Dietmar Holle

Konvergenzverhalten des Gibbs Samplers

Professor Dr. Gerold Alsmeyer

dem Institut für
Mathematische Statistik
am Fachbereich Mathematik und Informatik
der Westfälischen Wilhelms-Universität Münster
als Diplomarbeit eingereicht
im Mai 2010

Inhaltsverzeichnis

Einleitung

In praktischen Anwendungen stoßen Mathematiker regelmäßig auf höherdimensionale Integrale der Form

$$\frac{1}{K} \int f(x)\pi(x)\, dx \quad \text{mit } K = \int \pi(x)\, dx,$$

hierbei bezeichnet $f(x)$ eine beliebige Funktion und $\pi(x)/K$ kann als Dichte einer höherdimensionalen Verteilung angesehen werden. Die exakte Berechnung solcher Integrale ist in den meisten Fällen überhaupt nicht möglich. In dieser Situation können Markov Chain Monte Carlo Verfahren hilfreich sein: Gelingt es die Verteilung π zu simulieren, kann man eine stochastische Folge $X^{(1)}, \ldots, X^{(n)}$ erzeugen, mittels derer der Erwartungswert der Funktion $f(X)$ und damit das gewünschte Integral approximativ berechnet werden kann, das heißt:

$$E(f(X)) = \frac{1}{K} \int f(x)\pi(x)\, dx \approx \frac{1}{n} \sum_{i=1}^{n} X^{(i)}.$$

Die Gesetze der großen Zahlen sorgen im Falle einer unabhängigen Stichprobe $X^{(1)}, \ldots, X^{(n)}$ für eine beliebige Genauigkeit der Approximation, sofern n nur groß genug gewählt wird. Im Allgemeinen wird die Erzeugung unabhängiger Zufallsvariablen, die der Verteilung π genügen, nicht möglich sein. Finden wir aber ein Verfahren zur Erzeugung einer Markov-Kette, deren stationäre Verteilung gerade die gewünschte Zielverteilung π ist, so führt unsere Approximation weiterhin zum richtigen Ergebnis. Der Gibbs Sampler ist ein solches Verfahren. Populär geworden ist er durch einen Beitrag der Brüder Geman zur Restauration von Bildern aus dem Jahr 1984 (siehe [Ge/Ge]). Die Ursprünge der Methode liegen aber bereits im Jahr 1953: Der Metropolis-Hastings Algorithmus bildet das theoretische Fundament des Gibbs Samplers (siehe [Me]). Die Idee, das Ausgangsproblem durch eine Vielzahl einfacher Rechenschritte zu lösen, rechtfertigt sich im Zuge des technischen Fortschritts, welcher durch eine stetig zunehmende Rechenkapazität gekennzeichnet ist.

Dass eine Markov-Kette mit beliebiger Genauigkeit gegen ihre stationäre Verteilung strebt, besagt der Ergodensatz. Welche Aussagen lassen sich aber bezüglich der Konvergenzgeschwindigkeit treffen? Nachdem wir uns in den ersten beiden Kapiteln mit grundlegenden Fakten Harris-rekurrenter Markov-Ketten und der formalen Definition des Gibbs Samplers beschäftigen, werden wir im dritten Kapitel die geometrische Konvergenz desselben für eine bestimmte Klasse von Zielverteilungen, denen insbesondere die Normalverteilung angehört, zeigen. Im abschließenden vierten Kapitel werden wir uns nicht mehr mit dem Nachweis geometrischer Konvergenz begnügen: Hier werden

wir die Konvergenzrate als Spektralradius eines linearen Operators identifizieren, wobei dieses Vorhaben an restriktive Bedingungen geknüpft ist.

Die Ergebnisse des dritten Kapitels entstammen einer Veröffentlichung von Hwang und Sheu aus dem Jahr 1998 (siehe [Hw/Sh]). Das vierte Kapitel basiert auf einem Artikel von Li und Geng aus dem Jahr 2005 (siehe [Li/Ge]).

1 Harris-rekurrente Markov-Ketten

1.1 Definition einer Markov-Kette und Stationarität

Der Gibbs-Sampler ist ein Verfahren zur Konstruktion einer zeitlich homogenen Markov-Kette, deren stationäre Verteilung eine gewünschte Zielverteilung sein soll. Wenden wir uns daher zunächst einigen grundlegenden Begriffen und Eigenschaften Harris-rekurrenter Markov-Ketten zu. Als Zustandsraum beschränken wir uns auf den d-dimensionalen euklidischen Raum $(\mathbb{R}^d, \mathfrak{B}^d)$. Ziel dieses Kapitels ist es, wichtige Definitionen und wesentliche Resultate aus der Theorie solcher Markov-Ketten zusammenzustellen, da sie als grundlegende Voraussetzung für das Verständnis der Konvergenzresultate in den folgenden Kapiteln unabdingbar sind. Die Darstellung orientiert sich vor allem an [Num], siehe aber auch [Als2] und [Me/Tw]. Auf Beweise wird weitgehend verzichtet.

Definition 1.1 *Eine stochastische Folge $X = (X^{(n)})_{n\geq 0}$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, \mathbb{P})$ mit Zustandsraum $(\mathbb{R}^d, \mathfrak{B}^d)$ heißt Markov-Kette, wenn*

$$\mathbb{P}^{X^{(n+1)}|X^{(0)},\ldots,X^{(n)}} = \mathbb{P}^{X^{(n+1)}|X^{(n)}} \quad \mathbb{P}\text{-f.s.} \tag{1.1}$$

für alle $n \geq 0$.

Die bedingten Verteilungen $\mathbb{P}^{X^{(n+1)}|X^{(n)}}$ lassen sich durch stochastische Kerne $P_{n+1}(X^{(n)}, \cdot)$ darstellen. Ihre Existenz haben wir sichergestellt, da der Zustandsraum $(\mathbb{R}^d, \mathfrak{B}^d)$ polnisch ist (vgl. Satz 53.4 in [Als1]).

Definition 1.2 *Ein (Übergangs-)Kern P auf $(\mathbb{R}^d, \mathfrak{B}^d)$ ist eine Funktion $P : \mathbb{R}^d \times \mathfrak{B}^d \to [0,1]$ mit folgenden Eigenschaften:*

$$P(x, \cdot) : \mathfrak{B}^d \to [0,1] \quad \textit{ist ein W-Maß auf } (\mathbb{R}^d, \mathfrak{B}^d) \textit{ für alle } x \in \mathbb{R}^d \tag{1.2a}$$

und

$$P(\cdot, B) : \mathbb{R}^d \to [0,1] \quad \textit{ist } \mathfrak{B}^d\textit{-meßbar für alle } B \in \mathfrak{B}^d. \tag{1.2b}$$

Hängen diese Kerne nicht vom Zeitpunkt n ab, so spricht man auch von einer zeitlich homogenen Markov-Kette, welche sich alsdann vollständig durch ihre Anfangsverteilung P_0 und den Übergangskern P beschreiben lässt (siehe Gleichung (1.7) in [Als2]):

$$\mathbb{P}^X = P_0 \otimes \left(\bigotimes_{n=1}^{\infty} P \right) = P_0 \otimes P^{\infty}. \tag{1.3}$$

Definieren wir nun das Produkt zweier Kerne P_1 und P_2 als

$$P_1P_2(x,B) = \int P_2(y,B)\ P_1(x,dy), \tag{1.4}$$

so können wir in dieser Weise sukzessive fortfahren, um den sogenannten n-Schritt-Übergangskern P^n zu erhalten:

$$P^n(x,B) = PP^{n-1}(x,B), \tag{1.5}$$

welcher als Wahrscheinlichkeit für einen Übergang von x nach B in n-Schritten interpretiert werden darf. $P^0(x,B)$ bezeichnet konventionsgemäß das Dirac-Maß $\delta_x(B)$ für alle $x \in \mathbb{R}^d$ und $B \in \mathfrak{B}^d$. Offenbar gilt allgemeiner für eine beliebige Startverteilung λ aus der Menge der Verteilungen $\mathfrak{W}(\mathbb{R}^d)$ auf $(\mathbb{R}^d, \mathfrak{B}^d)$:

$$\mathbb{P}_\lambda^{X^{(1)}}(B) = \lambda P(B) = \int P(x,B)\ \lambda(dx) \tag{1.6}$$

für alle $B \in \mathfrak{B}^d$.

Es ist klar, dass die Zufallsgröße $X^{(n)}$ im Allgemeinen nicht gegen einen bestimmten Punkt konvergieren wird. Welche Aussagen lassen sich aber bezüglich ihrer Verteilung $\mathbb{P}_\lambda^{X^{(n)}}$ für $n \to \infty$ treffen?

Definition 1.3 *Ein nicht-triviales σ-endliches Maß π auf dem Zustandsraum $(\mathbb{R}^d, \mathfrak{B}^d)$ heißt stationäres oder invariantes Maß der Markov-Kette X, wenn*

$$\pi(B) = \int P(x,B)\ \pi(dx) \tag{1.7}$$

für alle $B \in \mathfrak{B}^d$. Gilt $\pi(\mathbb{R}^d) = 1$, so sprechen wir auch von einer stationären bzw. invarianten Verteilung von X.

Folgen wir einer Überlegung, welche auf Seite 22 in [Als2] ausgeführt wird und den Zusammenhang zwischen stationären Verteilungen und dem asymptotischen Verhalten einer Markov-Kette herstellt: Es existiere ein $\nu \in \mathfrak{W}(\mathbb{R}^d)$, sodass

$$\lim_{n\to\infty} \mathbb{P}_\lambda^{X^{(n)}}(B) = \nu(B) \tag{1.8}$$

für ein $\lambda \in \mathfrak{W}(\mathbb{R}^d)$ und alle $B \in \mathfrak{B}^d$. Mit dem Funktions-Erweiterungsargument erhalten wir für alle beschränkten, $\mathfrak{B}^d$-meßbaren, reellen Funktionen f:

$$\lim_{n\to\infty} \mathbb{E}_\lambda f(X^{(n)}) = \int f(x)\ \nu(dx).$$

Unter Verwendung der Markov-Eigenschaft lässt sich der Verteilungslimes ν bereits als stationäre Verteilung identifizieren:

$$\begin{aligned}\nu(B) &= \lim_{n\to\infty} \mathbb{P}_\lambda^{X^{(n+1)}}(B) = \lim_{n\to\infty} \int P(x,B)\ \mathbb{P}_\lambda^{X^{(n)}}(dx) \\ &= \lim_{n\to\infty} \mathbb{E}_\lambda P(X^{(n)},B) = \int P(x,B)\ \nu(dx).\end{aligned}$$

Ist der Verteilungslimes aus Gleichung (1.8) unabhängig von der Wahl der Startverteilung λ, so ist er bereits eindeutig festgelegt, denn für eine beliebige stationäre Verteilung μ und alle $B \in \mathfrak{B}^d$ gilt dann:

$$\mu(B) = P_\mu^{X^{(n)}}(B) \to \nu(B) \quad \text{falls } n \to \infty.$$

Im Folgenden wird es darum gehen, Bedingungen zu finden, die zum einen die Existenz und Eindeutigkeit der stationären Verteilung einer Markov-Kette sicherstellen und zum anderen für eine besonders schnelle Konvergenz gegen eben diesen Verteilungslimes sorgen.

1.2 Irreduzibilität und kleine Mengen

Nehmen wir an ϕ sei ein σ-endliches Maß auf $(\mathbb{R}^d, \mathfrak{B}^d)$ und $B \in \mathfrak{B}^d$ eine beliebige ϕ-positive Menge, das heißt eine Menge mit $\phi(B) > 0$.

Definition 1.4 *Ein Markov Kern P auf $(\mathbb{R}^d, \mathfrak{B}^d)$ heißt ϕ-irreduzibel, wenn für alle $x \in \mathbb{R}^d$ und alle ϕ-positiven Mengen $B \in \mathfrak{B}^d$ ein $n \in \mathbb{N}$ existiert, sodass*

$$P^n(x, B) > 0. \tag{1.9}$$

P heißt irreduzibel, so es ϕ-irreduzibel für irgendein ϕ ist. In diesem Fall nennen wir ϕ ein Irreduzibilitätsmaß von P. Ein Irreduzibilitätsmaß φ von P heißt schließlich maximal, wenn alle anderen Irreduzibilitätsmaße von P absolut-stetig bezüglich φ sind.

Diese Definition ist ein wenig allgemeiner als die herkömmliche Definition für diskrete Markov-Ketten, man denke zum Beispiel an eine Irrfahrt auf dem abzählbaren Zustandsraum $\{0, 1, \ldots\}$ mit absorbierendem Zustand 0. Nach Definition 1.4 liegt ϕ-Irreduzibilität vor mit $\phi = \delta_0$, das heißt, wir wählen gerade das Dirac-Maß in 0 als Irreduzibilitätsmaß. Ist P irreduzibel, so lässt sich bereits ein maximales Irreduzibilitätsmaß φ konstruieren (vgl. Proposition 2.4. in [Num]). Alle anderen maximalen Irreduzibilitätsmaße sind hierzu equivalent - in diesem Sinne liegt dann sogar Eindeutigkeit vor. Irreduzibilität lässt sich häufig leicht verifizieren. Hierzu geben wir ein hinreichendes Kriterium.

Satz 1.5 *P ist irreduzibel bezüglich ϕ, falls P^n eine positive Dichte bezüglich ϕ besitzt. Das heißt, es existieren ein $n \geq 1$ und ein $f : \mathbb{R}^d \times \mathbb{R}^d \to (0, \infty]$, sodass*

$$P^n(x, B) = \int_B f(x, y)\ \phi(dy) \tag{1.10}$$

für alle $x \in \mathbb{R}^d$ und $B \in \mathfrak{B}^d$.

Eine wichtige Klasse von Funktionen stellen die sogenannten "kleinen" Funktionen dar, weil sie in gewisser Hinsicht eine ähnliche Rolle für Markov-Ketten mit beliebigen Zustandsräumen spielen, wie die einzelnen Zustände im diskreten Fall. Sei $\mathfrak{B}_+^d$ die Klasse

der nicht-negativen meßbaren und φ-positiven Funktionen auf $(\mathbb{R}^d, \mathfrak{B}^d)$ [1], das heißt

$$\mathfrak{B}_+^d = \left\{ f : (\mathbb{R}^d, \mathfrak{B}^d) \to [0,\infty) \text{ , meßbar } \middle| \varphi(f) := \int f(x)\ \varphi(dx) > 0 \right\}$$

und $\mathcal{M}_+$ bezeichne die Klasse der positiven signierten Maße auf $(\mathbb{R}^d, \mathfrak{B}^d)$.

