aktuelle Wert x ein plausibler Wert für π ist und der Raum $\mathbb{X}$ hochdimensional, dann wird der vorgeschlagene Wert praktisch immer unplausibler sein als der aktuelle, d.h. man verwirft praktisch sicher. In solchen Fällen ist es wesentlich besser, wenn sich der vorgeschlagene Wert vom aktuellen nur in wenigen Komponenten unterscheidet.

Wir formulieren das Vorgehen im Fall eines Produktraums von zwei Komponenten (wobei die zweite Komponente natürlich wieder ein Produktraum sein kann). Sei also $\mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2$, d.h. $x \in \mathbb{X}$ hat die Form $(x_1 x_2)$ mit $x_k \in \mathbb{X}_k$. Wir nehmen ferner an, dass π absolut stetig ist bezüglich einem Produktmass $\mu_1(dx_1)\mu_2(dx_2)$ mit einer Dichte, die wir ebenfalls als π bezeichnen. Wir betrachten dann eine Vorschlagsverteilung Q, welche nur die erste Komponente modifiziert mit einer absolut stetigen Verteilung und die zweite unverändert lässt:

$$Q(x, A_1 \times A_2) = \int_{A_1} q(x, y_1)\mu_1(dy_1)\mathbf{1}_{A_2}(x_2).$$

Dann sind $\pi(dx)Q(x,dy)$ und $\pi(dy)Q(y,dx)$ konzentriert auf die Menge von Paaren mit $x_2 = y_2$, und die drei Komponenten (x_1, x_2, y_2) haben die Dichten $\pi((x_1x_2))q((x_1x_2), y_1)$ bzw. $\pi((y_1x_2))q((y_1x_2), x_1)$. Wenn wir setzen $\phi(x_1, x_2, y_1) = (x_1, x_2, y_1, x_2)$, dann zeigt die zweite Aussage von Lemma 4.1, dass die Bedingungen des Satzes 4.2 erfüllt sind und

$$r(x,y) = \frac{\pi((x_1x_2))q((x_1x_2), y_1)}{\pi((y_1x_2))q((y_1x_2), x_1)} \quad \text{auf } y_2 = x_2$$

falls Zähler und Nenner gleichzeitig nicht null sind. (Für Paare mit $x_2 \neq y_2$ ist die Dichte $r(x,y)$ nicht definiert, man braucht sie auch nicht). Dieser Ausdruck lässt sich auch schreiben mit Hilfe der bedingten Dichte $\pi_{1|2}(x_1 \mid x_2)$ der ersten Komponente gegeben die zweite:

$$r(x,y) = \frac{\pi_{1|2}(x_1 \mid x_2)q((x_1x_2), y_1)}{\pi_{1|2}(y_1 \mid x_2)q((y_1x_2), x_1)}.$$

Wenn man nur eine Komponente modifiziert, erhält man natürlich nie einen irreduziblen Kern. Man kann aber analog einen zweiten Kern betrachten, der nur die zweite Komponente modifiziert, und dann die beiden Kerne in abwechselnder Reihenfolge ausführen. Ebenso kann man statt 2 auch k Komponenten betrachten und für jede Komponente einen Kern, der den Rest festhält. Die Kombination kann gemäss einer festen oder einer zufälligen Reihenfolge erfolgen. Bei einer festen Reihenfolge geht meistens die Reversibilität verloren, aber die stationäre Verteilung ändert sich nicht.

Als Vorschlagsdichte können wir insbesondere

$$q(x, y_1) = q(x_2, y_1) = \pi_{1|2}(y_1 \mid x_2) \tag{4.5}$$

verwenden. Dann ist die Radon-Nikodym Dichte r stets eins, und damit auch die Akzeptierungswahrscheinlichkeit. Kombination dieser Kerne ist nichts anderes als der bereits im ersten Kapitel erwähnte Gibbs-Sampler.

Wir können aber auch die Idee vom random walk Metropolis Algorithmus verallgemeinern und eine Vorschlagsdichte der Form

$$q(x, y_1) = q(x_1, y_1) = q(y_1 - x_1)$$

verwenden. Wenn der random walk symmetrisch ist, sind die Akzeptierungswahrscheinlichkeiten gleich

$$\min\left(1, \frac{\pi_{1|2}(y_1 \mid x_2)}{\pi_{1|2}(x_1 \mid x_2)}\right).$$

4.2.2 Metropolis-Hastings auf dem Raum der Sprungfunktionen

Der komplizierteste Fall, den wir besprechen, ist der, wo $\mathbb{X}$ die Vereinigung von Teilräumen verschiedener Dimensionen ist und wo die Markovkette zwischen diesen Teilräumen hin und her springt. Wir erklären die Problemstellung und die Idee in diesem Abschnitt zunächst am Beispiel der Simulation zufälliger, stückweise konstanter Funktionen auf $[0,1]$. Das heisst, wir betrachten den Raum

$$\mathbb{X} = \cup_{k=0}^{\infty} \mathbb{X}_k,$$

wobei $\mathbb{X}_k$ der Raum der stückweise konstanten Funktionen mit genau k Sprüngen bezeichnet. Wir parametrisieren die Elemente von $\mathbb{X}_k$ durch die Sprungstellen $(t_i; i = 1, \ldots, k)$

und die Funktionswerte $(g_i; i = 1, \ldots, k+1)$, d.h.

$$x(t) = \sum_{i=1}^{k+1} g_i \mathbf{1}_{(t_{i-1}, t_i]}(t)$$

mit $t_0 = 0$ und $t_{k+1} = 1$. Der Raum $\mathbb{X}_k$ ist also eine Teilmenge des $\mathbb{R}^{2k+1}$, und wir werden x und $((t_i), (g_i))$ identifizieren. Wir nehmen an, dass die Verteilung π, von der wir simulieren wollen, auf jedem $\mathbb{X}_k$ eine Dichte π_k bezüglich des Lebesgue-Masses auf $\mathbb{R}^{2k+1}$ hat.

