# MCMC

Ausarbeitung zum Seminar: The Top 10 Algorithms

Claudia Spellecchia
Dübendorf, Schweiz
claudisp@student.ethz.ch
Matrikelnummer 04-722-583

Abgabe der Arbeit: 13.03.2008

# Inhaltsverzeichnis

# 1 Die Monte Carlo Methode

Die Monte Carlo Methode ist eine numerische Methode zur Lösung mathematischer Probleme mit Hilfe der Modellierung von Zufallsgrössen.

## 1.1 Berechnung der Zahl $\pi$

Dieses Beispiel zeigt, wie die Zahl $\pi$ mit Hilfe der Monte Carlo Methode ermittelt werden kann. Betrachte dazu das Einheitsquadrat $I = [0,1] \times [0,1]$ und das Teilgebiet $D = \{(\xi, \eta) \,|\xi^2 + \eta^2 \leq 1, \xi \geq 0, \eta \geq 0\}$. Das Verhältnis des Flächeninhaltes des Viertelkreises und dem Eineitsquadrat beträgt $\pi/4$. Die Idee ist nun zufällige Zahlen $(x, y)$ zu generieren, wobei $x \sim$ Uniform$[0,1]$ und $y \sim$ Uniform$[0,1]$. Somit ist die Wahrscheinlichkeit, dass ein Punkt in $[a,b] \times [c,d]$, $0 \leq a < b \leq 1$, $0 \leq c < d \leq 1$ liegt gleich $(b-a)\,(d-c)$. Generiert man mit Hilfe eines Generators (z.B. Linearer Kongruenzgenerator) mehrere zufällige Punkte im Einheitsquadrat und bestimmt man die relative Häufigkeit der Punkte, die im Gebiet D liegen, so erhält man einen Schätzwert für $\pi/4$. Die Zahl $\pi/4$ kann also als Erwartungswert einer Zufallsvariablen aufgefasst werden und somit durch den empirischen Erwartungswert (d.h. Aritmethischem Mittel) geschätzt werden.

Generieren wir $x \sim$ Uniform$[0,1]$ und $y \sim$ Uniform$[0,1]$, so können wir kontrollieren, ob $x^2 + y^2 \leq 1$ ist, und somit in D liegt. Sei $n'$ die Anzahl solcher in D liegenden Punkte, und sei $n$ die Anzahl aller generierten Punkte. Dann ist $n'/n$ eine Näherung für $\pi/4$.

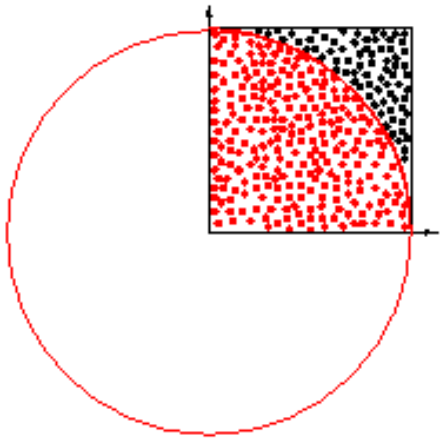

Abbildung 1: Die roten Punkte liegen in D, die schwarzen ausserhalb. Quelle: http://de.wikipedia.org/wiki/Monte-Carlo-Simulation

## 1.2 Geschichte der Monte Carlo Methode

Als Geburtsjahr der Monte Carlo Methode gilt das Jahr 1949, in dem eine Arbeit mit dem Titel 'The Monte Carlo Method' erschien. Als Gründer dieser Methode sind die Mathematiker J. v. Neumann, N. Metropolis und S. Ulam anzusehen.

Die theoretischen Grundlagen der Monte Carlo Methode waren schon seit langem bekannt, aber solange keine elektronischen Datenverarbeitungsmethoden zu Verfügung standen, konnte diese Methode keine breite Anwendung finden, da die Simulation von Zufallsgrössen von Hand viel zu aufwendig ist.

Die Bezeichnung 'Monte Carlo Methode' geht auf die Stadt Monte Carlo im Fürstentum Monaco zurück, die besonders durch ihr Spielcasino bekannt geworden ist. Eines der einfachsten mechanischen Geräten zur Realisierung von Zufallsgrössen ist nämlich das Roulette.