Definition 1.6 *Eine Funktion $s \in \mathfrak{B}_+^d$ bzw. ein Maß $\nu \in \mathcal{M}_+$ heißt klein, wenn ein $m_0 \geq 1$ und ein $\beta > 0$ existieren, sodass*

$$P^{m_0}(x, B) \geq \beta s(x) \nu(B) \tag{1.11}$$

für alle $x \in \mathbb{R}^d$ und $B \in \mathfrak{B}^d$. Eine φ-positive Menge $C \in \mathfrak{B}^d$ heißt klein, falls ihr Indikator 1_C eine kleine Funktion ist, das heißt, falls $m_0 \geq 1$, $\beta > 0$ und $\nu \in \mathcal{M}_+$ existieren, sodass

$$P^{m_0}(x, \cdot) \geq \beta \nu(\cdot) \quad \textit{für alle } x \in C. \tag{1.12}$$

Zunächst ist überhaupt nicht klar, ob überhaupt kleine Funktionen oder Maße existieren. Die Existenz kleiner Funktionen kann mittels der Irreduzibilität des Markov-Kerns P allerdings gezeigt werden (vgl. Theorem 2.1. in [Num]).

1.3 Rekurrenz und Aperiodizität

Irreduzibilität garantiert, dass alle interessanten Mengen mit positiver Wahrscheinlichkeit irgendwann aufgesucht werden und damit erreichbar sind. Rekurrenz ist eine Eigenschaft, die dafür sorgt, dass all diese Mengen unendlich oft aufgesucht werden, unabhängig vom Startpunkt der Kette. Bezeichnet

$$h_B^\infty(x) = \mathbb{P}_x(X^{(n)} \in B \quad \text{u.o.}) \tag{1.13}$$

die Wahrscheinlichkeit, dass eine Markov-Kette mit Startpunkt x die Menge B unendlich oft aufsucht, so definieren wir formal:

Definition 1.7 *Ein irreduzibler Markov-Kern P heißt rekurrent, wenn*

$$\begin{aligned} &h_B^\infty(x) > 0 \quad \textit{überall} \\ &\textit{und} \\ &h_B^\infty(x) = 1 \quad \textit{φ-f.s. für alle φ-positiven } B \in \mathfrak{B}^d. \end{aligned} \tag{1.14}$$

Er heißt Harris-rekurrent, wenn

$$h_B^\infty(x) \equiv 1 \quad \textit{für alle φ-positiven } B \in \mathfrak{B}^d. \tag{1.15}$$

[1] Achtung: $\mathfrak{B}_+^d$ bezeichnet sowohl einen Funktionenraum als auch ein Mengensystem - eine Menge $B \in \mathfrak{B}_+^d$ entspricht der Indikatorfunktion $1_B \in \mathfrak{B}_+^d$.

Jede Harris-rekurrente Markov Kette ist natürlich auch rekurrent. Im diskreten Fall gilt sogar die Umkehrung (vgl. S. 46 in [Als2]). Die Bedeutung der Harris-Rekurrenz wird im Zusammenhang mit der Ergodizität sehr bald zum Vorschein kommen. Vorher suchen wir aber noch nach einem geeigneten Kriterium für das Vorliegen dieser besonders starken Form der Rekurrenz.

Satz 1.8 *Ist P die Übergangsdichte einer Markov-Kette mit stationärer Verteilung π, welche eine echt positive und stetige d-dimensionale Lebesgue-Dichte f besitzt, so folgt bereits die Harris-Rekurrenz von P.*

Die Beweisidee soll an dieser Stelle kurz skizziert werden.

Beweis: Zunächst ist festzustellen, dass P unter diesen Umständen irreduzibel ist. Das Lebesgue-Maß kann als maximales Irreduzibilitätsmaß identifiziert werden. Bezeichne

$$h_K^m(x) = \mathbb{P}_x(X^{(n)} \in K \text{ für ein } n \geq m)$$

die Wahrscheinlichkeit, dass die Markov-Kette ein nicht-leeres Kompaktum $K \in \mathfrak{B}^d$ mindestens noch einmal zum Zeitpunkt m oder später aufsucht, so gilt:

$$h_K^m(x) \downarrow h_K^\infty(x)$$

für alle $x \in \mathbb{R}^d$ und $K \in \mathfrak{B}^d$, falls $m \to \infty$. Wir haben damit folgende Abschätzung (vgl. S. 242 in [Var]):

$$0 < \pi(K) = \int P^m(x,K)\ \pi(dx) \leq \int h_K^m(x)\ \pi(dx) \leq \int h_K^\infty(x)\ \pi(dx). \tag{1.16}$$

Nach einem Theorem von Varadhan (vgl. S. 235 in [Var]) kann nur $h_K^\infty(x) \equiv 0$ oder $h_K^\infty(x) \equiv 1$ gelten. Gleichung (1.16) schließt Ersteres aus, was gleichbedeutend mit der Harris-Rekurrenz von P ist. □

Die Irreduzibilität eines Markov-Kerns impliziert bereits die Existenz einer kleinen Funktion $s \in \mathfrak{B}_+^d$ und eines kleinen Maßes $\nu \in \mathcal{M}_+$. Zusammen mit der Rekurrenz kann ein stationäres Maß π konstruiert werden (vgl. Korollar 5.2. in [Num]).

Satz 1.9 *Sei P ein irreduzibler und rekurrenter Markov-Kern, dann definiert*

$$\pi_s(\cdot) = \sum_{n=0}^{\infty} \nu(P^{m_0} - s(x)\nu(\cdot))^n \tag{1.17}$$

ein stationäres Maß. Es ist das einzige stationäre Maß, das der Bedingung $\pi_s(s) = \int s(x)\pi_s(dx) = 1$ genügt.

Zu unserem Leidwesen stellen wir fest, dass dieses Maß nicht notwendig endlich sein muss. Wäre $\pi_s(\mathbb{R}^d) < \infty$, so könnten wir eine stationäre Verteilung $\pi = \pi_s/\pi_s(\mathbb{R}^d)$ durch Normierung konstruieren. Wir treffen folgende Klassifizierung rekurrenter Markov-Kerne:

Definition 1.10 *P sei ein irreduzibler und rekurrenter Markov-Kern mit einem stationären Maß π_s, welches durch Formel* (1.17) *gegeben ist. P heißt positiv rekurrent, falls $\pi_s(\mathbb{R}^d) < \infty$. Anderenfalls nennen wir P null rekurrent.*

Eine weitere bedeutsame Eigenschaft im Hinblick auf die Konvergenz der Markov-Kerne P^n stellt ihr periodisches Verhalten dar.

Definition 1.11 *Eine Folge $\{D_0, \ldots, D_{d-1}\}$ nicht-leerer, disjunkter Mengen nennen wir d-Zyklus (für den Kern P), wenn für alle $i = 0, \ldots, d-1$ und alle $x \in D_i$:*

$$P(x, D_j) = 1 \quad \text{für } j = i+1 \pmod{d}. \tag{1.18}$$

Nehmen wir an $s \in \mathfrak{B}_+^d$ sei eine kleine Funktion und $\nu \in \mathcal{M}_+$ ein kleines Maß, sodass die "Minorisationsbedingung" (1.11) erfüllt ist. Die Menge $\{m \geq 1,\ P^m(x,\cdot) \geq \beta_m s(x)\nu(\cdot)$ für ein $\beta_m > 0\}$ ist additiv abgeschlossen. Ihr größter gemeinsamer Teiler erweist sich als unabhängig von der speziellen Wahl der Funktion s und des Maßes ν, was uns zu folgender Definition veranlasst:

Definition 1.12 *Die natürliche Zahl*

$$d := \mathrm{ggT}\{m \geq 1,\ P^m(x,\cdot) \geq \beta_m s(x)\nu(\cdot) \text{ für ein } \beta_m > 0\} \tag{1.19}$$

heißt Periode des Markov-Kerns P. Im Fall $d = 1$ bezeichnen wir P als aperiodisch.

Wiederum ohne Beweis notieren wir (vgl. Theorem 2.2. in [Num]):

Satz 1.13 *Sei d die Periode eines irreduziblen Markov-Kerns P. Dann gilt:*

(*i*) *Es gibt einen d-Zyklus $\{D_0, \ldots, D_{d-1}\}$.*
(*ii*) *Falls ein weiterer d'-Zyklus $\{D'_0, \ldots, D'_{d'-1}\}$ existiert, so ist d ein Teiler von d'.*

Aperiodizität erweist sich als notwendige Bedingung, wenn eine positiv rekurrente Markov-Kette in Totalvariation konvergieren soll. Hat P nämlich die Periode $d \geq 2$, so nimmt die eindeutig bestimmte stationäre Verteilung π für $B \subset D_i$ mit $\varphi(B) > 0$ einen positiven Wert an, wie sich beispielsweise mit der Äquivalenz von stationärer Verteilung π und maximalem Irreduzibilitätsmaß φ begründen lässt. Es gilt aber auch

$$\lim_{n\to\infty} \mathbb{P}_x^{X^{(nd)}}(B) = \lim_{n\to\infty} P^{nd}(x, B) = 0 \neq \pi(B)$$

für alle $x \in \mathbb{R}^d \setminus D_i$.

Bemerkung: Aperiodizität nachzuweisen kann im Einzelfall sehr schwierig sein. Das Vorliegen einer positiven ϕ-Dichte wie wir sie aus dem Irreduzibilitätskriterium des Satzes 1.5 kennen, ist offenbar auch hinreichend für die Aperiodizität der Kette.

1.4 Geometrische Ergodizität

Damit haben wir die begrifflichen Grundlagen für ein bedeutsames Konvergenzresultat, den Ergodensatz, geschaffen (vgl. Korollar 6.7. (ii) in [Num]).

Satz 1.14 *Sei P ein Harris-rekurrenter und aperiodischer Markov-Kern mit eindeutig bestimmter stationärer Verteilung π. Dann gilt:*

$$\lim_{n\to\infty} \| P^n(x,\cdot) - \pi(\cdot) \| = 0 \tag{1.20}$$

für alle $x \in \mathbb{R}^d$.

Bemerkung: Der Ergodensatz erfordert Harris-Rekurrenz, damit das Konvergenzresultat (1.20) tatsächlich für alle $x \in \mathbb{R}^d$ Gültigkeit besitzt. Ohne Harris-Rekurrenz müssten wir die Einschränkung „π-fast sicher“ hinzufügen (vgl. Abschnitt 4.4 in [Tie]).

Ziehen wir ein kurzes Zwischenfazit: Wir haben bisher nach Eigenschaften einer Markov-Kette gesucht, die eine Konvergenz der Verteilungen $\mathbb{P}_x(X^{(n)} \in \cdot)$ gegen ein stationäres Maß $\pi(\cdot)$ garantieren. Namentlich sind dies die positive Harris-Rekurrenz und Aperiodizität eines irreduziblen Markov-Kerns P. Einen solchen Kern nennt man auch einfach ergodisch. Bisher haben wir aber keine Aussagen über die Qualität, sprich die Konvergenzgeschwindigkeit des asymptotischen Verhaltens getroffen. Hierzu werden wir jetzt eine besonders starke Form der Rekurrenz, die geometrische Rekurrenz, einführen. Diese wird alsdann für eine besonders zügige Konvergenz sorgen.

Definition 1.15 *Sei P ein Harris rekurrenter Markov-Kern. P heißt geometrisch rekurrent, wenn eine kleine Menge C mit $\varphi(C) > 0$ und eine Konstante $r > 1$ existieren, sodass*

$$\sup_{x\in C} \mathbb{E}_x[r^{S_C}] < \infty. \tag{1.21}$$

Geometrische Rekurrenz impliziert nicht nur die positive Rekurrenz, sondern ebenso (vgl. Proposition 5.19. in [Num]):

Satz 1.16 *Sei P ein geometrisch rekurrenter Markov-Kern, π seine stationäre Verteilung, C eine kleine Menge, welche* (1.21) *erfüllt, $x \in \mathcal{S}$ ein beliebiger Zustand der Markov-Kette, B eine φ-pos. Menge und $r > 1$ eine Konstante, die von π, C bzw. x abhängt. Dann gilt:*

$$(i) \quad \mathbb{E}_\pi[r^{S_B}] < \infty, \tag{1.22}$$

$$(ii) \quad \sup_{x\in C} \mathbb{E}_x[r^{S_B}] < \infty \tag{1.23}$$

und außerdem

$$(iii) \quad \mathbb{E}_x[r^{S_B}] < \infty \quad \pi\text{-f.s. für alle } x \in \mathbb{R}^d. \tag{1.24}$$

Eine geometrisch rekurrente, ergodische Markov-Kette wird auch als geometrisch oder exponentiell ergodisch bezeichnet. Es kann gezeigt werden, dass dann die n-Schritt Übergangskerne $P^n(x,B)$ π-fast sicher mit einer geometrischen Rate gegen ihre stationären Limiten $\pi(B)$ konvergieren. Diese Konvergenz erfolgt gleichmäßig über alle $B \in \mathfrak{B}^d$ (vgl. Theorem 6.14. [Num]).

Satz 1.17 *Sei P ein ergodischer Markov-Kern. P ist genau dann geometrisch ergodisch, wenn eine π-integrierbare Funktion $M \in L^1(\pi)$ und eine Konstante $\rho < 1$ existieren, sodass*

$$\| P^n(x,\cdot) - \pi(\cdot) \| \leq M(x)\rho^n \tag{1.25}$$

für alle $x \in \mathbb{R}^d$ und $n \geq 0$.

Satz 1.17 veranlasst uns sofort zu einer weiteren Definition.

Definition 1.18 *Die kleinste Konstante $\rho^* \in [0,1)$, für die eine Funktion M existiert, sodass Bedingung* (1.25) *erfüllt ist, nennen wir Konvergenzrate. Das heißt:*

$$\rho^* = \inf\{\rho \in [0,1) : \exists M \in \mathcal{L}(\pi),\ \textit{sodass}\ (1.25)\ \textit{erfüllt ist}\}. \tag{1.26}$$

Dem aufmerksamen Leser wird nicht entgangen sein, dass die Abschätzung aus Gleichung (1.25) immer noch eine Abhängigkeit von dem Startpunkt $x \in \mathbb{R}^d$ aufweist. Eine Verschärfung liefert somit die folgende Definition.

Definition 1.19 *Eine ergodische Markov-Kette heißt gleichmäßig exponentiell ergodisch, wenn eine endliche Konstante M und eine weitere Konstante $\rho \in [0,1)$ existieren, so dass*

$$\| P^n(x,\cdot) - \pi(\cdot) \| \leq M\rho^n \tag{1.27}$$

für alle $x \in \mathbb{R}^d$ und $n \geq 0$.

Geometrische Rekurrenz ist eine notwendige Voraussetzung für die geometrische Ergodizität. Häufig erweist sich ihr direkter Nachweis über Defintion 1.15 als schwierig. Deshalb scheint es wünschenswert, ein handliches Kriterium für das Vorliegen geometrischer Rekurrenz zu finden. Hilfreich ist die folgende Charakterisierung geometrischer Rekurrenz durch eine Driftbedingung (vgl. Proposition 5.21. in [Num]). I_B bezeichne den Kern $I(x,B) = 1_B(x)$.