Diese Situation tritt z.B. in der Bayes'schen nichtparametrischen Regression auf. Gegeben seien Beobachtungen

$$Y_i = x(s_i) + \varepsilon_i \quad (i = 1, \ldots, n)$$

mit vorgegebenen Beobachtungspunkten s_i (z.B. gleichabständig, d.h. $s_i = (i - 0.5)/n$) und i.i.d. Fehlern $\varepsilon_i \sim \mathcal{N}(0, 1)$. Unbekannt ist die Mittelwertfunktion x. Zur Vereinfachung haben wir die Varianz der Fehler ε_i als bekannt vorausgesetzt, aber es wäre kein Problem, diese als zusätzlichen unbekannten Parameter zu behandeln. Als a priori Verteilung für x wählen wir eine Verteilung auf unserem Raum $\mathbb{X}$ der Sprungfunktionen. Wir können z.B. für die Anzahl Sprünge eine Pois (λ)-Verteilung und – gegeben die Anzahl Sprünge – die Sprungstellen t_i uniform und die Funktionswerte g_i i.i.d. $\mathcal{N}\left(0, \tau^2\right)$-verteilt wählen. Die Dichte von π auf $\mathbb{X}_k$ ist dann

$$\pi_k(x) = \exp(-\lambda)\frac{\lambda^k}{k!}\ k!\ \mathbf{1}_{[t_1 < t_2 < \ldots < t_k]} \frac{1}{(\sqrt{2\pi}\tau)^{k+1}} \exp\Big(- \frac{1}{2\tau^2} \sum_{i=1}^{k+1} g_i^2 \Big).$$

Die a posteriori Verteilung gegeben die Beobachtungen $Y_j = y_j$ hat dann auf $\mathbb{X}_k$ die Dichte

$$\pi_k(x \mid (y_j)) = \text{const}\, \pi_k(x) \frac{1}{(2\pi)^{n/2}} \exp\Big(- \frac{1}{2} \sum_{i=1}^{k+1} \sum_{j=1}^{n} \mathbf{1}_{(t_{i-1}, t_i]}(s_j)(y_j - g_i)^2 \Big).$$

Wir wollen das Metropolis-Hastings Rezept verwenden. Das heisst wir schlagen Übergänge gemäss einer Verteilung Q vor und sorgen dann durch eine geeignete Akzeptierungswahrscheinlichkeit dafür, dass die vorgegebene Verteilung π reversibel wird. Damit wir uns im ganzen Raum $\mathbb{X}$ bewegen können, müssen wir Übergänge von $\mathbb{X}_k$ nach $\mathbb{X}_j$ mit $j \neq k$ haben. Die einfachsten Übergänge gehen von $\mathbb{X}_k$ nach $\mathbb{X}_{k-1}$ und $\mathbb{X}_{k+1}$ und modifizieren die Sprungfunktion x nicht überall, sondern es wird nur ein Sprung hinzugefügt oder weggelassen. (Weil die beiden Funktionen teilweise übereinstimmen, sind die Übergänge sicher nicht absolut stetig). Konkret schlagen wir für ein $x = ((t_i), (g_i)) \in \mathbb{X}_k$ ein $z = ((r_i), (h_i)) \in \mathbb{X}_{k+1}$ vor gemäss folgendem Algorithmus:

1. Wähle das j-te Intervall $I_j = (t_{j-1}, t_j]$ zur Unterteilung aus mit Wahrscheinlichkeit $t_j - t_{j-1}$ (Lange Intervalle haben eine grössere Wahrscheinlichkeit, unterteilt zu werden). Setze dann $r_i = t_i$ und $h_i = g_i$ für $i < j$, $r_i = t_{i-1}$ für $i > j$ und $h_i = g_{i-1}$ für $i > j + 1$.

2. Wähle den neuen Sprungpunkt r_j uniform auf I_j.

3. Wähle die beiden neuen Funktionswerte h_j und h_{j+1} gemäss einer Dichte f, untereinander und von r_j unabhängig.

Damit ist ein Übergangskern $Q_k^+(x, dz)$ von $\mathbb{X}_k$ nach $\mathbb{X}_{k+1}$ festgelegt. Die Verteilung $\pi_k(dx)Q_k^+(x, dz)$ auf $\mathbb{X}_k \times \mathbb{X}_{k+1}$ ist dabei konzentriert auf die Vereinigung der

$$A_{j,k} = \{(x, z) \mid x(t) = z(t)\ \forall t \notin (t_{j-1}, t_j]\} \quad (j = 1, 2, \ldots, k+1).$$

Jedes Paar (x, z) in A_{jk} hat $2k+4$ freie Komponenten, also ist A_{jk} das Bild einer offenen Teilmenge im $\mathbb{R}^{2k+4}$. Welche $2k+4$ Komponenten wir wählen, spielt keine Rolle. Wenn wir die Komponenten von x, den neuen Sprungpunkt und die beiden neuen Sprunghöhen wählen, dann hat $\pi_k(dx)Q_k^+(x, dz)$ für diese Komponenten die Dichte

$$\pi_k(t_1, \ldots t_k, g_1, \ldots g_{k+1})\ (t_j - t_{j-1})\ \mathbf{1}_{(t_{j-1}, t_j]}(r_j) \frac{1}{t_j - t_{j-1}}\ f(h_j)\ f(h_{j+1}). \tag{4.6}$$

Damit wir den Satz 4.2 anwenden können, muss $\pi_{k+1}(dz)Q_{k+1}^-(z, dx)$ auf den gleichen Mengen A_{jk} konzentriert sein und darauf ebenfalls eine Dichte haben. $Q_{k+1}^-(z, dx)$ muss also genau einen der $k+1$ Sprünge entfernen (wobei jeder Sprung positive Wahrscheinlichkeit haben muss, entfernt zu werden), und die neue Sprunghöhe muss gemäss einer Dichte gewählt werden.

Konkret legen wir einen Übergangskern $Q_{k+1}^-(z, dx)$ von $\mathbb{X}_{k+1}$ nach $\mathbb{X}_k$ fest gemäss folgendem Rezept

1. Wähle die zu eliminierende Sprungstelle r_j zufällig, d.h. j uniform auf $\{1, \ldots k+1\}$. Setze dann $t_i = r_i$ und $g_i = h_i$ für $i < j$, $t_i = r_{i+1}$ für $i \geq j$ und $g_i = h_{i+1}$ für $i > j$.
2. Wähle den Funktionswert g_j mit Dichte f.