## 1.3 Monte Carlo Integration

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

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

kann als Erwartungswert von i.i.d. (independent, identically distributed) gleichverteilten Zufallsgrössen auf $[0,1]$ aufgefasst werden, d.h. $E[f(U_1, ..., U_p)]$ mit, $U_1, ..., U_p$ i.i.d.$\sim$ Uniform$[0,1]$. Das Integral kann man approximieren, indem man $N \times p$ Werte $U_{i,1}, ..., U_{i,p}$, $i = 1..N$ erzeugt, und dann das arithmetische Mittel berechnet,

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

Aus dem zentralen Grenzwertsatz folgt, falls das zweite Moment existiert, dass die Konvergenzgeschwindigkeit $1/\sqrt{N}$ ist.

# 2 Markovketten

## 2.1 Grundbegriffe

Im allgemeinen Fall haben wir folgende Definition:

**Definition 1** *Sei $\mathbb{X}$ beliebiger Raum mit $\sigma$-Algebra $\mathcal{F}$. Ein **Kern** $P$ auf $(\mathbb{X}, \mathcal{F})$ ist eine Abbildung von $\mathbb{X} \times \mathcal{F} \longrightarrow [0,1]$ so dass,*

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

$P(x, A)$ gibt also die Wahrscheinlichkeit an, mit der man von einem Zustand $x$ zu einem beliebigen Ereignis $A$ gelangt. Jedem Wahrscheinlichkeitsmass $\nu$ wird durch den Kern ein Wahrscheinlichkeitsmass zugeordnet, durch

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

Im Fall, wo $\mathbb{X}$ endlich ist, ist der Kern einfach eine Matrix $P = (P(i,j))_{i,j \in \mathbb{X}}$ mit nichtnegativen, zwischen 0 und 1 liegenden Elementen, so dass die Zeilensumme gleich 1 ist. $P(i,j)$ gibt die Wahrscheinlichkeit an, von einem Zustand $i$ zu einem Zustand $j$ zu gelangen. $\nu P$ entspricht dann der Multiplikation von einem Zeilenvektor von Links.

**Definition 2** *Eine **Markovkette** auf $(\mathbb{X}, \mathcal{F})$ mit Startverteilung $\nu_0$ und Übergangskern $P$ ist eine Folge von Zufallsvariablen $(X_0, X_1, X_2, ...)$ mit Werten in $\mathbb{X}$ derart, dass*

- $P[X_0 \in A] = \nu_0(A)$
- $P[X_{t+1} \in A | X_t = x_t, ..., X_0 = x_0] = P[X_{t+1} \in A | X_t = x_t]$

Eine Markovkette beschreibt also eine diskrete Zeitentwicklung auf $\mathbb{X}$ mit kleiner Abhängigkeit, denn der nächste Zustand hängt nur vom jetzigen Zustand ab, jedoch nicht von der Vergangenheit. Der Kern beschreibt also die bedingte Verteilung des nächsten Zustandes gegeben der jetzige Zustand. Einfache Folgerungen sind die folgenden:

$$P[X_{t+k} \in A | X_t = x_t, ..., X_0 = x_0] = P^k(x_t, A)$$

und

$$P[X_t \in A] = \nu_0 P^t(A).$$

**Definition 3** *Eine Wahrscheinlichkeitsverteilung $\Pi$ auf $(\mathbb{X}, \mathcal{F})$ heisst **invariant** oder **stationär** für einen Übergangskern $P$, falls gilt*

$$\Pi P = \Pi.$$

**Definition 4** *Eine Wahrscheinlichkeitsverteilung $\Pi$ auf $(\mathbb{X}, \mathcal{F})$ heisst **reversibel** für einen Übergangskern $P$, falls gilt*

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

**Definition 5** *Ein Übergangskern $P$ heisst **irreduzibel**, wenn eine Wahrscheinlichkeitsverteilung $\Psi$ auf $(\mathbb{X}, \mathcal{F})$ existiert so dass $\sum_{k=1}^{\infty} P^k(x, A) > 0$ für alle $A \in F$ mit $\Psi(A) > 0$, für alle $x \in \mathbb{X}$.*

Nun zu den Bedeutungen dieser Definitionen. Falls wir als Startverteilung eine invariante Verteilung $\Pi$ wählen, dann haben alle weiteren $X_t$ die Verteilung $\Pi$.