Satz 1.20 *Die Markov-Kette $(X^{(n)})$ ist geometrisch rekurrent, wenn eine kleine Menge C, eine nicht-negative Funktion $g \in \mathfrak{B}_+^{\mathfrak{d}}$ und eine Konstante $r > 1$ existieren, sodass*

$$\sup_{x \in C^c} \mathbb{E}[rg(X_{n+1}) - g(X_n) | X_n = x] = \sup_{x \in C^c} (rPg(x) - g(x)) < 0 \tag{1.28}$$

und

$$\sup_{x \in C} \mathbb{E}[g(X_{n+1}); X_{n+1} \in C^c | X_n = x] = \sup_{x \in C} PI_{C^c} g(x) < \infty. \tag{1.29}$$

Beweis: Gleichung (1.28) lässt sich auch schreiben als

$$g(x) \geq rI_{C^c}Pg(x) + \gamma 1_{C^c}(x) \tag{1.30}$$

für ein $\gamma > 0$. Eine iterative Anwendung von g auf der rechten Seite der Gleichung liefert:

$$g(x) \geq r^{n+1}(I_{C^c}P)^{n+1}g(x) + \gamma \sum_{i=0}^{n} r^i (I_{C^c}P)^i 1_{C^c}(x).$$

Da sich g bezüglich des Kerns $I_{C^c}P$ superharmonisch verhält, folgt mittels des Rieszschen Zerlegungstheorems aus der Potentialtheorie, dass der erste Summand auf der rechten Seite für $n \to \infty$ verschwindet, sofern r klein genug gewählt wurde. Damit haben wir

$$\begin{aligned} g(x) &\geq \gamma \sum_{i=0}^{\infty} r^i (I_{C^c}P)^i 1_{C^c}(x) \\ &= \gamma 1_{C^c}(x) \sum_{i=0}^{\infty} r^i \mathbb{P}_x(X_i \in C^c, S_C \geq i) \\ &= \gamma 1_{C^c}(x) \sum_{i=0}^{\infty} r^i \mathbb{P}_x(S_C > i). \end{aligned}$$

$S_C = \inf\{n \geq 1 : X^{(n)} \in C\}$ bezeichnet hierbei die Ersteintrittszeit in die Menge C. Die Summe in der letzten Gleichung kann aber mit der Integrationsformel (A.11) aus [Als1], welche auf dem Satz 19.13. fußt, auch als Erwartungswert einer von S_C abhängigen Funktion geschrieben werden:

$$\sum_{i=0}^{\infty} r^i \mathbb{P}_x(S_C > i) = \frac{1}{r-1}\mathbb{E}_x(r^{S_C} - 1),$$

das heißt,

$$g(x) \geq \frac{\gamma}{r-1} 1_{C^c}(x)\mathbb{E}_x(r^{S_C} - 1).$$

Wenden wir nun ein weiteres Mal den Kern PI_{C^c} auf g an und verlangen zudem, dass die Kette in C starten soll, so erhalten wir mit Voraussetzung (1.29):

$$\frac{\gamma}{r-1} \sup_{x \in C} \mathbb{E}_x(r^{S_C - 1} - 1) \leq \sup_{x \in C} PI_{C^c}g(x) < \infty. \tag{1.31}$$

Man beachte, dass der Exponent von r gerade um den Wert 1 verschoben wird, da sich der Ersteintritt von C genau um eine Einheit verzögert. Damit folgt aber nun leicht das gewünschte Resultat

$$\sup_{x \in C} \mathbb{E}_x r^{S_C} < \infty,$$

welches die geometrische Rekurrenz beschreibt. □

Schauen wir noch einmal auf Ungleichung (1.28) oder (1.30). Die Funktion g kann als eine Art Energiefunktion interpretiert werden, deren Werte steigen, wenn sich x von der kleinen Menge C entfernt. Für Startpunkte $x \notin C$ beschränkt die Ungleichung (1.30)

die erwartete Energieänderung $Pg(x) - g(x)$ durch $-\gamma g(x)$, einem negativen Wert, welcher sich proportional zum gegenwärtigen Energieniveau verhält. Die negative Schranke bedeutet, dass die Kette sich tendentiell zu niedrigeren Engergiezuständen bewegen wird. Oder anders ausgedrückt, gegen C driftet. Die Proportionalität zum gegenwärtigen Energieniveau stellt sicher, dass die Verteilung der Rückkehrzeit zu C geometrische „Tails“ aufweist, was die geometrische Ergodizität für eine irreduzible und aperiodische Markov-Kette impliziert (vgl. die Ausführungen nach Theorem 4.7 in [Tie]).

2 Der Gibbs Sampler

2.1 Herkunft, Definition und Eigenschaften

Der Gibbs-Sampler ist ein Algorithmus zur Simulation multivariater Verteilungen $\pi(x_1,\ldots,x_d)$ auf Basis der univariaten bedingten Verteilungen (u.b.V.)

$$\pi_i(x_i \mid x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d) \quad i=1,\ldots,d, \tag{2.1}$$

den so genannten „full conditionals“. Der Name entstammt einer Arbeit von Geman und Geman aus dem Jahr 1984 (siehe [Ge/Ge]), die das Verfahren zur Restauration von Bildern benutzten und hiermit der Technik zu Popularität verhalfen. Sie analysierten Gibbs Verteilungen auf Gittern. Allerdings lassen sich auch andere Verteilungen verwenden, sodass die Bezeichnung Gibbs Sampling eher unglücklich gewählt ist. Der Ursprung der Methode liegt in der Statistischen Physik, wo sie bereits unter der Bezeichnung „heat bath algorithm“ bekannt war. Heute basieren die meisten statistischen Anwendungen von Markov Chain Monte Carlo Verfahren auf Gibbs Sampling.

Definition 2.1 *Sei π die zu simulierende Zielverteilung, deren univariate bedingte Verteilungen aus* (2.1) *bekannt sind. Der zugehörige Gibbs Sampler mit systematischer Abtastung (systematic scan) wird durch das folgende Übergangsschema erklärt: Generiere den Zufallsvektor $X^{(n+1)}$ zum Zeitpunkt $n+1$ durch den bereits gegebenen Wert $X^{(n)} = (x_1^{(n)},\ldots,x_d^{(n)})$ gemäß des folgenden Schemas:*

$$\begin{array}{lrcl}
1. & X_1^{(n+1)} & \sim & \pi_1(x_1 \mid x_2^{(n)},\ldots,x_d^{(n)}),\\
2. & X_2^{(n+1)} & \sim & \pi_2(x_2 \mid x_1^{(n+1)},x_3^{(n)},\ldots,x_d^{(n)}),\\
 & & \vdots & \\
d. & X_d^{(n+1)} & \sim & \pi_d(x_d \mid x_1^{(n+1)},\ldots,x_{d-1}^{(n+1)}).
\end{array}$$

Nach Vorgabe eines Anfangswertes $X^{(0)}$ liefert die iterative Anwendung des Schemas eine stochastische Folge $X^{(0)},X^{(1)},\ldots$, welche auch Gibbs-Sequenz genannt wird.

Bemerkung: (a) Der Anfangswert $X^{(0)}$ bzw. seine Verteilung wird mehr oder weniger willkürlich festgelegt. Anschließend verwirft man die ersten n Werte der Folge in der Annahme, dass die Zufallsvektoren $X^{(n)}$ sich nach einer gewissen Zeit nahezu unabhängig von $X^{(0)}$ verhalten. Die Festlegung dieser sogenannten „burn-in-Periode“ erfolgt ebenfalls nach heuristischen Verfahren (vgl. Abschnitt 1.4.6 in [Gi/Ri/Sp] oder siehe auch [Mac]).

(b) In Definition 2.1 verwenden wir die Begriffe „simulieren“ bzw. „generieren“ in synonymer Weise. Die Erzeugung von Realisierungen einer Zufallsvariablen mit einer vorgegebenen Verteilung stellt ein nicht-triviales Problem dar. Gibbs-Sampling reduziert dieses Problem auf den eindimensionlen Fall. Wenn die Pseudo-Inverse

$$F^{-1}(y) := \inf\{x \in \mathbb{R} : F(x) \geq y\}, \quad y \in (0,1) \tag{2.2}$$

existiert, kann man Zufallszahlen X erzeugen, die der Verteilung $F^{-1}(U)$ genügen, sofern U eine auf $(0,1)$ gleichverteilte Zufallsgröße U darstellt (vgl. Lemma 36.8. in [Als1]). Leider existiert in den meisten Fällen keine Pseudo-Inverse. Dennoch vereinfacht der Gibbs Sampler die Simulation einer mehrdimensionalen Verteilung, indem er die Simulation eines d-dimensionalen Zufallsvektors X auf d Simulationen eindimensionaler Verteilungen $X_1, \ldots, X_d$ zurückführt.

(c) Auf den ersten Blick mag sich ein potentieller Anwender Fragen, ob es überhaupt realistische Anwendungen gibt, bei denen einerseits die multivariate Zielverteilung π unbekannt ist, während andererseits sämtliche univariaten bedingten Verteilungen vorliegen. Eine solche Informationslage besteht aber des Öfteren, insbesondere bei Zufallsprozessen auf Gittern in den Naturwissenschaften. Häufig besitzen die univariaten bedingten Verteilungen nämlich eine äußerst einfache Form, da lediglich den engsten Nachbarn gegenüber ein Abhängigkeitsverhältnis besteht (vgl. Abschnitt 2.3 in [Gel/Sm]). Sei $S = \{1, \ldots, d\}$ und bezeichne T_i eine von i abhängige Teilmenge von S, so nehmen die bedingten Verteilungen regelmäßig die einfachere Gestalt

$$\pi_i(x_i | x_j, j \in S \setminus \{i\}) = \pi_i(x_i | x_j, j \in T_i) \tag{2.3}$$

an. Dieser plausible Zusammenhang kommt in der Markov Random Field Theorie zum Tragen.

(d) Genaugenommen kann der Gibbs Sampler als Spezialfall des Metropolis-Hastings-Algorithmus angesehen werden. Bekanntlich wird hierbei ein neuer Wert $Y = X^{(n+1)}$ mit einer Akzeptanzwahrscheinlichkeit $\alpha(x,y)$ angenommen, die sowohl vom anvisierten neuen als auch vom aktuellen Wert $X = X^{(n)}$ abhängt. Der Kandidat Y wird mittels eines Kandidatenkerns $Q(X^{(n)}, \cdot)$ („proposal“ distribution) generiert. Über seine Annahme entscheidet schließlich ein Bernoulli-Experiment mit Akzeptanzwahrscheinlichkeit

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

wobei π bzw. q die Dichten der Zielverteilung bzw. des Kandidatenkerns bezeichnen. Anstatt den Zufallsvektor X en bloc zu aktualisieren, empfielt es sich, ein komponentenweises Vorgehen zu wählen, so wie es Metropolis in seinem Ursprungswerk von 1953 auch vorgeschlagen hat (siehe [Me]). Jede Komponente besitzt somit eine eigene Kandidatendichte $q_i(y_i | y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d)$. Im Falle des Gibbs Sampling stehen die univariaten bedingten Verteilungen zur Verfügung. Verwendet man diese als Kandidatenkerne, so erhält man in jeder Komponente eine Akzeptanzwahrscheinlichkeit $\alpha(x,y) = 1$. Die univariaten bedingten Verteilungen enthalten implizit ein hohes Maß an Information über die Zielverteilung, sodass sie als Kandidaten-Verteilungen besser geeignet sind,

als willkürlich gewählte Verteilungen, die mit der Zielverteilung wenig gemein haben. Insofern erklärt sich die hohe Effizienz (vgl. Theorem 7.1.17 in [Ro/Ca] oder siehe auch S.10-12 in [Gi/Ri/Sp] und [Gh/Gr]).

Für die Gibbs-Sequenz $X^{(0)}, X^{(1)}, \ldots$ halten wir das folgende fundamentale Resultat fest:

Satz 2.2 *Die Gibbs-Sequenz stellt eine zeitlich homogene Markov-Kette mit stationärer Verteilung π dar.*

Beweis: Die Übergangswahrscheinlichkeit P von x nach y lässt sich unmittelbar aus der Definition des Gibbs Samplers ablesen und besitzt folgende Gestalt:

$$P(x,y) = \prod_{i=1}^{d} \pi_i(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d). \tag{2.4}$$

Insbesondere sei die Unabhängigkeit vom Zeitindex erwähnt. Die Stationarität von π ergibt sich bereits aus der Konstruktion des Übergangsschemas in Definition 2.1 und kann natürlich auch explizit nachgerechnet werden (vgl. etwa Theorem 7.1.9 in [Ro/Ca]). □

Bemerkung: In der Regel prüft man die Stationarität einer Markov-Kette über die detaillierte Gleichgewichtsgleichung $\pi(x)P(x,y) = \pi(y)P(y,x)$ für alle $x, y \in \mathcal{S}$ nach (P bezeichnet natürlich die Dichte des Übergangskerns P und wird ebenfalls mit P bezeichnet. Identische Bezeichnungen für Verteilung und Dichte sind in der Literatur üblich (siehe Abschnitt 4.2 in [Tie])). Ist die detaillierte Gleichgewichtsgleichung erfüllt, so folgt die Reversibilität der Kette und hieraus die Stationarität von π. Umgekehrt wird die Stationarität einer Markov-Kette mit dem Metropolis-Hastings-Algorithmus erzwungen, indem die Akzeptanzwahrscheinlichkeit gerade so zu wählen ist, dass die detaillierte Gleichgewichtsgleichung erfüllt ist (vgl. Satz 2.4 in [Str]).

Korollar 2.3 *Der Gibbs Sampler mit systematischer Abtastung ist offensichtlich nicht reversibel.*

Reversibilität lässt sich aber einfach herstellen, zum Beispiel indem man das Schema aus Definition 2.1 zum Übergang von $X^{(n)}$ nach $X^{(n+1)}$ in folgender Weise erweitert:

$$\begin{aligned}
1.\quad & X_1^* && \sim \pi_1(x_1 \mid x_2^{(n)}, \ldots, x_d^{(n)}),\\
2.\quad & X_2^* && \sim \pi_2(x_2 \mid x_1^*, x_3^{(n)}, \ldots, x_d^{(n)}),\\
& \vdots\\
d.\quad & X_d^{(n+1)} && \sim \pi_d(x_d \mid x_1^*, \ldots, x_{d-1}^*),\\
d+1.\quad & X_{d-1}^{(n+1)} && \sim \pi_{d-1}(x_{d-1} \mid x_1^*, \ldots, x_{d-2}^*, x_d^{(n+1)}),\\
& \vdots\\
2d-1.\quad & X_1^{(n+1)} && \sim \pi_1(x_1 \mid x_2^{(n+1)}, \ldots, x_d^{(n+1)}),
\end{aligned}$$

Eine Abkehr von der systematischen Abtastung des zugrundeliegenden Raumes $\mathbb{R}^d$ führt ebenfalls zur Reversibilität und findet Ausdruck in folgender Definition:

Definition 2.4 *Sei π wiederum eine gewünschte Zielverteilung, deren univariate bedingte Verteilungen vorliegen und sei ξ eine Wahrscheinlichkeitsverteilung auf dem Raum $(D,\mathfrak{P}(D)) = (\{1,\ldots,d\},\mathfrak{P}(\{1,\ldots,d\}))$. Dann wird das folgende Übergangsschema von $X^{(n)}$ nach $X^{(n+1)}$ als Gibbs Sampler mit zufälliger Abtastung (random scan) bezeichnet: Generiere zunächst i mittels ξ und dann $X^{(n+1)}$ mit gegebenem $X^{(n)} = (x_1^{(n)},\ldots,x_d^{(n)})$ gemäß des folgenden Schemas:*

1. $X_i^{(n+1)} \sim \pi_i(x_i \mid x_1^{(n)},\ldots,x_{i-1}^{(n)},x_{i+1}^{(n)},\ldots,x_d^{(n)})$

und setze

2. $X_j^{(n+1)} := x_j^{(n)}$ *für alle* $j \neq i$.