Die Dichte von $\pi_{k+1}(dz)Q_{k+1}^-(z, dx)$ auf A_{jk} ist dann gleich

$$\pi_{k+1}(t_1, \ldots, t_{j-1}, r_j, t_j, \ldots, t_k, g_1, \ldots g_{j-1}, h_j, h_{j+1} g_{j+1}, \ldots g_{k+1}) \frac{1}{k+1} f(g_j). \tag{4.7}$$

(Wir nehmen die gleichen Variablen zur Beschreibung von A_{jk} wie oben bei (4.6).)

Um zu entscheiden, ob wir einen Sprung hinzufügen oder eliminieren, werfen wir eine Münze mit Parameter β_k (natürlich ist $\beta_0 = 1$). Danach verfahren nach obigen Algorithmen. Damit ist unsere Vorschlagsverteilung Q vollständig definiert. In Formeln gilt

$$Q(x, dz) = \beta_k Q_k^+(x, dz)\mathbf{1}_{\mathbb{X}_{k+1}}(z) + (1 - \beta_k)Q_k^-(x, dz)1_{\mathbb{X}_{k-1}}(z) \text{ für } x \in \mathbb{X}_k.$$

Die Bedingung von Satz 4.2 ist dann erfüllt: $\pi(dx)Q(x, dz)$ und $\pi(dz)Q(x, dz)$ sind konzentriert auf den Mengen A_{jk}, und gemäss Lemma 4.1 (ii) ist die Radon-Nikodym-Dichte auf einem A_{jk} gleich dem Quotienten von β_k mal die Dichte (4.6) und von $(1 - \beta_{k+1})$ mal die Dichte (4.7). Damit lautet die Akzeptierungswahrscheinlichkeit

$$a(x, z) = \min\Big(1, \frac{\pi_{k+1}(z)(1 - \beta_{k+1})f(g_j)}{\pi_k(x)\beta_k f(h_j)f(h_{j+1})(k+1)}\Big) \quad ((x, z) \in A_{jk}).$$

Für die Simulation von der a-posteriori Verteilung ersetzen wir einfach $\pi_k(x)$, bzw. $\pi_{k+1}(z)$ durch $\pi_k(x \mid y)$, bzw. $\pi_{k+1}(z \mid y)$.

Der oben beschriebene Übergangsmechanismus ist natürlich nicht der einzig mögliche. Manchmal ist es von Vorteil, nur Modifikationen vorzuschlagen, welche das Mittel $\int_0^1 x(t)dt$ konstant lassen. Dies leistet der folgende Algorithmus:

1. Wähle das j-te Intervall $I_j = (t_{j-1}, t_j]$ zur Unterteilung aus mit Wahrscheinlichkeit $t_j - t_{j-1}$. Setze dann $r_i = t_i$ und $h_i = g_i$ für $i < j$, $r_i = t_{i-1}$ für $i > j$ und $h_i = g_{i-1}$ für $i > j + 1$.

2. Wähle den neuen Sprungpunkt r_j uniform auf I_j.

3. Wähle die beiden neuen Funktionswerte

$$h_j = g_j + \frac{u}{r_j - t_{j-1}}, \quad h_{j+1} = g_j - \frac{u}{t_j - r_j},$$

wobei u die Dichte f hat und von r_j unabhängig ist.

Damit ist ein anderer Übergangskern $Q_k^+(x, dz)$ von $\mathbb{X}_k$ nach $\mathbb{X}_{k+1}$ festgelegt. Die Verteilung $\pi_k(dx)Q_k^+(x, dz)$ auf $\mathbb{X}_k \times \mathbb{X}_{k+1}$ ist jetzt konzentriert auf die Vereinigung der

$$B_{j,k} = \{(x, z) \mid x(t) = z(t)\ \forall t \notin (t_{j-1}, t_j],\ \int_0^1 x(t)dt = \int_0^1 z(t)dt\}$$

Jedes Paar (x, z) in B_{jk} hat $2k + 3$ freie Komponenten, z.B. können wir die Komponenten von x, den neuen Sprungpunkt r_j und die oben definierte Variable u wählen. Für diese Variablen hat $\pi_k(dx)Q_k^+(x, dz)$ die Dichte

$$\pi_k(t_1, \ldots t_k, g_1, \ldots g_{k+1})\ (t_j - t_{j-1})\ \mathbf{1}_{(t_{j-1}, t_j]}(r_j)\frac{1}{t_j - t_{j-1}}\ f(u). \tag{4.8}$$

Damit wir den Satz 4.2 anwenden können, muss wieder $\pi_{k+1}(dz)Q_{k+1}^-(z, dx)$ auf den gleichen Mengen B_{jk} konzentriert sein und darauf ebenfalls eine Dichte haben. $Q_{k+1}^-(z, dx)$ muss also genau einen der $k + 1$ Sprünge entfernen und die neue Sprunghöhe muss gleich dem gewichteten Mittel der beiden alten Sprunghöhen sein. Bis auf die Wahl des zu entfernenden Sprungs, welche z.B. gleichverteilt sein kann, ist der Übergang also deterministisch, und die Dichte von $\pi_{k+1}(dz)Q_{k+1}^-(z, dx)$ auf B_{jk} ist damit im Wesentlichen die Dichte $\pi_{k+1}(z)$. Zur Berechnung der Radon-Nikodym-Dichte müssen wir jedoch die gleichen Variablen zur Beschreibung von B_{jk} wie oben bei (4.8) nehmen, d.h. wir müssen die Dichte $\pi_{k+1}(z)$ für die Variablen $((r_i), (h_i))$ umrechnen auf die Dichte für die Variablen $((t_i), (g_i), r_j, u)$. Dazu benutzen wir die Umrechnungsformel für Dichten (Satz 3.4). Für die meisten Variablen besteht die Transformation einfach in einem Wechsel der Bezeichnung. Das Einzige, was nicht trivial ist, ist der Zusammenhang zwischen (h_j, h_{j+1}) und (g_j, u), und der ergibt die Funktionaldeterminante $(t_j - t_{j-1})/((t_j - r_j)(r_j - t_{j-1}))$. Damit ist die gesuchte Dichte gleich

$$\pi_{k+1}(z)\frac{1}{k+1}\frac{t_j - t_{j-1}}{(t_j - r_j)(r_j - t_{j-1})}, \tag{4.9}$$

wobei man z ausdrücken soll mit Hilfe von $((t_i), (g_i), r_j, u)$.