Bei Startverteilung $\Pi$ bedeutet Reversibilität, dass $(X_0, X_1)$ und $(X_1, X_0)$ dieselbe Verteilung haben. Es gilt auch, dass Reversibilität die Invarianz impliziert.

Ist ein Übergangskern irreduzibel, so kann man bei beliebiger Startverteilung mit positiver Wahrscheinlichkeit jeden anderen Zustand erreichen.

**Satz 6** *Sei $P$ irreduzibler Kern mit invarianter Verteilung $\Pi$. Dann ist $\Pi$ die einzige invariante Verteilung und es gilt:*

$$P\left[\frac{1}{n+1}\sum_{t=0}^{n} h(x_t) \to \int h(x)\Pi(dx) | X_0 = x\right] = 1$$

*für $\Pi$-fast alle $x \in \mathbb{X}$, $\forall h$ mit $\int |h(x)|\Pi(dx) < \infty$.*

Zu bemerken ist, dass $\int h(x)\Pi(dx)$ einfach der ausgeschriebene Erwartungswert $E_\Pi[h(X)]$ ist. Der Satz besagt also, dass uns die Irreduzibilität eine eindeutige invariante Verteilung liefert, und dass das Gesetz der Grossen Zahlen gilt.

## 2.2 Beispiele

1. Eine Markovkette kann man als gewichteten, orientierten Graphen darstellen. Die Knoten stellen die Zustände der Markovkette dar, und ein Pfeil mit dem Gewicht $P(i, j)$ verbindet den Zustand $i$ zum Zustand $j$, falls $P(i, j) > 0$. Für das Beispiel, das in Abbildung 2 abgedildet ist, lautet die stochastische Matrix
$$P_1 = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}.$$
Ausserdem sieht man dem Graphen an, dass jedes Element erreicht werden kann, d.h. $P_1$ ist irreduzibel.

2. Für das Beispiel, das in Abbildung 3 abgedildet ist, lautet die stochastische Matrix
$$P_2 = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1/4 & 3/4 \end{pmatrix}.$$

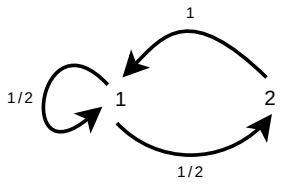


Abbildung 2: Der Graph von der Markovkette mit $P_1$

Ausserdem sieht man dem Graphen an, dass nicht jedes Element erreicht werden kann, beispielsweise kann man vom Zustand 1 oder 2 die Zustände 3 und 4 nicht erreichen, d.h. $P_2$ ist reduzibel.

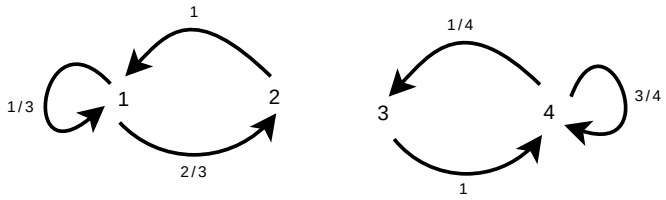


Abbildung 3: Der Graph von der Markovkette mit $P_2$

# 3 Markovketten Monte Carlo (MCMC)

In diesem Kapitel besprechen wir die Standardmethode zur Simulation von Verteilungen in hohen Dimensionen. Die uns interessierende Verteilung bezeichnen wir mit $\Pi$. Die Grundidee besteht darin, eine rekursive Folge $X_t$ so zu erzeugen, dass $X_t$ für grosse $t$ die Verteilung $\Pi$ hat. Wir wollen also eine Markovkette bilden, wobei die Übergänge so gewählt werden, dass, falls $X_r$ die gewünschte Verteilung $\Pi$ hat, auch alle weiteren Variablen $X_{r+1}, X_{r+2}, ..$ die Verteilung $\Pi$ haben, d.h. $\Pi$ ist eine invariante Verteilung. Anschliessend wenden wir das Prinzip, das wir im Abschnitt 1.3 kennen gelernt haben an, und erhalten folgende Approximation

$$\int h(x)\Pi(dx) = E_\pi[h(X)] \approx \frac{1}{N-r+1}\sum_{t=r}^{N} h(X_t), \quad \text{wobei } X_t \sim \Pi.$$

Damit diese Approximation sinnvoll ist, müssen wir einen Übergangskern wählen, der folgende Bedingungen erfüllt:

- $P$ ist irreduzibel,

- $\Pi$ ist invariant für $P$ (reversibel genügt),
- Die Simulation gemäss $P(x, \cdot)$ soll einfach sein für alle $x$.