Es ergibt sich folgendes Resultat:

Lemma 2.5 *Die Markov-Kette eines Gibbs Samplers mit zufälliger Abtastung ist reversibel, wenn ξ die Gleichverteilung auf $(D,\mathfrak{P}(D))$ darstellt.*

Beweis: Bezeichne C_j die Menge aller $y \in \mathbb{R}^d$, die sich nur in der j-ten Komponente von x unterscheiden, das heißt

$$C_j = \{y \in \mathbb{R}^d \mid x_i = y_i \text{ für alle } i \neq j\}. \tag{2.5}$$

C_j bildet eine Äquivalenzklasse. Wir notieren für die Übergangsdichte P des Gibbs Samplers:

$$P(x,y) = \begin{cases} \frac{1}{d}\frac{\pi(y)}{\int_{C_j}\pi(z)\,dz} & \text{falls } x \in C_j \text{ und} \\ 0 & \text{sonst.} \end{cases} \tag{2.6}$$

Nun brauchen wir nur noch die detaillierte Gleichgewichtsgleichung nachrechnen:

$$\pi(x)P(x,y) = \frac{1}{d}\frac{\pi(x)\pi(y)}{\int_{C_j}\pi(z)\,dz} = \frac{1}{d}\frac{\pi(y)\pi(x)}{\int_{C_j}\pi(z)\,dz} = \pi(y)P(y,x) \tag{2.7}$$

für alle $x,y \in \mathbb{R}^d$. Die Kette ist somit reversibel mit stationärer Verteilung π. □

Neben den hier vorgestellten Abtastungsschemata existieren auch noch andere. Es soll keineswegs überraschen, dass die Konvergenzgeschwindigkeit gegen die stationäre Verteilung durchaus von der Wahl des Übergangsschemas abhängt (siehe z.B. [Ro/Sa]).

2.2 Informationsgehalt der univariaten bedingten Verteilungen und Versagen des Gibbs Samplers

Eine erstaunliche Eigenschaft des Gibbs Sampler besteht darin, dass die univariaten bedingten Verteilungen $\pi_i(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d)$ tatsächlich genügend Information enthalten, um eine Stichprobe, die der Verteilung π genügt, zu generieren (vgl. Abschnitt 7.1.5 und Theorem 7.1.19 in [Ro/Ca]). Liegen etwa im Fall $d = 2$ die beiden bedingten

Verteilungen $\pi_1(x_1|x_2)$ und $\pi_2(x_2|x_1)$ vor und bezeichne $\pi^{(1)}$ bzw. $\pi^{(2)}$ die Marginalverteilungen von X_1 bzw. X_2, so lässt sich die gemeinsame Verteilung als

$$\pi(x_1,x_2) = \pi_1(x_1|x_2)\pi^{(2)}(x_2) = \pi_2(x_2|x_1)\pi^{(1)}(x_1) \tag{2.8}$$

schreiben. Aus

$$\pi^{(2)}(x_2) = \frac{\pi_2(x_2|x_1)}{\pi_2(x_1|x_2)}\pi^{(1)}(x_1)$$

folgt

$$\int \frac{\pi_2(x_2|x_1)}{\pi_2(x_1|x_2)}\, dx_2 = \frac{1}{\pi^{(1)}(x_1)},$$

da $\pi^{(2)}(x_2)$ eine Dichte ist, für die natürlich $\int \pi^{(2)}(x_2)\ dx_2 = 1$ gilt. Damit hätten wir dann eine Darstellung für die gemeinsame Verteilung gefunden, sofern diese überhaupt existiert:

$$\pi(x_1,x_2) = \frac{\pi_2(x_2|x_1)}{\int \pi_2(v|x_1)/\pi_2(x_1|v)}\, dv. \tag{2.9}$$

Im allgemeinen Fall ($d > 2$) führen ähnliche Manipulationen der univariaten bedingten Verteilung zu einem Resultat, dass unter dem Namen Hammersley-Clifford-Theorem bekannt ist.

Erinnern wir noch einmal kurz an die Definition der Marginalverteilung

$$\pi^{(1)}(x_1) = \int \pi(x_1,x_2)\ dx_2.$$

Setzen wir die Identität aus Gleichung (2.8) ein, erhalten wir

$$\pi^{(1)}(x_1) = \int \pi_1(x_1|x_2)\pi^{(2)}(x_2)\ dx_2.$$

Ebenso kann die andere Marginalverteilung $\pi^{(2)}(x_2)$ ausgedrückt werden, sodass

$$\begin{aligned}
\pi^{(1)}(x_1) &= \int \pi_1(x_1|x_2) \int \pi_1(x_2|v)\pi^{(1)}(v)\ dv\ dx_2 \\
&= \int \left[\int \pi_1(x_1|x_2)\pi_1(x_2|v)\ dx_2\right] \pi^{(1)}(v)\ dv \\
&= \int h(x_1,v)\pi^{(1)}(v)\ dv,
\end{aligned} \tag{2.10}$$

wobei

$$h(x_1,v) = \int \pi_1(x_1|x_2)\pi_1(x_2|v)\ dx_2. \tag{2.11}$$

Gleichung (2.11) definiert eine Fixpunktgleichung mit der eindeutig bestimmten Lösung $\pi^{(1)}(x_1)$ (vgl. hierzu auch Abschnitt 2.1 in [Gel/Sm] oder Abschnitt 4.1 in [Ca/Ge]).

Eine hinreichende Bedingung für die Irreduzibilität der Gibbs-Sequenz ist die Positivitätsbedingung, welche im Einzelfall nicht einfach zu verifizieren ist (vgl. den Abschnitt oberhalb von Gleichung (2.2) in [Bes]):

Definition 2.6 *Sei $X = (X_1,\ldots,X_d)$ ein Zufallsvektor mit Dichte $\pi(x_1,\ldots,x_d)$ und bezeichne $\pi^{(i)}$ die Marginalverteilung von X_i. Falls $\pi^{(i)}(x_i) > 0$ für alle $i = 1,\ldots,d$ bereits $\pi(x_1,\ldots,x_d) > 0$ impliziert, so genügt π der Positivitäts-Bedingung.*

Mit dieser Bedingung können wir nun das bereits angesprochene Hammersley-Clifford-Theorem formulieren (vgl. Theorem 7.1.20 in [Ro/Ca].

Satz 2.7 *Unter der Positivitätsbedingung kann π explizit berechnet werden und besitzt für ein beliebig vorgegebenes $y \in \mathbb{R}^d$ folgende Gestalt:*

$$\pi(x_1,\ldots,x_d) = \prod_{j=1}^{d} \frac{\pi_j(x_j|x_1,\ldots,x_{j-1},y_{j+1},\ldots,y_d)}{\pi_j(y_j|x_1,\ldots,x_{j-1},y_{j+1},\ldots,y_d)} \pi(y_1,\ldots,y_d).$$

Beweis: Es gilt

$$\begin{aligned}
\pi(x_1,\ldots,x_d) &= \pi_d(x_d|x_1,\ldots,x_{d-1})\pi^d(x_1,\ldots,x_{d-1}) \\
&= \frac{\pi_d(x_d|x_1,\ldots,x_{d-1})}{\pi_d(y_d|x_1,\ldots,x_{d-1})} \pi(x_1,\ldots x_{d-1},y_d) \\
&= \frac{\pi_d(x_d|x_1,\ldots,x_{d-1})}{\pi_d(y_d|x_1,\ldots,x_{d-1})} \frac{\pi_{d-1}(x_{d-1}|x_1,\ldots,x_{d-2},y_d)}{\pi_{d-1}(y_{d-1}|x_1,\ldots,x_{d-2},y_d)} \\
&\quad \times \pi(x_1,\ldots,y_{d-1},y_d).
\end{aligned}$$

Ein Rekursionsargument liefert das gewünschte Ergebnis. □

Klar, dass π die stationäre Verteilung darstellen muss, wenn man sich ihrer univariaten bedingten Verteilungen bedient, um den Gibbs Sampler zu initialisieren (vgl. Satz 2.2). Umgekehrt ist es aber keineswegs sicher, dass zu beliebig vorgegebenen univariaten Verteilungen tatsächlich eine stationäre Verteilung existiert. Betrachten wir drei Beispiele, in denen das zuvor entwickelte Konzept an seine Grenzen stößt:

(a) Sei π die Gleichverteilung auf dem Raum $([-1,0)\times[-1,0)) \cup ([0,1]\times[0,1]) \subset \mathbb{R}^2$. Dann ist $\pi(x|y) = 0$, falls x und y unterschiedliche Vorzeichen aufweisen. Der Quadrant, in dem der Startwert y_0 liegt, wird daher niemals verlassen. Das Beispiel verdeutlicht, dass die Gibbs-Sequenz natürlich nicht unbedingt irreduzibel sein muss. Folglich existiert keine stationäre Verteilung. Die Korrelationsrate ρ^* nimmt in diesem Fall den Wert 1 an (vgl. Seite 52 in [Rob] oder auch Example 7.1.10 in [Ro/Ca]).

(b) Liegen wiederum im bivariaten Fall die bedingten Dichten

$$\begin{aligned}
\pi(x|y) &= ye^{-yx}, \quad 0 < x < \infty \quad \text{und} \\
\pi(y|x) &= xe^{-xy}, \quad 0 < y < \infty
\end{aligned}$$

vor, liefert Gleichung (2.10)

$$\pi^{(1)}(x) = \left[\int ye^{-yx}ve^{-vy}\,dy\right] \pi^{(1)}(v)\,dv \tag{2.12}$$

$$= \int \left[\frac{v}{(x+v)^2}\right] \pi^{(1)}(v)\,dv \tag{2.13}$$

Mit der Substitution $\pi^{(1)}(v) = \frac{1}{v}$ erhalten wir

$$\frac{1}{x} = \int \left[\frac{v}{(x+v)^2} \right] \frac{1}{v} \, dt$$

als Lösung der Fixpunktgleichung (2.12). Allerdings ist $1/x$ offensichtlich keine Wahrscheinlichkeitsdichte. Eine hinreichende Bedingung für die Konvergenz des Gibbs Samplers ist durch $\int \pi^{(1)}(x) \, dx < \infty$ gegeben (vgl. Example 2 (continued) in [Ca/Ge]).

(c) Sei π in diesem letzten Beispiel eine diskrete Verteilung auf dem Raum $\{0,1\}^d$, wobei d sehr groß gewählt wird. Die Hälfte der Masse konzentriert sich auf 0, die andere Hälfte verteilt sich gleichmäßig auf die übrigen Vektoren:

$$\pi(x_1,\ldots,x_d) = \begin{cases} \frac{1}{2}, & \text{falls } (x_1,\ldots,x_d) = (0,\ldots,0) \\ \frac{1}{2(2^d-1)}, & \text{falls } (x_1,\ldots,x_d) \neq (0,\ldots,0). \end{cases}$$

Wenden wir in dieser Situation den Gibbs Sampler mit systematischer Abtastung an, so stellen wir fest, dass dieser zwischen Phasen, in denen er für lange Zeit ausschließlich den Wert 0 ausgibt, und solchen, in denen der 0-Vektor überhaupt nicht mehr auftritt, alterniert. Könnten wir die Verteilung von π direkt simulieren, so würden wir sehr schnell bemerken, dass sich die Wahrscheinlichkeit für $\pi(X^{(n)})$ dem Wert $1/2$ annähert. Der Gibbs Sampler erweist sich als vollkommen ineffizient. Das Problem entsteht, da wir einem einzelnen Zustand eine relativ hohe Wahrscheinlichkeit zuweisen.

Mit den Resultaten aus dem ersten Kapitel formulieren wir schließlich ein hinreichendes Kriterium für die Ergodizität der Gibbs-Sequenz.

Korollar 2.8 *Besitzt der durch den Gibbs Sampler induzierte Übergangskern P eine positive und stetige Lesbegue-Dichte, so verhält sich die Gibbs-Sequenz ergodisch.*

Beweis: Satz 1.5, Bemerkung 1.3 und Satz 1.8 sorgen für Irreduzibilität, Aperiodizität und Harris-Rekurrenz der zugrundeliegenden Markov-Kette. □

3 Geometrische Konvergenz des Gibbs Samplers

3.1 Eigenschaften der Zielverteilung μ

Nachdem wir uns im vorangegangenen Kapitel mit den grundlegenden Eigenschaften des Gibbs Samplers beschäftigt haben, versuchen wir jetzt, seine geometrische Ergodizität zu beweisen. In voller Allgemeinheit wird dies natürlich nicht gelingen, sodass wir uns veranlasst sehen, gewisse Einschränkungen bezüglich der Klasse der Zielverteilungen vorzunehmen. Sei μ eine Wahrscheinlichkeitsverteilung auf $(\mathbb{R}^d, \mathfrak{B}^d)$ mit einer Lebesgue-Dichte, welche ebenfalls die Bezeichnung μ trage und folgende Gestalt besitze:

$$\mu(x) = \frac{1}{M}\exp(-V(x)).$$

M bezeichne dabei eine Normalisierungskonstante und $V : \mathbb{R}^d \to \mathbb{R}$ sei eine 2 mal stetig differenzierbare, strikt konvexe Abbildung für die $\alpha_1, \alpha_2 > 0$ existieren, sodass

$$\alpha_1 \leq \left(\frac{\partial^2 V}{\partial x_i \partial x_j}(x)\right) \leq \alpha_2 \quad \forall x \in \mathbb{R}^d \text{ und } i, j \in \{1, \ldots, d\}. \tag{3.1}$$

Mit anderen Worten: Egal an welcher Stelle die Hesse-Matrix ausgewertet wird, sämtliche Einträge sind stets positiv und befinden sich im Intervall $[\alpha_1, \alpha_2]$. Unter diesen Annahmen besitzt V ein isoliertes Minimum, das sich gleichsam als globales Minimum entpuppt. Ohne Beschränkung der Allgemeinheit wird angenommen, dass $V(0) = \min V = 0$.[1]

Bemerkung: Insbesondere die d-dimensionale Normalverteilung $\mathcal{N}_d(\mu, \Sigma)$ mit Erwartungswert $\mu = 0$ erfüllt die genannten Voraussetzungen. Bedingung (3.1) ist geradezu auf Polynome zweiten Grades zugeschnitten, da die 2. Ableitungen von Polynomen höheren Grades weiterhin von x abhängen und die 2. Ableitungen affin linearer Funktionen bereits verschwinden. So gesehen zielt unser Vorhaben auf den Nachweis der geometrischen Ergodizität eines Gibbs Samplers, dessen Zielverteilung eine Normalverteilung darstellt.