4.2.3 Allgemeine Übergänge zwischen Räumen unterschiedlicher Dimension

Wir verallgemeiner nun das Vorgehen, das wir im vorangegangenen Abschnitt gesehen haben. Es sei

$$\mathbb{X} = \cup_{k=0}^{\infty}\mathbb{R}^k,$$

und π habe auf jedem Teil $\mathbb{R}^k$ eine strikt positive Dichte π_k. Wir betrachten Übergangskerne $Q_{km}(x, dz)$ von $\mathbb{R}^k$ nach $\mathbb{R}^m$ der folgenden Art:

$$x \to z = z(x, U_{km}),$$

wobei U_{km} eine d_{km} dimensionale Zufallsvariable mit überall positiver Dichte f_{km} bezüglich des Lebesguemasses ist und der Zusammenhang $z = z(x, u_{km})$ deterministisch ist. Dann ist also die Verteilung $\pi_k(dx)Q_{km}(x, dz)$ konzentriert auf eine $k+d_{km}$-dimensionale Fläche im $\mathbb{R}^{k+m}$, nämlich auf alle Paare der Form $(x, z(x, u_{km}))$, und die Dichte von (x, u_{km}) ist $\pi_k(x)f_{km}(u_{km})$.

Um die Voraussetzungen des Satzes 4.2 zu erfüllen, müssen wir daher einen Übergang $Q_{mk}(z, dx)$ von $\mathbb{R}^m$ nach $\mathbb{R}^k$ vorsehen, so dass $\pi_m(dz)Q_{mk}(z, dx)$ auf die gleiche Fläche konzentriert ist. Wenn $Q_{mk}(z, dx)$ die gleiche Struktur hat wie $Q_{km}(x, dz)$, d.h. wenn sich x deterministisch aus (z, u_{mk}) berechnet, ist $\pi_m(dz)Q_{mk}(z, dx)$ konzentriert auf die Fläche $(z, x(z, u_{mk}))$. Also müssen zwei Parametrisierungen der gleichen Fläche vorliegen: Die Dimensionen müssen übereinstimmen:

$$k + d_{km} = m + d_{mk},$$

und es muss eine Bijektion

$$(x, u_{km}) \leftrightarrow (z, u_{mk})$$

geben. Für die Radon-Nikodym-Dichte schliesslich müssen wir die Dichte $\pi_m(z)f_{mk}(u_{mk})$ für (z, u_{mk}) umrechnen auf die Dichte für (x, u_{km}), d.h. wir müssen noch mit der Funktionaldeterminante

$$\left|\frac{\partial(z, u_{mk})}{\partial(x, u_{km})}\right|$$

multiplizieren.

Zur Komplettierung der Vorschlagsverteilung gehört noch die Wahl der Dimension m . Dies soll mit einer stochastischen Matrix (β_{km}) geschehen, was zu folgendem Übergangskern von $\mathbb{X}$ in sich führt:

$$Q(x, dz) = \sum_{j=0}^{\infty} \beta_{kj} Q_{kj}(x, dz)\mathbf{1}_{\mathbb{R}^j}(z) \quad (x \in \mathbb{R}^k).$$

Zusammenfassend geben wir die Formel für die Akzeptierungswahrscheinlichkeiten an, die sich aus obigen Überlegungen ergibt:

$$a(x, z) = \min\left(1, \frac{\pi_m(z)\beta_{mk}f_{mk}(u_{mk})}{\pi_k(x)\beta_{km}f_{km}(u_{km})}\left|\frac{\partial(z, u_{mk})}{\partial(x, u_{km})}\right|\right) \quad (x \in \mathbb{R}^k, z = z(x, u_{km}) \in \mathbb{R}^m).$$

4.3 Genauigkeit von MCMC Approximationen

Wir betrachten nun den Fall, wo $X_1, X_2, X_3, \ldots$ abhängig und nicht identisch verteilt sind. Wenn X_t nur asymptotisch für $t \to \infty$ die Verteilung π hat und wir trotzdem den Schätzer

$$\hat{\theta}_N = \frac{1}{N}\sum_{t=1}^{N} h(X_t)$$

für $\theta = \int h(x)\pi(dx)$ verwenden, dann machen wir einen systematischen Fehler:

$$\mathbf{E}\left[\hat{\theta}_N\right] = \frac{1}{N}\sum_{t=1}^{N}\mathbf{E}\left[h(X_t)\right] \neq \theta.$$

Dieser systematische Fehler ist von der Grössenordnung $O(1/N)$, sofern

$$\sum_{t=1}^{\infty}\left|\mathbf{E}\left[h(X_t)\right] - \int h(x)\pi(dx)\right| < \infty.$$

Abhängigkeit der X_t bewirkt ferner, dass

$$\mathrm{Var}\left(\hat{\theta}_N\right) = \frac{1}{N^2}\sum_{s=1}^{N}\sum_{t=1}^{N}\mathrm{Cov}\left(h(X_s), h(X_t)\right),$$

d.h. auch die Verteilung des Zufallsfehlers ist nicht mehr gleich wie im vorherigen Abschnitt.