Satz 6 liefert uns dann die Gültigkeit des Gesetz der Grossen Zahlen.

## 3.1 Metropolis-Hastings Algorithmus

Der Metropolis-Hastings Algorithmus liefert uns ein Rezept, um solche Kerne zu konstruieren. Wir wollen einen Kern finden, so dass $\Pi$ reversibel ist für $P$ (daraus folgt, dass $\Pi$ auch invariant ist für $P$). Im diskreten Fall müssen wir also das folgende Problem lösen:

$$\Pi(i)P(i,j) = \Pi(j)P(j,i) \quad \forall i, j.$$

Zu beachten ist, dass wir für jedes $i < j$ entweder die Wahrscheinlichkeit $P(i,j)$ oder $P(j,i)$ wählen müssen, denn die andere ergibt sich aus der Reversibilitätsgleichung. Weiter ist zu beachten, dass die Zeilensumme des Übergangskerns gleich 1 sein muss, d.h. $\sum_j P(i,j) = 1$. Falls die Zeilensumme über $i \neq j$ kleiner ist als 1, können wir $P(i,i)$ so wählen, dass die Summe 1 ergibt.
Folgende Konstruktion liefert uns einen Übergangskern $P$, für den unsere Zielverteilung $\Pi$ invariant ist.

1. Nehme eine beliebige Übergangsmatrix $Q$ s.d.

   $$Q(i,j) > 0 \Leftrightarrow Q(j,i) > 0$$

   d.h. falls ich nicht vorschlage von $i$ nach $j$ zu gehen, schlage ich auch nicht vor von $j$ nach $i$ zu gehen.

2. Für jedes $i < j$ setze entweder

   $$\text{a)} \;\; P(i,j) = Q(i,j) \;\text{ und }\; P(j,i) = \frac{\Pi(i)}{\Pi(j)} Q(i,j)$$

   oder

   $$\text{b)} \;\; P(j,i) = Q(j,i) \;\text{ und }\; P(i,j) = \frac{\Pi(j)}{\Pi(i)} Q(j,i)$$

3. Wir wollen, dass $P(i,j) \leq Q(i,j)$ und $P(j,i) \leq Q(j,i)$, damit die Zeilensumme

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

   Dann können wir $P(i,i) = 1 - \sum_{j \neq i} P(i,j)$ wählen.

4. Aus 2. folgt somit, falls wir a) nehmen, dass $P(j,i) = \frac{\Pi(i)}{\Pi(j)}Q(i,j) \leq Q(j,i)$ ist, genau dann wenn $\Pi(i)Q(i,j) \leq \Pi(j)Q(i,j)$. Ist dies nicht erfüllt, so führt 2b) zum Ziel.

Fassen wir nun zusammen, ergibt sich für $P$ somit:

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

wobei

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

$Q$ heisst Vorschlagsdichte und $a$ heisst Akzeptierungswahrscheinlichkeit.
Ich nehme also an, ich kann von $Q_{i\cdot}$ simulieren, wie kann ich denn mit $P_{i\cdot}$ simulieren?

**Algorithmus 7**

1. Generiere $U \sim$Uniform$(0,1)$ und gegeben $X_t = i$, schlage $Y$ mit $P[Y = j] = Q(i,j)$ vor, d.h. $Y \sim Q(i,\cdot)$.
2. Falls $U \leq a(i,Y)$, setzte $X_{t+1} = Y$, sonst $X_{t+1} = X_t = i$.

Wir haben also nach dem Durchlauf des Algorithmus

$$P[Y = j | X_t = i] = Q(i,j)a(i,j) \text{ für } j \neq i$$

und

$$P[Y = i | X_t = i] = 1 - \sum_{j, j\neq i} Q(i,j)a(i,j)$$

### 3.1.1 Das Ising Modell

Das Ising Modell realisiert eine sehr einfache Vorstellung von einem Ferromagneten. Wir betrachten hier das Ising Modell in 2D. Atome sitzen in den Ecken von einem 2-dimensionalen Gitter. In jeder Ecke $i$ hat ein Atom ein positiver Spin (d.h. $\sigma_i = 1$) oder negativer Spin (d.h. $\sigma_i = -1$). Jedes Paar von Orte $i$ und $j$ hat eine Interaktionsenergie $J_{i,j}\sigma_i\sigma_j$, welche nicht nur von den jeweiligen Spins abhängt, sondern auch von dem Abstand zwischen denen.
Dann hat das Ising Modell folgende Energiefunktion