Unter den zuvor gemachten Voraussetzungen genügt der Betrag des Gradienten $\nabla V(x)$ folgender Abschätzung:

[1] Gelte $W(y) = \min W \neq 0$, so erreicht man die gewünschte Form durch Übergang zu der Funktion $V(x) := W(x+y) - \min W$. Die Verschiebung der Funktion um eine Konstante, in diesem Fall $-\min W$, lässt sich hierbei der Normierungskonstante M zuschlagen.

Lemma 3.1 $\alpha_1|x| \leq |\nabla V(x)| \leq \alpha_2|x|$.

Beweis: Die Einträge der Hesse-Matrix $\operatorname{Hess} V(x)$ sind stetig in x, das heißt

$$\frac{\partial V}{\partial x_i}(x) = \int_0^{x_i} \frac{\partial^2 V}{\partial x_i^2}(x_1,\ldots,x_{i-1},t,x_{i+1}\ldots,x_d)\,dt.$$

Folglich gilt

$$\alpha_1|x_i| \leq \left|\frac{\partial V}{\partial x_i}(x)\right| \leq \alpha_2|x_i|$$

für alle i, sodass die Ungleichung bereits in jeder Komponente erfüllt ist. □

$V(x)$ lässt sich in ähnlicher Weise abschätzen.

Lemma 3.2 $\frac{1}{2}\alpha_1|x|^2 \leq V(x) \leq \frac{1}{2}\alpha_2|x|^2$.

Beweis: Exemplarisch wird die erste Ungleichung bewiesen. Die zweite Ungleichung lässt sich völlig analog beweisen. Gemäß der Taylorschen Formel für Funktionen mehrerer Veränderlicher gilt an der Entwicklungsstelle 0:

$$V(0+x) = V(0) + \langle a,x\rangle + \frac{1}{2}\langle x,Ax\rangle + o(|x|^2),$$

wobei $a := \nabla V(0)$ und $A := (\operatorname{Hess} V)(0)$. Ferner befindet sich in 0 ein lokales Extremum, das heißt $\nabla V(0) = 0$. Annahmegemäß gilt $V(0) = 0$:

$$V(x) = \frac{1}{2}\langle x,Ax\rangle + o(|x|^2).$$

Die stetige Funkion $x \mapsto \langle x,Ax\rangle$ nimmt auf dem Kompaktum $S := \{x \in \mathbb{R}^d : |x| = 1\}$ (Sphäre vom Radius 1) ihr Minimum an:

$$k_1 + 2\varepsilon := \inf\{\langle x,Ax\rangle : x \in S\} > 0, \quad \varepsilon > 0.$$

Aufgrund der positiven Definitheit der Matrix A ist sichergestellt, dass es sich bei k_1 um einen positiven Wert handelt. Sei nun $x \neq 0$, dann ist der normierte Vektor $y := \frac{1}{|x|}x$ Element der Einheitssphäre S, das heißt $\langle y,Ay\rangle \geq k_1 + 2\varepsilon$. Aus

$$\langle y,Ay\rangle = \left\langle \frac{1}{|x|}x, \frac{1}{|x|}Ax \right\rangle = \frac{1}{|x|^2}\langle x,Ax\rangle$$

folgt

$$\langle x,Ax\rangle \geq (k_1 + 2\varepsilon)|x|^2$$

für alle $x \in \mathbb{R}^d$. $o(|x|^2)$ steht natürlich gemäß der Landau-Notation für eine Funktion φ mit der Eigenschaft, dass es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, derart dass $|\varphi(x)| \leq \varepsilon|x|^2$, sofern $|x| < \delta$. Für eben diese x haben wir damit $V(x) \geq \frac{1}{2}k_1|x|^2$ bewiesen. Durch das in Lemma (3.1) in Verbindung mit Formel (3.1) vorgegebene Wachstumsverhalten lässt

sich k_1 einerseits als α_1 identifizieren und lässt sich andererseits die Gültigkeit für beliebige x erklären. □

An dieser Stelle sollen einige grundlegende Bezeichnungsweisen und Notationen eingeführt werden. Sei $X^{(0)}, X^{(1)}, \ldots$ die Markov-Kette in $\mathbb{R}^d$ mit Übergangskern P. Diese Kette sei vektorwertig, das heißt, zu einem beliebigen Zeitpunkt n beobachten wir den Wert

$$X^{(n)} = (x_1^{(n)}, \ldots, x_d^{(n)}) \in \mathbb{R}^d.$$

Mit $Z^{(n)}$ bezeichnen wir den um die 1. Komponente gestutzten Zufallsvektor

$$\begin{aligned} Z^{(n)} &= (x_2^{(n)}, \ldots, x_d^{(n)}) \\ &= (z_2^{(n)}, \ldots, z_d^{(n)}) \in \mathbb{R}^{d-1}, \end{aligned}$$

welcher in den übrigen Komponenten $X^{(n)}$ entspricht. Die zugrundeliegende Markov-Kette ist offensichtlich zeitlich homogen. Dieser Umstand führt zu der Bezeichnungsweise

$$X = X^{(n)} \quad \text{und} \quad Y = X^{(n+1)}.$$

Speziell für den Gibbs-Sampler setzen wir

$$\begin{aligned} W^{(k)} &= (y_1, \ldots, y_{k-1}, x_{k+1}, \ldots x_d) \in \mathbb{R}^{d-1}, \\ Y^{(k)} &= (y_1, \ldots, y_{k-1}, y_k, x_{k+1}, \ldots x_d) \in \mathbb{R}^d, \end{aligned}$$

und

$$P_k(y_k | W^{(k)}) = \frac{1}{Z_k(W^{(k)})} e^{-V(Y^{(k)})}$$

mit

$$Z_k(W^{(k)}) = \int e^{-V(Y^{(k)})}\, dy_k.$$

Vorsicht: Der Index k bei $W^{(k)}$ bezieht sich auf die Komponente, während n zum Beispiel bei $X^{(n)}$ die zeitliche Entwicklung beschreiben soll. Manchmal verwenden wir die Schreibweise $W^{(n,k)}$, sofern es im Kontext sinnvoll erscheint.[2] Der Gibbs-Sampler ist eine Markov-Kette mit Übergangsdichte

$$P(X,Y) = \prod_{k=1}^{d} P_k(y_k | W^{(k)}).$$

3.2 Die Abbildungen ϕ und Ψ

Gemäß Korollar 2.8 erweist sich $\mu(X)$ als stationäre Verteilung. Die Übergangswahrscheinlichkeit $P(X,Y)$ beschreibt den stochastischen Übergang von X nach Y. Welchen

[2] Der Vektor $W^{(k)}$ enthält per definitionem implizit eine zeitliche Komponente, da er gerade den Übergang von $X = X^{(n)}$ nach $Y = X^{(n+1)}$ beschreibt.

Wert wird ein Anwender Y zuweisen, unter der Prämisse, dass X und $P(X,Y)$ bekannt sind? Sehen wir einmal von der Erwartungswertbildung ab, so bleibt die Möglichkeit ein heuristisches Verfahren zu wählen, das im Prizip der Maximum-Likelihood-Methode aus der mathematischen Statistik zur Auffindung eines Schätzers gleicht. Hierzu definiere die Abbildungen $\phi_k : \mathbb{R}^{d-1} \to \mathbb{R}$, deren Wert sich in eindeutiger Weise aus der Gleichung

$$\begin{aligned} & V(y_1, \ldots, y_{k-1}, \phi_k(W^{(k)}), x_{k+1}, \ldots x_d) \\ = & \min_{y_k} V(y_1, \ldots, y_{k-1}, \quad y_k \quad , x_{k+1}, \ldots x_d) \end{aligned} \tag{3.2}$$

ergibt. Wohldefiniertheit liefert die Theorie über implizite Funktionen. Die Funktion aus Formel (3.2) bezeichnen wir auch als

$$\hat{V}_k(W^{(k)}) = V(\tilde{Y}^{(k)}) = V(y_1, \ldots, y_{k-1}, \phi_k(W^{(k)}), x_{k+1}, \ldots x_d)$$

mit $\tilde{Y}^{(k)} = (y_1, \ldots, y_{k-1}, \phi_k(W^{(k)}), x_{k+1}, \ldots x_d)$. Für $k = 1$ schreiben wir einfacher $\hat{V}(\cdot)$ statt $\hat{V}_1(\cdot)$:

$$\begin{aligned} \hat{V}(Z) &= V(\phi_1(Z), Z) \\ &= \min_{x_1} V(x_1, Z). \end{aligned}$$

Für $\hat{V}(Z)$ gibt es eine Abschätzung, welche jener aus Lemma 3.2 entspricht und sich mit eben diesem Lemma auch sofort beweisen lässt.

Lemma 3.3 $\frac{1}{2}\alpha_1|Z|^2 \leq \hat{V}(Z) \leq \frac{1}{2}\alpha_2|Z|^2$.

Beweis: Nach Lemma 3.2 gilt einerseits

$$\begin{aligned} \hat{V}(Z) &= V(\phi_1(Z), Z) \\ &\geq \frac{1}{2}\alpha_1(\phi_1(Z)^2 + |Z|^2) \\ &\geq \frac{1}{2}\alpha_1|Z|^2 \end{aligned}$$

und andererseits

$$\begin{aligned} \hat{V}(Z) &\leq V(0, Z) \\ &\leq \frac{1}{2}\alpha_2|Z|^2. \end{aligned}$$

□

Nur ϕ_1 hängt ausschließlich von $W^{(1)} = Z$ ab; für die übrigen Abbildungen $\phi_2, \ldots, \phi_d$ ist eine teilweise Aktualisierung des Vektors X erforderlich. Möchte man allerdings Y ausschließlich auf Basis von X bzw. Z schätzen, so empfiehlt es sich, die folgenden Abbildungen sukzessive für $Z = (x_2, \ldots, x_d) \in \mathbb{R}^{d-1}$ zu definieren:

$$\begin{aligned} \psi_2(Z) &= \phi_2(\phi_1(Z), z_3, \ldots, z_d), \\ \psi_3(Z) &= \phi_3(\phi_1(Z), \psi_2(Z), z_4, \ldots, z_d), \\ &\vdots \\ \psi_d(Z) &= \phi_d(\phi_1(Z), \psi_2(Z), \ldots, \psi_{d-1}(Z)), \\ \\ \Psi &= (\psi_2, \ldots, \psi_d). \end{aligned}$$

Von besonderem Interesse werden im Folgenden Aussagen über den Betrag des Fehler-Vektors $Y-(\phi_1(Z),\psi_2(Z),\ldots,\psi_d(Z))$ sein. Hilfreich erweist sich eine simple Abschätzung der Normalisierungskonstante $Z_k(W^{(k)})$ mittels $\hat{V}_k(W^{(k)})$.

Lemma 3.4 *Es gibt $c_1,c_2>0$, sodass*

$$c_1\exp(-\hat{V}_k(W^{(k)}))\le Z_k(W^{(k)})\le c_2\exp(-\hat{V}_k(W^{(k)})).$$

Beweis: Bei festem $W^{(k)}$ wird durch $\tilde{V}(y_k)=V(Y^{(k)})-\hat{V}_k(W^{(k)})$ eine Funktion von $\mathbb{R}$ nach $\mathbb{R}$ definiert, die gerade in $y_k=\phi_k(W^{(k)})$ ihr Minimum, den Wert 0, annimmt. Ansonsten übernimmt sie das Wachstumsverhalten von V, sodass ihre 2. Ableitung ebenfalls im Intervall $[\alpha_1,\alpha_2]$ liegt. Damit lässt sich Lemma 3.2 anwenden:

$$\frac{1}{2}\alpha_1(y_k-\phi_k(W^{(k)}))^2\le V(Y^{(k)})-\hat{V}_k(W^{(k)})\le\frac{1}{2}\alpha_2(y_k-\phi_k(W^{(k)}))^2. \tag{3.3}$$

Addition von $\hat{V}_k(W^{(k)})$, Multiplikation mit (-1), Anwendung der Exponentialfunktion und Integration bezüglich y_k liefern

$$\begin{aligned}&\exp(-\hat{V}_k(W^{(k)}))\overbrace{\int\exp(-\frac{1}{2}\alpha_1(y_k-\phi_k(W^{(k)}))^2)\,dy_k}^{=:c_2}\\ \ge&\int\exp(-V(Y^{(k)})\,dy_k=Z_k(W^{(k)})\\ \ge&\exp(-\hat{V}_k(W^{(k)}))\underbrace{\int\exp(-\frac{1}{2}\alpha_2(y_k-\phi_k(W^{(k)}))^2)\,dy_k}_{=:c_1}.\end{aligned}$$

Das uneigentliche Integral $\int\exp(-t^2)\,dt$ entspricht dem Wert der Gamma-Funktion an der Stelle $\frac{1}{2}$, das heißt $\int\exp(-t^2)\,dt=\Gamma(\frac{1}{2})=\sqrt{\pi}$. Mit Hilfe der Substitution $t=\sqrt{\frac{\alpha_1}{2}}(y_k-\phi_k)$ erhält man schließlich:

$$c_2=\int\exp(-\frac{1}{2}\alpha_1(y_k-\phi_k)^2)\,dy_k=\sqrt{\frac{2\pi}{\alpha_1}}. \tag{3.4}$$

c_1 errechnet sich ebenso. □

Lemma 3.5

$$\sup_{Z\neq0}\frac{\hat{V}(\Psi(Z))}{\hat{V}(Z)}=\gamma<1.$$

Beweis: Es ist klar, dass $\hat{V}(\Psi(Z))<\hat{V}(Z)$ und damit $\gamma\le1$ für $Z\neq0$, da V strikt konvex ist. Was passiert aber, wenn $|Z|$ sehr groß wird? Angenommen es gibt ein $\delta>0$, sodass

$$\left|\frac{\partial V}{\partial x_{i+1}}(X^{(i)})\right|<\delta|Z| \tag{3.5}$$

für alle $X^{(i)} = (\phi_1(Z), \psi_2(Z), \ldots, \psi_i(Z), z_{i+1}, \ldots, z_d),\ i = 1, \ldots, d-1$. Dann lässt sich diese partielle Ableitung offenbar auch als Wegintegral auffassen:

$$\begin{aligned}
\delta|Z| &> \frac{\partial V}{\partial x_{i+1}}(X^{(i)}) - \overbrace{\frac{\partial V}{\partial x_{i+1}}(X^{(i+1)})}^{=0} \\
&= \int_0^1 \underbrace{\frac{\partial^2 V}{\partial x_{i+1}^2}\left(X^{(i+1)} + \lambda\,(X^{(i)} - X^{(i+1)})\right)}_{>\alpha_1} \underbrace{\left(z_{i+1} - \psi_{i+1}(Z)\right)}_{\text{unabh. von } \lambda}\, d\lambda \\
&\geq \alpha_1\ |z_{i+1} - \psi_{i+1}(Z)| \quad \text{für alle } i = 1, \ldots, d-1.
\end{aligned} \tag{3.6}$$