Der mittlere quadratische Fehler (MSE) berücksichtigt sowohl den Bias als auch den Zufallsfehler:

$$\mathbf{E}\left[(\hat{\theta}_N - \theta)^2\right] = \left(\mathbf{E}\left[\hat{\theta}_N\right] - \theta\right)^2 + \mathrm{Var}\left(\hat{\theta}_N\right).$$

Für eine Genauigkeitsangabe braucht man eine Abschätzung sowohl für den Bias als auch für die Varianz. Dies sind leider ziemlich schwierige Probleme. Wir werden sehen, dass typischerweise $\mathrm{Var}\left(\hat{\theta}_N\right)$ immer noch von der Ordnung $O(1/N)$ ist. Wie oben erwähnt, ist der Bias typischerweise von der $O(1/N)$, und dann ist sein Beitrag zum mittleren quadratischen Fehler asymptotisch vernachlässigbar (denn dort geht der Bias im Quadrat ein). Das muss aber im endlichen Fall noch nicht viel heissen.

Für den Bias begnügt man sich oft mit grafischen Hilfsmitteln, das heisst man versucht aus einem Plot von $h(X_t)$ gegen t herauszufinden, von welchem Zeitpunkt t_0 an keine systematischen Abweichungen mehr auftreten. Dann verwendet man nur die Werte von diesem t_0 an und ignoriert dafür den Bias.

Wir machen hier zuerst ein paar theoretische Überlegungen zu Bias und Varianz im Fall von Markovketten Monte Carlo, und diskutieren dann die Behandlung der Abhängigkeit im stationären Fall, wo alle X_t die gewünschte Verteilung π haben, aber nicht unabhängig sind.

4.3.1 Konvergenzresultate bei Markovketten

Wir diskutieren hier den Bias einer Markovketten Monte Carlo Methode. Sei (X_i) eine Markovkette mit Startverteilung ν_0, Übergang P und invarianter Verteilung π. Wir möchten abschätzen, wie rasch

$$\mathbf{E}\left[h(X_t)\right] - \int h(x)\pi(dx) = \int P^t h(x)\nu_0(dx) - \int h(x)\pi(dx)$$

gegen Null knovergiert.

Wir beschränken uns auf den einfachsten Fall, wo π eine Wahrscheinlichkeit auf dem diskreten Raum $\{1, 2, \ldots n\}$ ist. Ein algebraischer Zugang benutzt ein Resultat über die

Eigenwerte der Übergangsmatrix P: Der Satz von Frobenius besagt, dass der Eigenwert mit maximalem Absolutbetrag einer irreduziblen und aperiodischen stochastischen Matrix gleich 1 ist und dass dessen Vielfachheit 1 ist. Die Konvergenzgeschwindigkeit ist dann bestimmt durch den Eigenwert mit dem zweitgrössten Absolutbetrag.

Wir behandeln hier eine stochastische Methode, die *Kopplung* von Markovketten. Dies bedeutet, dass man einen Markovprozess $(X_t^{(\mu)}), X_t^{(\nu)})$ auf dem Produkt $\{1, 2, \dots n\}^2$ mit den folgenden Eigenschaften konstruiert: Für sich betrachtet sind $(X_t^{(\mu)})$ und $(X_t^{(\nu)})$ Markovketten ist Übergangsmatrix P und Startverteilungen μ, bzw. ν. Diese beiden Ketten sind jedoch *abhängig*, und zwar so, dass sie zusammen bleiben, nachdem sie sich das erste Mal getroffen haben, d.h.

$$\mathrm{P}\left[X_t^{(\mu)} = X_t^{(\nu)} = j \mid X_{t-1}^{(\mu)} = X_{t-1}^{(\nu)} = i\right] = P(i,j).$$

Wie wir die Übergänge durchführen solange $X_{t-1}^{(\mu)} \neq X_{t-1}^{(\nu)}$ ist, ist nicht festgelegt. Am einfachsten führen wir die Übergänge der beiden Ketten unabhängig voneinander durch. Wir können sie aber auch abhängig machen und so die Chancen zu erhöhen, dass sie sich treffen. Auf jeden Fall haben wir die folgende Abschätzung für den Unterschied zwischen den Verteilungen μP^t und νP^t.

Lemma 4.2. *Wenn $(X_t^{(\mu)}, X_t^{(\nu)})$ die obigen Eigenschaften hat, dann gilt*

$$\sum_j |\mu P^t(j) - \nu P^t(j)| \leq 2\,\mathrm{P}\left[X_t^{(\mu)} \neq X_t^{(\nu)}\right].$$

Wenn wir $\mu = \pi$ und $\nu = \nu_0$ wählen, dann ergibt das Lemma 4.2 die gewünschte Abschätzung für den Bias

$$|\,\mathbf{E}\,[h(X_t)] - \int h(x)\pi(dx)| = |\sum_j (\pi P^t(j) - \nu_0 P^t(j))h(j)| \leq \max_i |h(i)| \sum_j |\mu P^t(j) - \nu P^t(j)|.$$

Der Beweis von Lemma 4.2 beruht auf folgendem Lemma

Lemma 4.3. *Für zwei Wahrscheinlichkeiten P und Q auf $\mathbb{N}$ gilt*

$$\frac{1}{2}\sum_j |p(j) - q(j)| = \sum_j (q(j) - p(j))_+ = (1 - \sum_j \min(p(j), q(j)))$$

$$= \sup_A |P(A) - Q(A)| = \min\left\{\sum_{i \neq j} r(i,j);\ r \geq 0,\ \sum_j r(i,j) = p(i),\ \sum_j r(j,i) = q(i)\right\}.$$

(x_+ ist der positive Teil von x, d.h. $x_+ = \max(x, 0)$.)

Der letzte Ausdruck ist nichts anderes als das Minimum von $\mathrm{P}\,[X \neq X']$ über alle gemeinsamen Verteilungen von (X, X') derart, dass $X \sim P$ und $X' \sim Q$. Daraus folgt sofort das Lemma 4.2. Diejenige Verteilung r, welche das Minimum im letzten Ausdruck realisiert, nennen wir die optimale Kopplung von P und Q.