$$E(\sigma) = -\sum_{i,j} J_{i,j}\sigma_i\sigma_j - B\sum_k \sigma_k,$$

wobei $\sigma$ eine Konfiguration ist, d.h. $\sigma$ ist eine beliebige Wahl von den Spins (d.h. $\sigma \in \{-1, 1\}^m$, wobei wir $m$ Atome haben), und $B$ bezeichnet eine Konstante, die vom äusseren Magnetfeld abhängt.

Es ist zu bemerken, dass die erste Summe über die Paare $\{i, j\}$ geht, die miteinander interagieren, üblich sind die direkten Nachbarn.

In vielen Anwendungen interessieren wir uns für eine Funktion $h(\sigma)$, deren geschätzter Erwartungswert uns schon fundamentale Informationen liefert. Im Ising Modell wird der Erwartungswert $\theta$ über alle Konfigurationen gebildet:

$$\theta = \frac{1}{Z(T)} \sum_{\sigma} h(\sigma) \exp(-E(\sigma)/\kappa T).$$

Der Normierungsfaktor ist die Zustandssumme

$$Z(T) = \sum_{\sigma} \exp(-E(\sigma)/\kappa T),$$

wobei $T = Temperatur$ und $\kappa = BoltzmanKonstante$. Wir möchten also einen Schätzer für $\theta$ finde.

Eine Option wäre, alle Konfigurationen von der Verteilung $\Pi$, wobei

$$\Pi(\sigma) = \frac{\exp(-E(\sigma)/\kappa T)}{Z(T)}$$

zu wählen, so dass das arithmetische Mittel von $M$ Proben

$$\hat{\theta} = \frac{1}{M} \sum_{k} h(\sigma_k) \to \theta.$$

Das Problem besteht aber darin, dass es eben nicht so einfach ist, direkt von $\Pi$ Zufallsgrössen zu erzeugen.

Eine weitere Option wäre eine aperiodische, symmetrische Markovkette zu konstruieren, so dass die Konvergenzverteilung $\Pi$ ist. Man kann eine solche Markovkette konstruieren, wenn man beachtet, dass durch die Änderung nur eines Spins, der Energie nur eine einfach zu berechnende Änderung $\Delta E$ zugeführt wird (da sich nur einige Terme in der Summe ändern).

Zum Schluss geben wir noch einige Beschreibungen, was mit dem Ferromagneten passiert, wenn sich die Temperatur verändert. Bei kleinen Temperaturen liegt der Ferromagnet in geordneter Phase vor (alle Spins $\sigma_i$ sind entweder 1 oder -1) , und es liegt eine Magnetisierung vor. Wird eine kritische Temperatur überschritten, so findet ein Phasenübergang statt, nämlich in eine ungeordnete Phase, in der die Spins willkürlich orientiert sind, vergleiche dazu Abbildung 4.

## 3.2 Schlussbemerkungen

Obwohl MCMC beeindruckende Fortschritte gemacht hat, ist etwas im Hinterkopf zu behalten: Monte Carlo ist ein Ausweg, den man nur wählen sollte, falls keine exakte analytische Methode oder kein endlicher numerischer Algorithmus zur Verfügung steht. Man sollte auch nur für eine gewisse Zeit simulieren, nämlich bis dahin, wo man denkt, man hat die Zielverteilung erreicht, denn es gibt auch Probleme, für die bewiesen wurde, dass die Konvergenzgeschwindigkeit extrem langsam ist. Ein Beispiel ist das Ising Modell.