Die letzte Zeile ist gleichbedeutend mit

$$|X^{(i)} - X^{(i+1)}| \leq \frac{1}{\alpha_1}\delta|Z|, \tag{3.7}$$

da nur die $(i+1)$-ten Komponenten verschieden sind. Es gilt folglich auch

$$|X^{(i)} - X^{(1)}| \leq (i-1)\frac{1}{\alpha_1}\delta|Z| \leq \frac{d}{\alpha_1}\delta|Z|. \tag{3.8}$$

Wir wissen, dass $\partial V/\partial x_{i+1}$ einer Lipschitz-Bedingung genügt, das heißt, es gibt ein $c > 0$, sodass

$$\left|\frac{\partial V}{\partial x_{i+1}}(X^{(i)}) - \frac{\partial V}{\partial x_{i+1}}(X^{(1)})\right| \leq c\left|(X^{(i)}) - (X^{(1)})\right|. \tag{3.9}$$

Mit der Dreiecksungleichung und der Annahme aus (3.5) kann die linke Seite der Ungleichung weiter abgeschätzt werden:

$$\begin{aligned}
\left|\frac{\partial V}{\partial x_{i+1}}(X^{(i)}) - \frac{\partial V}{\partial x_{i+1}}(X^{(1)})\right| &\geq \left|\left|\frac{\partial V}{\partial x_{i+1}}(X^{(i)})\right| - \left|\frac{\partial V}{\partial x_{i+1}}(X^{(1)})\right|\right| \\
&\geq \left|\frac{\partial V}{\partial x_{i+1}}(X^{(1)})\right| - \delta|Z|
\end{aligned} \tag{3.10}$$

und damit

$$\left|\frac{\partial V}{\partial x_{i+1}}(X^{(1)})\right| \leq c\frac{d}{\alpha_1}\delta|Z| + \delta|Z| = \left(1 + cd\frac{1}{\alpha_1}\right)\delta|Z|. \tag{3.11}$$

Folglich lässt sich der Gradient von V an der Stelle $X^{(1)}$ durch

$$\left|\nabla V(X^{(1)})\right| \leq d\left(\left(1 + cd\frac{1}{\alpha_1}\right)\right)\delta|Z| \tag{3.12}$$

abschätzen, was für ein klein gewähltes δ im Widerspruch zu der Aussage des Lemmas 3.1 steht. Fixieren wir dieses δ, so stellen wir im Gegensatz zu unserer ursprünglichen Annahme aus Gleichung (3.5) fest, dass

$$\left|\frac{\partial V}{\partial x_{i+1}}(X^{(i)})\right| \geq \delta|Z|. \tag{3.13}$$

Bemühen wir noch einmal Ungleichung (3.6) und schätzen wir diesmal in die andere Richtung:

$$\begin{aligned}\delta|Z| &\leq \frac{\partial V}{\partial x_{i+1}}(X^{(i)}) - \overbrace{\frac{\partial V}{\partial x_{i+1}}(X^{(i+1)})}^{=0} \\ &= \int_0^1 \underbrace{\frac{\partial^2 V}{\partial x_{i+1}^2}\left(X^{(i+1)} + \lambda(X^{(i)} - X^{(i+1)})\right)}_{<\alpha_2} \underbrace{\left(z_{i+1} - \psi_{i+1}(Z)\right)}_{\text{unabh. von } \lambda} d\lambda \\ &\leq \alpha_2 \ |z_{i+1} - \psi_{i+1}(Z)| \quad \text{für alle } i = 1,\ldots,d-1.\end{aligned} \tag{3.14}$$

Haben wir bisher die Idee mit dem Wegintegral dazu verwendet, um Aussagen über das Verhalten von $\partial V/\partial x_{i+1}$ zu gewinnen, so kann mit Hilfe einer zweifachen Anwendung auch auf die Funktion V selbst geschlossen werden. Wir notieren

$$\begin{aligned}&V(X^{(i)}) - V(X^{(i+1)} \\ &\quad = \left(z_{i+1} - \psi_{i+1}(Z)\right)^2 \int_0^1 \int_0^1 \frac{\partial^2 V}{\partial x_{i+1}^2}\left(X^{(i+1)} + \lambda\mu(X^{(i)} - X^{(i+1)})\right) \lambda \ d\mu \ d\lambda \\ &\quad \geq \frac{1}{2}\alpha_1 \ \left(z_{i+1} - \psi_{i+1}(Z)\right)^2 \quad \text{für alle } i = 1,\ldots,d-1.\end{aligned} \tag{3.15}$$

Zusammen mit der letzten Zeile aus (3.14) erhalten wir

$$V(X^{(i)}) - V(X^{(i+1)} \geq \frac{1}{2}\alpha_1 \left(\frac{\delta}{\alpha_2}\right)|Z|^2 \tag{3.16}$$

und schließlich ebenso

$$\hat{V}(Z) - \hat{V}(\Psi(Z)) = V(X^{(1)}) - V(X^{(d)} \geq \frac{1}{2}\alpha_1 \left(\frac{\delta}{\alpha_2}\right)|Z|^2. \tag{3.17}$$

Nach Lemma 3.3 ist $\hat{V}(Z) \leq 1/2\alpha_2|Z|^2$, sodass eine Division durch $\hat{V}(Z) \neq 0$ zu folgendem Resultat führt:

$$1 - \frac{\hat{V}(\Psi(Z))}{\hat{V}(Z)} \geq \frac{\alpha_1}{\alpha_2}\left(\frac{\delta}{\alpha_2}\right)^2 =: \delta_0 \tag{3.18}$$

ist gleichbedeutend mit

$$\frac{\hat{V}(\Psi(Z))}{\hat{V}(Z)} \leq 1 - \delta_0 =: \gamma < 1. \tag{3.19}$$

□

Korollar 3.6 *Es gibt ein $m \in \mathbb{Z}$, sodass*

$$|\Psi^{(m)}(Z)| \leq \frac{1}{2}|Z|.$$

Beweis: m-malige Anwendung der Abbildung Ψ auf sich selbst liefert mit Lemma 3.5

$$\hat{V}(\Psi^{(m)}(Z)) \leq \gamma^m \hat{V}(Z).$$

Dies wiederum impliziert mittels Lemma 3.3

$$\begin{aligned} & \frac{1}{2}\alpha_1 |\Psi^{(m)}(Z)|^2 \leq \gamma^m \frac{1}{2}\alpha_2 |Z|^2 \\ \Leftrightarrow \qquad & |\Psi^{(m)}(Z)|^2 \leq \frac{\alpha_2}{\alpha_1}\gamma^m |Z|^2. \end{aligned}$$

Für

$$m \geq \frac{\ln(\frac{\alpha_1}{4\alpha_2})}{\ln \gamma} \quad \text{folgt dann} \quad |\Psi^{(m)}(Z)| \leq \frac{1}{2}|Z|.$$

□

Lemma 3.7 *Es gibt ein $c > 0$, sodass $|\nabla \phi_k| \leq c$ für alle k.*

Beweis: V hat in $\tilde{Y}^{(k)}$ ein lokales Minimum, das heißt,

$$\frac{\partial V}{\partial x_k}(y_1, \ \dots \ , y_{k-1}, \phi_k(W^{(k)}), x_{k+1}, \ \dots \ , x_d) = 0.$$

Differentiation nach y_j bzw. x_j unter Beachtung der Kettenregel ergibt

$$\frac{\partial^2 V}{\partial y_j x_k} + \frac{\partial^2 V}{\partial x_k^2}\frac{\partial \phi_k}{\partial y_j} = 0 \quad \text{bzw.} \quad \frac{\partial^2 V}{\partial x_j x_k} + \frac{\partial^2 V}{\partial x_k^2}\frac{\partial \phi_k}{\partial x_j} = 0 \quad \text{falls } j \leq k-1 \text{ bzw. } j \geq k.$$

Mit Hilfe der in Formel (3.1) formulierten Bedingung gelten dann die Abschätzungen:

$$\left|\frac{\partial \phi_k}{\partial x_j}\right| \leq \frac{\alpha_2}{\alpha_1} \quad \text{und damit natürlich bleibt} \quad |\nabla \phi_k| \leq \frac{\alpha_2}{\alpha_1}\sqrt{d} \quad \text{für alle } k.$$

□

3.3 Abschätzung der Fehler ξ und η

Erinnern wir uns: Mit dem Gibbs-Sampler werden die einzelnen Komponenten sukzessive aktualisiert. Welchen Fehler begehen wir aber, wenn wir die neue Komponente $x_k^{(n+1)}$ durch $\phi_k(W^{(n,k)})$ ersetzen? Eine befriedigende Antwort auf die Frage nach der Güte unserer Schätzung liefert Lemma 3.8. Der Schätzfehler für den Wert $x_k^{(n+1)}$ wird durch die Zufallsgröße

$$\xi_k^{(n)} = x_k^{(n+1)} - \phi_k(W^{(n,k)})$$

beschrieben.

Lemma 3.8 *Es gibt $\beta,\ c>0$, sodass*

$$\mathbb{E}\left[e^{\beta\left|\xi_k^{(n)}\right|^2}\ \middle|\ \mathcal{F}_n\right] \le c$$

für alle X und n. $\mathcal{F}_n$ bezeichnet die von $X^{(0)},\ \dots\ ,X^{(n)}$ erzeugte σ-Algebra.

Beweis: Zunächst sei bemerkt, dass aufgrund der Markov-Eigenschaft

$$\mathbb{E}\left[e^{\beta\left|\xi_k^{(n)}\right|^2}\ \middle|\ \mathcal{F}_n\right] = \mathbb{E}\left[e^{\beta\left|\xi_k^{(n)}\right|^2}\ \middle|\ X^{(n)}\right].$$

Eine Version der faktorisierten bedingten Erwartung ist gegeben durch

$$\begin{aligned}&\mathbb{E}\left[e^{\beta\left|\xi_k^{(n)}\right|^2}\ \middle|\ X^{(n)}=(x_1^{(n)},\dots,x_d^{(n)})\right]\\ &=\int e^{\beta|x_k^{(n+1)}-\phi_k(W^{(n,k)})|^2}\underbrace{\prod_{i=1}^{k}P_i(x_i^{(n+1)}|W^{(n,i)})}_{=:P^k(X^{(n)},X^{(n+1,k)})}\ dx_1^{(n+1)}\dots\ dx_k^{(n+1)}\end{aligned} \tag{3.20}$$

für ein beliebiges k. $P^k(X^{(n)},X^{(n+1,k)})$ aktualisiert nur die ersten k-Komponenten und hat folgende Gestalt:

$$P^k(X^{(n)},X^{(n+1,k)}) = \prod_{i=1}^{k}\frac{e^{-V(x_1^{(n+1)},\dots,x_i^{(n+1)},x_{i+1}^{(n)},\dots,x_k^{(n)})}}{Z_i(W^{(n,i)})}. \tag{3.21}$$

Das Integral im Nenner kann mit Lemma 3.4 abgeschätzt werden:

$$\begin{aligned}Z_i(W^{(n,i)}) &= \int e^{-V(x_1^{(n+1)},\dots,x_i^{(n+1)},x_{i+1}^{(n)},\dots,x_k^{(n)})}\ dx_i^{(n+1)}\\ &\ge c_1 e^{-\hat{V}_i(W^{(n,i)})}.\end{aligned}$$

Einsetzen in Formel (3.21) führt zu

$$\begin{aligned}P^k(X^{(n)},X^{(n+1,k)}) &\le \prod_{i=1}^{k}\frac{e^{-V(x_1^{(n+1)},\dots,x_i^{(n+1)},x_{i+1}^{(n)},\dots,x_k^{(n)})}}{c_1 e^{-\hat{V}_i(W^{(n,i)})}}\\ &= \prod_{i=1}^{k}\frac{1}{c_1}\ e^{-(V(X^{(n+1,k)})-\hat{V}_i(W^{(n,i)}))}\\ &\le \left(\frac{1}{c_1}\right)^k\prod_{i=1}^{k}e^{-\frac{1}{2}\alpha_1(x_i^{(n+1)}-\phi_i(W^{(n,i)}))^2},\end{aligned} \tag{3.22}$$

wobei sich Letzteres sofort aus Formel (3.3) ergibt.

Kehren wir zurück zu unserer faktorisierten bedingten Erwartung aus Formel (3.20). Mit dem Satz von Fubini und Formel (3.22) erhalten wir

$$\begin{aligned}
&\mathbb{E}\Big[e^{\beta\left|\xi_k^{(n)}\right|^2}\ \Big|\ X^{(n)}=(x_1^{(n)},\dots,x_d^{(n)})\Big]\\
&\leq\left(\frac{1}{c_1}\right)^k\int\dots\int e^{\beta|x_k^{(n+1)}-\phi_k(W^{(n,k)})|^2}\ e^{-\frac{1}{2}\alpha_1(x_k^{(n+1)}-\phi_k(W^{(n,k)}))^2}\ dx_k^{(n+1)}\dots\\
&\quad\dots\ e^{-\frac{1}{2}\alpha_1(x_{k-1}^{(n+1)}-\phi_{k-1}(W^{(n,k-1)}))^2}\ dx_{k-1}^{(n+1)}\dots\ e^{-\frac{1}{2}\alpha_1(x_1^{(n+1)}-\phi_1(W^{(n,1)}))^2}\ dx_1^{(n+1)}.
\end{aligned}$$

Wir zeigen nun, dass dieses Mehrfachintegral eine obere Schranke c besitzt, indem die Integrale von innen nach außen sukzessive durch eine Konstante abgeschätzt werden, die alsdann nach vorne gezogen wird und die Berechnung der weiteren Integrale nicht mehr tangiert. Lediglich das innere Integral

$$\begin{aligned}
&\int e^{\beta|x_k^{(n+1)}-\phi_k(W^{(n,k)})|^2}\ e^{-\frac{1}{2}\alpha_1(x_k^{(n+1)}-\phi_k(W^{(n,k)}))^2}\ dx_k^{(n+1)}\\
=&\int e^{(\beta-\frac{1}{2}\alpha_1)(x_k^{(n+1)}-\phi_k(W^{(n,k)}))^2}\ dx_k^{(n+1)}
\end{aligned}\tag{3.23}$$

unterscheidet sich von den übrigen, welche die Gestalt

$$\int e^{-\frac{1}{2}\alpha_1(x_i^{(n+1)}-\phi_i(W^{(n,i)}))^2}\ dx_i^{(n+1)}\quad(i=1,\dots,k-1)$$

besitzen. Letzteres haben wir aber bereits errechnet, nämlich in Lemma 3.4, Formel (3.4):

$$\int\exp(-\frac{1}{2}\alpha_1(y_k-\phi_k)^2)\ dy_k=\sqrt{\frac{2\pi}{\alpha_1}}=c_2.$$