Beweis. Aus $\sum_j (p(j) - q(j)) = 0$ folgt, dass

$$\sum_j (q(j) - p(j))_+ = \sum_j (p(j) - q(j))_+ = \frac{1}{2}\sum_j |p(j) - q(j)|,$$

woraus die erste Gleichung folgt. Für die zweite Gleichung beachtet man, dass

$$(q(j) - p(j))_+ = q(j) - \min(p(j), q(j)).$$

Die dritte Gleichung gilt, weil für beliebiges A

$$-\sum_{j;p(j)<q(j)} (q(j) - p(j)) \leq \sum_{j\in A}(p(j) - q(j)) \leq \sum_{j;p(j)>q(j)} (p(j) - q(j)).$$

Andererseits folgt aus $\sum_j (p(j) - q(j)) = 0$, dass

$$\sum_{j;p(j)<q(j)} (q(j) - p(j)) = \sum_{j;p(j)>q(j)} (p(j) - q(j)) = \frac{1}{2}\sum_j |p(j) - q(j)|.$$

Die letzte Gleichung beweisen wir über zwei Ungleichungen. Für jedes r, das den angegebenen Bedingungen genügt, und für jedes A gilt

$$|P(A) - Q(A)| = |\sum_{i,j} r(i,j)(1_A(i) - 1_A(j))| \leq \sum_{i,j} r(i,j)|1_A(i) - 1_A(j)| \leq \sum_{i\neq j} r(i,j).$$

Damit haben wir die eine Ungleichung. Für die andere Ungleichung wählen wir r wie folgt

$$r(i,j) = \min(p(i), q(i))\delta_{i,j} + \frac{(p(i) - q(i))_+(q(j) - p(j))_+}{\sum_k (p(k) - q(k))_+}.$$

Man rechnet leicht nach, dass dieses r die Bedingugen erfüllt. Offensichtlich gilt

$$\sum_{i\neq j} r(i,j) = 1 - \sum_i r(i,i) = 1 - \sum_i \min(p(i), q(i)).$$

□

Um gemäss der optimalen Kopplung r zu simulieren, beachten wir, dass sich dieses r als eine Mischung schreiben lässt

$$r(i,j) = \gamma\frac{\min(p(i)q(i))\delta_{ij}}{\gamma} + (1-\gamma)\frac{(p(i) - q(i))_+}{1-\gamma}\frac{(q(j) - p(j))_+}{1-\gamma}$$

mit $\gamma = \sum_i \min(p(i), q(i))$. Das heisst, mit Wahrscheinlichkeit γ erzeugen wir $X = X'$ gemäss der Verteilung $(\min(p(i), q(i))/\gamma)$, und mit Wahrscheinlichkeit $1 - \gamma$ sind X und X' unabhängig mit den Verteilungen $((p(i) - q(i))_+/(1-\gamma))$, bzw. $((q(i) - p(i))_+/(1-\gamma))$.

Um zu einer konkreten Abschätzung von $\mathrm{P}\left[X_t^{(\mu)} \neq X_t^{(\nu)}\right]$ zu kommen, nehmen wir an, dass es einen Zustand j_0 gibt, der von allen andern Zuständen mit positiver Wahrscheinlicheit erreicht werden kann:

$$P(i, j_0) \geq \varepsilon > 0 \quad \forall i.$$

Wenn die Übergänge unabhängig erfolgen, so lange sich die Ketten noch nicht getroffen haben, dann treffen sich die Ketten bei jedem Schritt mit Wahrscheinlicheit grösser oder gleich ε^2 im Zustand j_0. Damit folgt

$$\mathrm{P}\left[X_t^{(\mu)} \neq X_t^{(\nu)}\right] \leq (1 - \varepsilon^2)^t,$$

d.h. wir haben exponentielle Konvergenz. Schärfere Aussagen bekommt man, wenn man sowohl zur Zeit $t = 0$ als auch bei jedem Übergang die optimale Kopplung vornimmt. Dann erhalten wir

$$\mathrm{P}\left[X_t^{(\mu)} \neq X_t^{(\nu)}\right] \leq \frac{1}{2}\sum_{j=1}^{n} |\mu(j) - \nu(j)|(1-\alpha)^t$$

wobei

$$\alpha = \frac{1}{2}\max_{i,k}\sum_{j=1}^{n} |P(i,j) - P(k,j)| \geq \varepsilon.$$

Was passiert, wenn es keinen Zustand gibt, der von überall mit positiver Wahrscheinlichkeit erreicht werden kann ? Wir diskutieren diesen Fall anhand des folgenden Beispiels:

Beispiel 4.3. *Betrachte die Irrfahrt auf* $\{1,2,3,4,5\}$ *mit Reflektion am Rand:*

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

und $\pi = (\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$. *Wenn wir die Übergänge für alle Ketten mit der gleichen Schrittrichtung* U_t *durchführen, dann treffen sich die Ketten sicher, sobald vier Mal hintereinander die gleiche Richtung gewählt wird. In andern Worten, für beliebige* i, j

$$\mathrm{P}\left[X_t^{(\mu)} = X_t^{(\nu)} \mid X_{t-4}^{(\mu)} = i, X_{t-4}^{(\nu)} = j\right] \geq 2(\frac{1}{2})^4 = \frac{1}{8} > 0.$$

Dies ergibt wieder eine exponentielle Konvergenz. Das Vorgehen, mehrere Schritte auf einmal anzusehen, hilft ganz allgemein.

Für komplexere Übergänge bleibt jedoch das Problem, scharfe Abschätzungen für den Bias a priori anzugeben, schwierig. Alternativ kann man versuchen, nach der Simulation aus dem Plot von $h(X_t)$ gegen t abzulesen "wann die Kette konvergiert hat".

4.3.2 Schätzung der Varianz im stationären Fall

Als nächstes betrachten wir den stochastischen Fehler unter der Annahme, dass $(X_1, X_2, \ldots, X_k)$ und $(X_{i+1}, X_{i+2}, \ldots, X_{i+k})$ die gleiche Verteilung haben für alle i und für alle k, d.h (X_i) ist stationär.

Falls (X_i) eine Markovkette ist mit einem Übergangskern, der nicht von der Zeit abhängt, dann haben wir Stationarität genau dann, wenn $X_1 \sim \pi$ (π ist die stationäre Verteilung).