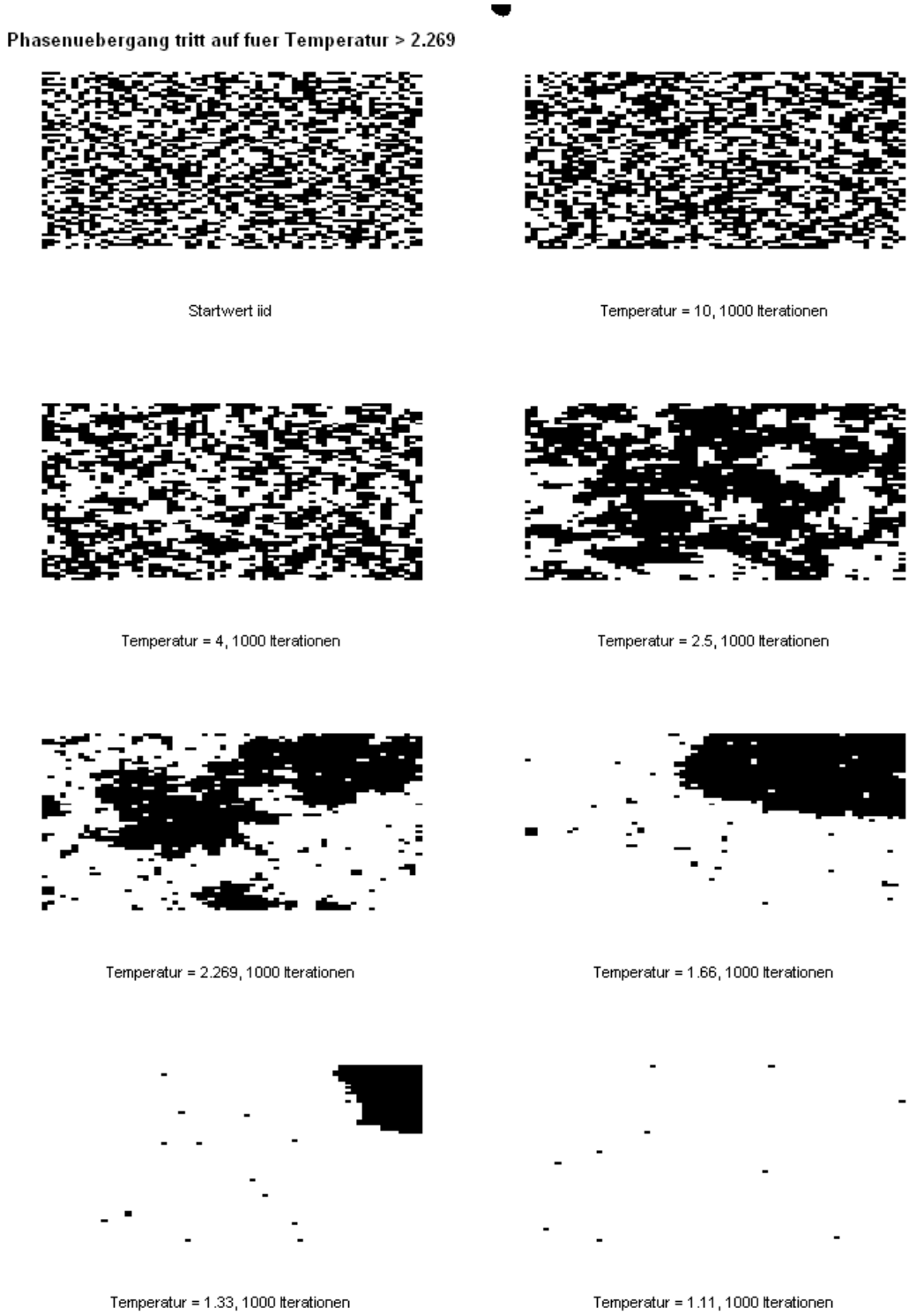


Abbildung 4: Spinrichtung bei verschiedenen Temperaturen, kritische Temperatur ist 2.269. Bilder wurden produziert mit R-Code von http://www.stat.math.ethz.ch/∼kuensch/

# Literatur

[BeSu00] I. Beichl, F. Sullivan, *the Top 10 Algorithms: The Metropolis Algorithm.* IEEE Jan/Feb 2000

[Fish96] G.S. Fishman, *Monte Carlo, Concepts, Algorithms, and Applications.* Springer, 1996

[HeTh78] W. Hengartner, R. Theodorescu, *Einführung in die Monte-Carlo-Methode.* Carl Hanser Verlag München Wien, 1978

[Kuen03] H. Künsch, *Stochastische Simulation.* Skript zur Vorlesung im WS03/04, Oktober 2003

[Sobo91] I.M. Sobol, *Die Monte-Carlo-Methode.* Deutscher Verlag der Wissenschaften, Berlin 1991