Die Berechnung des Integrals aus Formel (3.23) verläuft analog. Setzt man $t=\sqrt{\frac{\alpha_1}{2}-\beta}\ (y_k-\phi_k)$ sofern $\beta<\frac{\alpha_1}{2}$, erhält man

$$\begin{aligned}
&\int e^{\beta|x_k^{(n+1)}-\phi_k(W^{(n,k)})|^2}\ e^{-\frac{1}{2}\alpha_1(x_k^{(n+1)}-\phi_k(W^{(n,k)}))^2}\ dx_k^{(n+1)}\\
=&\int e^{-t^2}\sqrt{\frac{\alpha_1}{2}-\beta}\ dt\\
=&\sqrt{\frac{\alpha_1}{2}-\beta}\ \sqrt{\pi}
\end{aligned}\tag{3.24}$$

und schließlich

$$\begin{aligned}
\mathbb{E}\Big[e^{\beta\left|\xi_k^{(n)}\right|^2}\ \Big|\ X^{(n)}=(x_1^{(n)},\dots,x_d^{(n)})\Big]&\leq\left(\frac{1}{c_1}\right)^k\ c_2^{k-1}\ \left(\frac{\alpha_1}{2}-\beta\right)^{\frac{1}{2}}\ \pi^{\frac{1}{2}}\\
&=\sqrt{\left(\frac{\alpha_2}{\alpha_1}\right)^k\left(\frac{\alpha_1}{2}\right)\ \left(\frac{\alpha_1}{2}-\beta\right)}\\
&=:c.
\end{aligned}$$

□

Während das gerade bewiesene Lemma 3.8 eine Aussage über die Qualität des Fehlers im Zuge der Aktualisierung einer einzelnen Komponente k trifft, versuchen wir nun dieses Ergebnis auf eine komplette Aktualisierung aller d Komponenten auszudehnen. In der ersten Komponente begehen wir den bereits behandelten Fehler

$$\xi_1^{(n)} = x_1^{(n+1)} - \phi_k(W^{(n,1)}).$$

Dieser Fehler beeinflusst die Schätzfehler in den übrigen Komponenten

$$\begin{aligned}\eta_i^{(n)} =& Z^{(n+1)} - \psi_i(Z^{(n)}) \\ =& Z^{(n+1)} - \phi_i(\phi_1(Z^{(n)}), \psi_2(Z^{(n)}), \ldots, \psi_{i-1}(Z^{(n)}), z_{i+1}^n, \ldots, z_d^n).\end{aligned}$$

Der Zufallsvektor $\eta^{(n)}$ entfernt sich im Allgemeinen in jeder Komponente durch die kumulierten Fehler in den vorangegangenen Komponenten mehr und mehr von dem teilaktualisierten Pendant $\xi^{(n)}$. Eine geometrische Fehlerabschätzung, vergleichbar derer aus Lemma 3.8, ist dennoch möglich.

Satz 3.9 *Für die Zufallsvektoren $\eta^{(0)}, \eta^{(1)}, \ldots$ aus $\mathbb{R}^{d-1}$ gibt es $\beta,\ c > 0$, sodass*

$$\mathbb{E}[e^{\beta|\eta^{(n)}|^2} | \mathcal{F}_n^Z] \leq c$$

für alle n. $\mathcal{F}_n^Z$ bezeichnet die von $Z^{(0)}, \ldots, Z^{(n)}$ erzeugte σ-Algebra.

Beweis: Die Vorgehensweise entspricht jener aus Lemma 3.8. Eine Version der faktorisierten bedingten Erwartung ist gegeben durch

$$\begin{aligned}&\mathbb{E}\left[e^{\beta|\eta^{(n)}|^2} \,\middle|\, Z^{(n)} = (z_2^{(n)}, \ldots, z_d^{(n)})\right] \\ =& \int e^{\beta \Sigma_{i=2}^d |z_i^{(n+1)} - \psi_i(Z^{(n)})|^2} \underbrace{\prod_{i=1}^d P_i(x_i^{(n+1)} | W^{(n,i)})}_{=: P^d(X^{(n)}, X^{(n+1,i)})} \, dx_1^{(n+1)}, dz_2^{(n+1)} \ldots \, dz_d^{(n+1)}. \qquad (3.25)\end{aligned}$$

Obwohl wir unter $\mathcal{F}_n^Z$ bedingen, müssen wir die Komponente $x_1^{(n+1)}$ in der Integration berücksichtigen, da sie Bestandteil der Übergangswahrscheinlichkeit ist.[3] Schätzen wir diese wiederum so ab, wie wir es bereits in Formel (3.22) getan haben, so erhalten wir mit dem Satz von Fubini

$$\begin{aligned}&\mathbb{E}\left[e^{\beta|\eta^{(n)}|^2} \,\middle|\, Z^{(n)} = (z_2^{(n)}, \ldots, z_d^{(n)})\right] \\ \leq& \left(\frac{1}{c_1}\right)^k \int \ldots \int e^{\beta|z_d^{(n+1)} - \psi_d(Z^{(n)})|^2} \, e^{-\frac{1}{2}\alpha_1(z_d^{(n+1)} - \phi_d(W^{(n,d)}))^2} \, dz_d^{(n+1)} \ldots \\ &\ldots \, e^{\beta|z_2^{(n+1)} - \psi_2(Z^{(n)})|^2} \, e^{-\frac{1}{2}\alpha_1(z_2^{(n+1)} - \phi_2(W^{(n,2)}))^2} \, dz_2^{(n+1)} \, e^{-\frac{1}{2}\alpha_1(x_1^{(n+1)} - \phi_1(W^{(n,1)}))^2} \, dx_1^{(n+1)}.\end{aligned}$$

[3] Alternativ könnten wir für $x_1^{(n+1)}$ auch einen beliebigen festen Wert annehmen, um später einzusehen, dass sich unsere Abschätzung als unabhängig von demselben erweist.

Betrachten wir das innere Integral

$$\int e^{\beta|z_d^{(n+1)}-\psi_d(Z^{(n)})|^2}\ e^{-\frac{1}{2}\alpha_1(z_d^{(n+1)}-\phi_d(W^{(n,d)}))^2}\ dz_d^{(n+1)}, \tag{3.26}$$

so lassen sich die beiden Exponenten im Integranden aufgrund der Verschiedenartigkeit von $\psi_d(Z^{(n)})$ und $\phi_d(W^{(n,d)})$ nicht so elegant zusammenfassen, wie wir es in Gleichung (3.23) getan haben. Allerdings sind wir mit Lemma 3.7 in der Lage, den Abstand zwischen diesen beiden Werten relativ grob einzuschränken. Nach besagtem Lemma lässt sich nämlich der Gradient eines jeden ϕ_k betragsmäßig gegen eine Konstante c abschätzen. Daher gilt:

$$\left|\ \phi_2(W^{(n,2)}) - \psi_2(Z^{(n)})\ \right| \leq (x_1^{(n+1)} - \phi_1(W^{(n,1)})\ c\ =\ \xi_1^{(n)} c$$

und damit für die 3. Komponente

$$\left|\ \phi_3(W^{(n,3)}) - \psi_3(Z^{(n)})\ \right| \leq\ (\xi_1^{(n)} c + \xi_2^{(n)})c = \xi_1^{(n)} c^2 + \xi_2^{(n)} c.$$

Fährt man in dieser Weise sukzessive fort, erhält man schließlich für die letzte Komponente:

$$\left|\ \phi_d(W^{(n,d)}) - \psi_d(Z^{(n)})\ \right| \leq \sum_{k=2}^{d}\sum_{j=1}^{k-1} \xi_j^{(n)}\ c^{k-j}.$$

Wenn wir zur Vereinfachung der Darstellung auf sämtliche Argumente, Zeit- und Komponentenindizes verzichten, erhalten wir für das innere Integral aus Formel (3.26)

$$\int e^{\beta(z-\psi)^2-\frac{1}{2}\alpha_1(z-\phi)^2}\ dz.$$

Mittels quadratischer Ergänzung ergibt sich für den Exponenten des Integranden auch

$$\left(\beta-\frac{\alpha_1}{2}\right)\Big(z-\underbrace{\frac{\psi+\phi}{\beta-\frac{\alpha_1}{2}}}_{=:p}\Big)^2 + \underbrace{\Big((\psi^2+\phi^2)-\frac{(\psi+\phi)^2}{\sqrt{\beta-\frac{\alpha_1}{2}}}\Big)}_{=:q}.$$

Das Integral aus Formel (3.26) ist also identisch mit jenem, das wir in Formel (3.24) unter Verwendung einer geeigneten Substitution bereits berechnet hatten:

$$\begin{aligned}
&\int e^{\beta|z_d^{(n+1)}-\psi_d(Z^{(n)})|^2}\ e^{-\frac{1}{2}\alpha_1(z_d^{(n+1)}-\phi_d(W^{(n,d)}))^2}\ dz_d^{(n+1)}\\
=&e^q\ \int e^{(\beta-\frac{\alpha_1}{2})(z_d^{(n+1)}-p)^2}\ dz_d^{(n+1)}\\
=&e^q\ \sqrt{\frac{\alpha_1}{2}-\beta}\ \sqrt{\pi},
\end{aligned}$$

falls $\beta < \frac{\alpha_1}{2}$. Sukzessive lassen sich nun sämtliche Integrale von innen nach außen errechnen, sodass die Aussage des Satzes als bewiesen betrachtet werden kann. □

3.4 Nachweis der geometrischen Ergodizität mittels Driftbedingung

Mit den getroffenen Vorbereitungen sind wir nun in der Lage, das Hauptergebnis dieses Kapitels zu beweisen.

Satz 3.10 *Sei* $S_q = \inf\{n \geq 1; |Z^{(n)}| \leq q\}$ *und* $q \in \mathbb{R}$ *groß genug gewählt, dann gibt es ein* $\beta > 0$*, derart dass*

$$\sup_{|Z| \leq q} \mathbb{E}_Z[e^{\beta S_q}] < \infty.$$

Dies impliziert die geometrische Rekurrenz der Markov-Kette. Folglich konvergiert die Kette exponentiell gegen ihre Gleichgewichtsverteilung.

Beweis: Unsere Bemühungen werden zunächst darauf abzielen, eine Funktion g zu finden, die der Driftbedingung aus Satz 1.20 genügt. In diesem Bestreben zeigen wir zunächst, dass $e^{\lambda \hat{V}(Z^{(n)})+n}$ vor dem Zeitpunkt S_q die Supermartingal-Eigenschaft besitzt, unter der Prämisse, dass λ klein genug gewählt worden ist. Die Funktion $g(z) := e^{\lambda \hat{V}(z)}$ wird sich für unser Vorhaben als geeignet erweisen.

Gemäß der Taylorschen Formel für Funktionen mehrerer Veränderlicher gilt an der Entwicklungsstelle $U = (\phi_1(\Psi(Z^{(n)}), \Psi(Z^{(n)}))$:

$$\begin{aligned}
\hat{V}(Z^{(n+1)}) &= V(\phi_1(Z^{(n+1)}), Z^{(n+1)}) \\
&= V(U) + \sum_{j=2}^{d} \frac{\partial V}{\partial x_j}(U)\eta_j^{(n)} + \frac{1}{2}\langle(\xi_1^{(n)}, \eta^{(n)}), A(\xi_1^{(n)}, \eta^{(n)})\rangle \\
&= \hat{V}(\Psi(Z^{(n)})) + \langle \nabla V(U), (\xi_1^{(n)}, \eta^{(n)})\rangle + \frac{1}{2}\langle(\xi_1^{(n)}, \eta^{(n)}), A(\xi_1^{(n)}, \eta^{(n)})\rangle.
\end{aligned}$$

Dabei bezeichnet A per definitionem die Hesse-Matrix von V, ausgewertet an einer Zwischenstelle $U + \theta(\xi_1^{(n)}, \eta^{(n)})$, $\theta \in [0,1]$. Ferner beachte man, dass $\frac{\partial V}{\partial x_1}(U) = 0$ (Extremum in der 1. Komponente). Diese Formel nutzen wir für die Berechnung der faktorisierten bedingten Erwartung

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda \hat{V}(Z^{(n+1)})} \,\middle|\, Z^{(n)} = (z_2^{(n)}, \ldots, z_d^{(n)})\right] \\
=&\int e^{\lambda \hat{V}(Z^{(n+1)})} P^d(X^{(n)}, X^{(n+1,i)})\, dx_1^{(n+1)}\, dz_2^{(n+1)} \ldots\, dz_d^{(n+1)} \\
=&e^{\lambda \hat{V}(\Psi(Z^{(n)}))} \int e^{\lambda \langle \nabla V(U), (\xi_1^{(n)}, \eta^{(n)})\rangle}\, e^{\frac{1}{2}\langle(\xi_1^{(n)}, \eta^{(n)}), A(\xi_1^{(n)}, \eta^{(n)})\rangle} \ldots \\
&\ldots\, P^d(X^{(n)}, X^{(n+1,i)})\, dx_1^{(n+1)}\, dz_2^{(n+1)} \ldots\, dz_d^{(n+1)}.
\end{aligned}$$

Wir möchten zu einer Abschätzung gelangen, welche von der 1. Komponente unabhängig ist. Daher halten wir diese Komponente für einen Augenblick fest und weisen ihr

den Wert $\phi_1(\Psi(Z^{(n)})$ zu. Schätzen wir zudem die quadratische Form im Exponenten des Integranden mit den üblichen Matrix- bzw. Vektornormen ab, heißt das

$$\lambda\frac{1}{2}\langle(0,\eta^{(n)}),A(0,\eta^{(n)})\rangle \leq \lambda\frac{1}{2}\|A\|\ |\eta^{(n)}|^2,$$

und wählen wir λ so klein, dass $\lambda\frac{1}{2}\|A\| < \beta$, so definieren wir $\alpha := \beta - \lambda\frac{1}{2}\|A\|$ und erhalten

$$\begin{aligned}
&\mathbb{E}\Big[e^{\lambda\hat{V}(Z^{(n+1)})}\ \Big|\ X_1^{(n+1)} = \phi_1(\Psi(Z^{(n)})\ ;\ Z^{(n)} = (z_2^{(n)},\dots,z_d^{(n)})\Big]\\
\leq& e^{\lambda\hat{V}(\Psi(Z^{(n)}))}\int e^{\lambda\langle\nabla V(U),(0,\eta^{(n)})\rangle}\ e^{-\alpha|\eta^{(n)}|^2}\ \dots\\
&\dots\ e^{\beta|\eta^{(n)}|^2}\prod_{i=2}^{d}P_i\left(z_i^{(n+1)}\ \Big|\ \phi_1(\Psi(Z^{(n)}),z_2^{(n+1)},\dots,z_{i-1}^{(n+1)},z_{i+1}^{(n)},\dots,z_d^{(n)}\right)\\
&\dots\ dz_2^{(n+1)}\dots\ dz_d^{(n+1)}\\
\leq& c\ e^{\lambda\hat{V}(\Psi(Z^{(n)}))}\int e^{\lambda\langle\nabla V(U),(0,\eta^{(n)})\rangle}\ e^{-\alpha|\eta^{(n)}|^2}\ dz_2^{(n+1)}\dots\ dz_d^{(n+1)},
\end{aligned}$$

wie sich nun leicht unter Berufung auf Satz 3.9 und unter Hinzunahme der Hölder-Ungleichung im ausgearteten Fall ergibt. Offensichtlich hängt dieses Resultat nicht mehr von seiner 1. Komponente $\phi_1(\Psi(Z^{(n)})$ ab. Substituieren wir z durch η und verzichten auf den Zeitindex n, so erhalten wir:

$$\begin{aligned}
&\mathbb{E}\Big[e^{\lambda\hat{V}(Z^{(n+1)})}\ \Big|\ Z^{(n)} = (z_2^{(n)},\dots,z_d^{(n)})\Big]\\
&\leq c\ e^{\lambda\hat{V}(\Psi(Z^{(n)}))}\int e^{-\alpha|\eta|^2+\lambda\sum_{j=2}^{d}\frac{\partial V}{\partial x_j}(U)\eta_j}\ d\eta_2\dots\ d\eta_d.
\end{aligned}\tag{3.27}$$