Lemma 4.4. *Sei* (X_i) *stationär, sei* $Y_i = h(X_i)$, *und sei* $R(k) = \mathrm{Cov}\,(Y_t, Y_{t+k})$. *Dann gilt*

a)

$$\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{N}\sum_{k=-N+1}^{N-1}(1 - \frac{|k|}{N})R(k).$$

b) Falls $\sum_{k=1}^{\infty} |R(k)| < \infty$*, dann*

$$N \operatorname{Var}(\hat{\theta}_N) \to \sigma_\infty^2 = \sum_{k=-\infty}^{\infty} R(k) \quad (N \to \infty).$$

c) Falls $\sum_{k=1}^{\infty} |R(k)| < \infty$*, dann*

$$\operatorname{Corr}\left(\frac{1}{N} \sum_{i=1}^{N} Y_i, \frac{1}{N} \sum_{i=N+1}^{2N} Y_i \right) \to 0.$$

Beweis. Aussage a) folgt aus

$$\begin{aligned} \operatorname{Var}\left(\sum_{i=1}^{N} Y_i \right) &= \sum_{i=1}^{N} \sum_{j=1}^{N} \underbrace{\operatorname{Cov}(Y_i, Y_j)}_{=R(i-j)} \\ &= \sum_{k=-N+1}^{N-1} R(k) \cdot \underbrace{(\text{Anzahl Paare mit } i-j=k)}_{(=N-|k|)} \end{aligned}$$

Für b) schreiben wir

$$\begin{aligned} N \operatorname{Var}\left(\hat{\theta}_N\right) &= \sum_{k=-N+1}^{N-1} (1 - \frac{|k|}{N}) R(k) \\ &= \sum_{k=-\infty}^{\infty} \underbrace{\max(0, 1 - \frac{|k|}{N}) R(k)}_{\to R(k) \text{ für } N \to \infty} \end{aligned}$$

Die Behauptung folgt also mit dem Konvergenzsatz von Lebesgue.

Für c) gehen wir aus von

$$\operatorname{Cov}\left(\sum_{i=1}^{N} Y_i, \sum_{i=N+1}^{2N} Y_i \right) = \sum_{k=1}^{2N-1} \min(k, 2N-k) \cdot R(k)$$

Wegen b) muss man nun zeigen, dass der ganze Ausdruck weniger schnell wächst als N. Das ergibt sich aus folgender Abschätzung:

$$\begin{aligned} \Big| \sum_{k=1}^{2N-1} \min(k, 2N-k) \cdot R(k) \Big| &\leq \sqrt{N} \sum_{k=1}^{\sqrt{N}} |R(k)| + N \sum_{k=\sqrt{N}}^{\infty} |R(k)| \\ &\leq \sqrt{N} \sum_{k=1}^{\infty} |R(k)| + N \sum_{k=\sqrt{N}}^{\infty} |R(k)| = o(N) \end{aligned}$$

□

Unter der Annahme $\sum |R(k)| < \infty$ gilt daher (mit Chebyshev)

$$\mathrm{P}\left[|\hat{\theta}_N - \theta| > \epsilon \right] \leq \frac{\operatorname{Var}\left(\hat{\theta}_N\right)}{\epsilon^2} \sim \frac{\sigma_\infty^2}{N\epsilon^2}$$

Man braucht also eine Schätzung von σ_∞. Zudem weiss man, dass die Chebyshev-Ungleichung eine schlechte Abschätzung ist und möchte sie deshalb ersetzen durch eine Normalapproximation.

Damit stellen sich folgende Fragen:

1. Wann ist $\sum |R(k)| < \infty$?
2. Wie schätzt man σ_∞?
3. Gilt ein zentraler Grenzwertsatz?

Zu 1: Sei (X_i) eine stationäre Markovkette mit Übergangskern P. Es gilt dann

$$\begin{aligned} \operatorname{Cov}(h(X_0), h(X_t)) &= \mathbf{E}\left[(h(X_0)-\theta)(h(X_t)-\theta)\right] \\ &= \mathbf{E}\left[(h(X_0)-\theta)\,\mathbf{E}\left[h(X_t)-\theta|X_0\right]\right] \\ &= \int (h(x)-\theta)(P^t h(x)-\theta)\pi(dx). \end{aligned}$$

Also ist entscheidend, wie rasch $P^t h(x) - \theta$ gegen null geht (analog wie beim Bias). Insbesondere ist die folgende Bedingung hinreichend:

$$\sup_x \sum_t |P^t h(x) - \theta| < \infty.$$

Zu 2: Ein natürlicher Schätzer für $R(k)$ ist:

$$\hat{R}(k) \;=\; \frac{1}{N}\sum_{i=1}^{N-|k|} (Y_i - \hat{\theta}_N)(Y_{i+|k|} - \hat{\theta}_N)$$

(eigentlich müsste man durch $N-|k|$ teilen, aber für $|k|$ klein spielt es keine Rolle und für $|k|$ gross, schätzt man $\hat{R}(k)$ ein wenig kleiner).

Aber $\sum_{k=-N+1}^{N-1} \hat{R}(k)$ ist ein unbrauchbarer Schätzer von σ_∞^2: Man kann nämlich nachrechnen, dass

$$\sum_{k=-N+1}^{N-1} \hat{R}(k) \;=\; \left(\sum_{i=1}^{N}(Y_i - \hat{\theta}_N)\right)^2 \;=\; 0.$$

Ein besserer Schätzer ist

$$\hat{\sigma}_\infty^2 = \sum_{k=-m}^{m} w_k \hat{R}(k), \tag{4.10}$$

wobei w_k die in Abbildung 4.1 gezeigte Form hat. Man gewichtet also die Kovarianzen mit wachsendem Abstand herunter. Die Wahl des Punktes m, von dem an man den geschätzten Kovarianzen Gewicht null gibt, ist dabei der heikle Punkt. Von der Theorie her sollte $m \to \infty$ und $m = o(N)$ gelten, d.h. m wächst, aber langsamer als N. Empirisch hat sich $m \approx N^{1/3}$ als eine häufig vernünftige Wahl erwiesen.

Zu 3: Es gibt eine ganze Literatur zum Problem, Bedingungen für die Gültigkeit des Zentralen Grenzwertsatzes bei stationären Zufallsvariablen zu formulieren und zu beweisen. Eines der einfachsten und für Markovketten Monte Carlo wichtiges Resultat ist das folgende: Ist (X_i) eine Markovkette mit P reversibel für π, dann ist $\frac{1}{N}\sum h(X_i)$ asymptotisch normal, falls $\sum_k |R(k)| < \infty$.

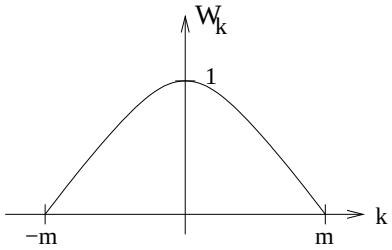

Abbildung 4.1: Form der Gewichte beim Schätzer (4.10) von σ_∞^2

Betrachten wir zum Abschluss noch das Vertrauensintervall von θ: Aufgrund von dem bisher Gesagten ist folgendes Intervall naheliegend:

$$\hat{\theta}_N \ \pm \ \Phi^{-1}(1-\frac{\alpha}{2})\frac{1}{\sqrt{N}}\widehat{\sigma}_\infty$$

Eine andere Möglichkeit ist die sogenannte "batch means" Methode. Dort berechnet man die Mittel von jeweils b aufeinanderfolgenden Y_i's:

$$\hat{\theta}_{i,b} = \frac{1}{b}\sum_{j=(i-1)b+1}^{ib} Y_j$$

Diese Mittelwerte $\hat{\theta}_{i,b}$, $i = 1, 2, \ldots, k = N/b$ betrachtet man als unabhängig und normalverteilt, vgl. Lemma 4.4c). Das übliche t-Vertrauensintervall lautet dann

$$\hat{\theta}_N \ \pm \ \frac{1}{\sqrt{k}} t_{k-1,1-\frac{\alpha}{2}} \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\hat{\theta}_{i,b}-\hat{\theta}_N)^2}$$

Der Vorteil ist, dass man so σ_∞ nicht schätzen muss. Die Wahl von b ist jedoch gleich schwierig wie die Wahl von m in (4.10).

4.3.3 Kopplung aus der Vergangenheit

Dies ist eine neuere Idee, wie man das Problem des Bias bei Markovketten Monte Carlo vermeiden kann. Es wäre schön, wenn man in endlich vielen Schritten die stationäre Verteilung realisieren könnte. Dazu nehmen wir das Konzept der Kopplung wieder auf, das wir bereits beim Beweis der Konvergenz gegen die stationäre Verteilung eingeführt haben. Wenn sich alle Pfade mit allen möglichen Startwerten gekoppelt haben, dann kennt man insbesondere auch den Zustand der stationären Markovkette, die mit π startet. Allerdings kann man daraus nicht schliessen, dass nach der Kopplung die Verteilung des gemeinsamen Zustandes gleich π ist. Der Zeitpunkt der Kopplung ist zufällig, und zu einem zufälligen Zeitpunkt ist die Verteilung auch bei der stationären Kette verschieden von π.

Am deutlichsten sieht man das bei der Irrfahrt mit Spiegelung am Rand. Die Ketten treffen sich fast sicher, aber zum Zeitpunkt T der Kopplung sind alle $X_T^{(i)} \in \{1, 5\}$, d.h.

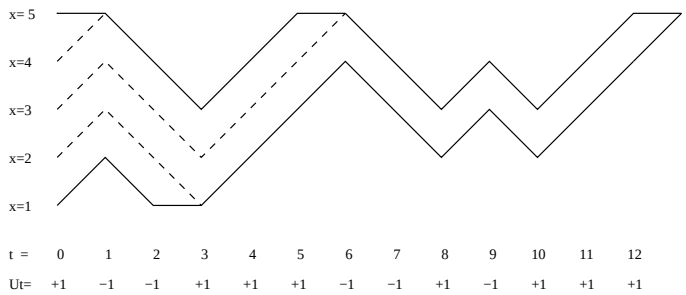

Abbildung 4.2: Kopplung vorwärts

die Verteilung ist sicher nicht gleich π. Man kann also Kopplung vorwärts nicht benutzen, um eine Zufallsvariable zu erzeugen, die exakt die Verteilung π hat.

Der Ausweg besteht darin, die feste Zeit $t = 0$ zu betrachten, und die Kopplung rückwärts (aus der Vergangenheit) einzuführen.

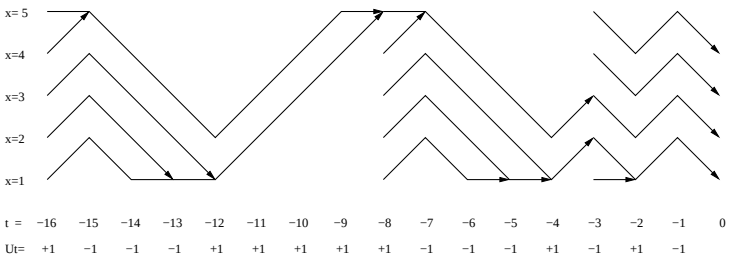

Abbildung 4.3: Kopplung rückwärts

Der Algorithmus geht wie folgt: Wir gehen zuerst rückwärts zu -2 und betrachten die Ketten mit den Startwerten 1 bis 5. Dies bringt aber im Beispiel noch nichts, da sich die Ketten zur Zeit $t = 0$ nicht koppeln. Als nächstes machen wir das Gleiche von -4, -8 etc. aus, bis wir einen Zeitpunkt finden, von dem aus wir Kopplung zur Zeit $t = 0$ erhalten. Im Beispiel ist das der Fall bei $t = -16$: Dann koppeln sich alle Ketten zum Zeitpunkt -8 und enden in 2. Die Zufallsvariable, die wir auf diese Weise generieren, hat exakt die Verteilung π. Es ist dabei wesentlich, dass man beim weiteren Zurückgehen in der vorderen Hälfte die gleichen Zufallszahlen verwendet.