Die 1. Komponente des Gradienten von V verschwindet an der Stelle U (lokales Extremum). Bezeichnen wir deshalb den um seine 1. Komponente gestutzten, $\mathbb{R}^{d-1}$-wertigen Vektor wiederum mit ∇V, so lässt sich der Exponent des Integranden mittels Cauchy-Schwarz-Ungleichung abschätzen:

$$\begin{aligned}
-\alpha|\eta|^2+\lambda\sum_{j=2}^{d}\frac{\partial V}{\partial x_j}(U)\eta_j &= -\alpha|\eta|^2+\lambda\langle\nabla V(U),\eta\rangle\\
&\leq -\alpha|\eta|^2+\lambda|\nabla V(U)|\ |\eta|.
\end{aligned}$$

Mit einer quadratischen Ergänzung erhält man für die letzte Zeile auch

$$-\alpha|\eta|^2+\lambda|\nabla V(U)|\ |\eta| = -\alpha\left(|\eta|-\frac{\lambda}{2\alpha}|\nabla V(U)|\right)^2+\frac{\lambda^2}{4\alpha}|\nabla V(U)|^2.$$

Kehren wir zurück zu Formel (3.27). Es gilt

$$\begin{aligned}
&\mathbb{E}\Big[e^{\lambda\hat{V}(Z^{(n+1)})}\ \Big|\ Z^{(n)} = (z_2^{(n)},\dots,z_d^{(n)})\Big]\\
&\leq c\ e^{\lambda\hat{V}(\Psi(Z^{(n)}))}\int e^{-\alpha\left(|\eta|-\frac{\lambda}{2\alpha}|\nabla V(U)|\right)^2+\frac{\lambda^2}{4\alpha}|\nabla V(U)|^2}\ d\eta_2\dots\ d\eta_d\\
&= c\ e^{\lambda\hat{V}(\Psi(Z^{(n)}))}e^{\frac{\lambda^2}{4\alpha}|\nabla V(U)|^2}\int e^{-\alpha\left(|\eta|-\frac{\lambda}{2\alpha}|\nabla V(U)|\right)^2}\ d\eta_2\dots\ d\eta_d.
\end{aligned}$$

Das Mehrfachintegral in der letzten Zeile lässt sich mittels eines Satzes über rotationssymmetrische Funktionen auf ein eindimensionales Integral zurückführen (siehe hierzu § 8 in [For]). Insbesondere ist die Berechenbarkeit damit sichergestellt. Sein Wert kann mit der Konstanten c zu einer neuen Konstanten c_1 zusammengefasst werden:

$$\mathbb{E}[e^{\lambda\hat{V}(Z^{(n+1)})}|\mathcal{F}_n^Z] \leq c_1 e^{\lambda\hat{V}(\Psi(Z^{(n)}))} e^{\frac{\lambda^2}{4\alpha}|\nabla V(U)|^2}. \tag{3.28}$$

Betrachten wir den Exponenten der rechten Seite noch einmal genauer. Mit Hilfe der Lemmata 3.1 und 3.2 kann eine Konstante c_2 definiert werden, mittels derer der Gradient von V aus eben diesem Exponenten eliminiert werden kann. Es gilt

$$|\nabla V(U)|^2 \leq \frac{2\alpha_2^2}{\alpha_1} V(U) \leq \frac{2\alpha_2^2}{\alpha_1}\hat{V}(\Psi(Z^{(n)})).$$

Setze $c_2 := \frac{2\alpha_2^2}{\alpha_1}\ \frac{1}{4\alpha} = \frac{\alpha_2^2}{2\alpha_1\alpha}$ und erhalte mit Lemma 3.5

$$\begin{aligned}
\lambda\hat{V}(\Psi(Z^{(n)})) + \frac{\lambda^2}{4\alpha}|\nabla V(U)|^2 &\leq \lambda(1+c_2\lambda)\hat{V}(\Psi(Z^{(n)})) \\
&\leq \lambda(1+c_2\lambda)\gamma\hat{V}(Z^{(n)}) \\
&= \lambda\hat{V}(Z^{(n)}) - \underbrace{\lambda(1-\underbrace{(1+c_2\lambda)\gamma}_{a(\lambda)})\hat{V}(Z^{(n)})}_{b(\hat{V}(Z^{(n)}))}.
\end{aligned}$$

Wähle nun λ so klein bzw. q so groß, dass

$$a(\lambda) \leq 1 \quad \text{bzw.} \quad b(\hat{V}(Z^{(n)})) > 1 + \ln c_1.$$

Damit erhalten wir

$$\begin{aligned}
\mathbb{E}[e^{\lambda\hat{V}(Z^{(n+1)})}|\mathcal{F}_n^Z] &\leq c_1 e^{\lambda\hat{V}(Z^{(n)})-(1+\ln c_1)} \\
&= e^{\lambda\hat{V}(Z^{(n)})}\ e^{-1}.
\end{aligned}$$

Multiplikation von e^n auf beiden Seiten liefert dann die gewünschte Martingaleigenschaft

$$\mathbb{E}[e^{\lambda\hat{V}(Z^{(n+1)})+(n+1)}|\mathcal{F}_n^Z] \leq e^{\lambda\hat{V}(Z^{(n)})+n}. \tag{3.29}$$

Damit wiederum gilt nun auch

$$\mathbb{E}[e^{\lambda\hat{V}(Z^{(S_q)})+S_q}|\mathcal{F}_1^Z] \leq e^{\lambda\hat{V}(Z^{(1)})+1},$$

wenn $|Z^{(1)}| > q$. Die geometrische Rekurrenz erhalten wir mit der Driftbedingung 1.20. Setzen wir nämlich in Ungleichung (3.29) für $z \in \mathbb{R}^{d-1}$ die Funktion $g(z) := e^{\lambda\hat{V}(z)}$, $C := \{|z| \leq q\}$ und $n = 0$, so haben wir lediglich Formel (1.29) zu verifizieren. Den ersten, schwierigen Teil der Driftbedingung (1.28) haben wir mit Ungleichung 3.29 bereits bestätigt ($n = 0$ und $r = e$). Somit ist lediglich noch zu zeigen, dass

$$\sup_{z\in C}\mathbb{E}[g(Z^{(1)}); |Z^{(1)}| > q | Z^{(0)} = z] = \sup_{z\in C} PI_{C^c}g(z) < \infty. \tag{3.30}$$

Es gilt für alle $z \in C$:

$$\begin{aligned}&\mathbb{E}[g(Z^{(1)}); |Z^{(1)}| > q | Z^{(0)} = z] \\ \leq &\mathbb{E}[g(Z^{(1)}) | Z^{(0)} = z] \\ \leq &c_1 e^{\lambda \hat{V}(\Psi(Z^{(0)}))} e^{\frac{\lambda^2}{4\alpha} |\nabla V(U(Z^{(0)})|^2},\end{aligned}$$

wie wir bereits in Gleichung 3.28 festgestellt hatten. Zur Erinnerung: $U(Z^{(n)}) = (\phi_1(\Psi(Z^{(n)}), \Psi(Z^{(n)}))$. Die Lemmata 3.5 und 3.3 erlauben zusammen mit der Voraussetzung $|Z^{(0)}| < q$ die Abschätzung

$$\hat{V}(\Psi(Z^{(0)})) \leq \gamma \hat{V}(Z^{(0)}) \leq \frac{\gamma}{2} \alpha_2 |Z^{(0)}|^2 \leq \frac{\gamma}{2} \alpha_2 q^2 < \infty,$$

wobei $\gamma < 1$ gemäß Lemma 3.5 gewählt wird. $\nabla V(U(Z^{(0)})|$ kann anschließend mittels Lemma 3.1 gegen $\alpha_2 |Z^{(0)}| \leq \alpha_2 q < \infty$ abgeschätzt werden, sodass Bedingung 3.30 offenbar erfüllt ist. □

Die Konvergenz im Ergodensatz 1.17 erfolgt gleichmäßig über alle $A \in \mathfrak{S}$. Welche Aussage lässt sich aber bezüglich der Anfangsverteilung $\delta_x \in \mathfrak{W}$ treffen? Dürfen wir auf die besonders starke, gleichmäßige exponentielle Ergodizität aus Definition 1.19 hoffen? Leider werden wir enttäuscht, aber es gibt dennoch eine Verschärfung: Die Funktion $M \in L^1(\pi)$ mittels derer der Abstand der Übergangskerne $P^n(x,A)$ von der stationären Verteilung $\pi(A)$ in Totalvariation abgeschätzt wird, kann nämlich durch eine reellwertige Funktion ersetzt werden.

Lemma 3.11 *Die Markov-Kette aus Satz 3.10 ist geometrisch ergodisch mit beliebigem Anfangszustand, das heißt, die Funktion $M(x)$ aus Theorem 1.17 ist reell.*

Beweis: Es gilt mit Satz 1.17:

$$\| P^n(X,\cdot) - \mu(\cdot) \| \leq M(X) \rho^n,$$

für $M(\cdot) \in L^1(\mu),\ 0 < \rho < 1$. Eine weitere Anwendung von P liefert:

$$\begin{aligned}\| P^{n+1}(X,\cdot) - \mu(\cdot) \| &\leq \int \| P^n(Y,\cdot) - \mu(\cdot) \| \ P(X,dY) \\ &\leq \rho^n \int M(Y) \ P(X,dY).\end{aligned}$$

Für die Übergangsdichte $P(X,Y)$ gilt:

$$P(X,Y) \leq c^d e^{-\sum_{k=1}^d (V(Y^{(k)}) - V(\tilde{Y}^{(k)}))},$$

wobei

$$\begin{aligned}Y^{(k)} &= (y_1, \ \dots \ , y_k, x_{k+1}, \ \dots \ , x_d), \\ \tilde{Y}^{(k)} &= (y_1, \ \dots \ , y_{k-1}, \phi_k(y_1, \ \dots \ , y_{k-1}, x_{k+1}, \ \dots \ , x_d), x_{k+1}, \ \dots \ , x_d).\end{aligned}$$

Per Definition von V gilt

$$V(Y^{(k-1)}) \geq V(\tilde{Y}^{(k)}), \quad k = 2, \ldots, d,$$

sodass

$$\begin{aligned} P(X,Y) &\leq c^d e^{-(V(Y)-V(\tilde{Y}^{(1)}))} \\ &= c^d e^{\hat{V}(x_2, \ldots, x_d)} \; e^{-V(Y)} \end{aligned}$$

folgt. Damit erhalten wir insgesamt

$$\| P^{n+1}(X,\cdot) - \mu(\cdot) \| \leq c^d e^{\hat{V}(x_2, \ldots, x_d)} \int M(Y) \; \mu(dY) \; \rho^n.$$

Die Funktion

$$M^*(x) := c^d e^{\hat{V}(x_2, \ldots, x_d)} \int M(Y) \; \mu(dY)$$

hat die gewünschte Eigenschaft. □

4 Bestimmung der Konvergenzrate des Gibbs Samplers

4.1 Die Pearsonsche-χ^2-Abstandsfunktion und der Hilbertraum $H_0^2(\pi)$

Im vorangegangenen Kapitel haben wir die geometrische Konvergenz des Gibbs-Samplers unter restriktiven Bedingungen an die Zielverteilung π nachgewiesen. Wesentlich schwieriger ist es, die Konvergenzrate ρ^* explizit zu bestimmen. In diesem Kapitel werden wir nun ein Konzept vorstellen, das es erlaubt, wiederum unter einschränkenden Bedingungen, die Konvergenzrate als Spektralradius eines mit der Gibbs-Sequenz korrespondierenden Operators zu identifizieren.

Wenn wir von Konvergenz sprechen, benötigen wir stets eine Metrik oder Norm, bezüglich derer sich eine Folge ihrem Limes nähert. Bisher haben wir stillschweigend die Totalvariationsnorm verwendet. In den nachfolgenden Betrachtungen legen wir ein anderes Maß zur Abstandsmessung zweier Verteilungen zugrunde, die χ^2-Abstandsfunktion von Pearson (χ^2-Funktion). Hierzu müssen die Dichten der Verteilungen vorliegen:

Definition 4.1 *Seien $\pi(x)$ und $p(x)$ die Dichten zweier Verteilungen auf $(\mathbb{R}^d, \mathfrak{B}^d)$ und $\pi(x) > 0$. Falls $\int \frac{p^2(x)}{\pi(x)}\, dx$ existiert, ist die χ^2-Funktion von $p(x)$ und $\pi(x)$ definiert als*

$$d_\pi^2(\pi, p) = \int \frac{p^2(x)}{\pi(x)}\, dx - 1. \tag{4.1}$$

Bemerkung: Natürlich ist die χ^2-Funktion keine echte Metrik, bereits die erforderliche Symmetrie ist verletzt. Dennoch eignet sie sich für unsere Zwecke. Unter Beachtung der Cauchy-Schwartz-Ungleichung lässt sich etwa folgende Beziehung zwischen L_1-Norm und χ^2-Funktion herstellen:

$$\| p - \pi \|_{L_1} \leq \int \left| \frac{p(x) - \pi(x)}{\sqrt{\pi(x)}} \right| \sqrt{\pi(x)}\, dx \leq \left(\int \frac{(p(x) - \pi(x))^2}{\pi(x)}\, dx \right)^{\frac{1}{2}} = d_\pi(\pi, p).$$

Wenn die χ^2-Funktion von p und π einen kleinen Wert annimmt, so gilt dies ebenso für die Totalvariationsnorm. In diesem Sinn ist die Konvergenz bezüglich der χ^2-Funktion stärker als jene bezüglich der Totalvariationsnorm (vgl. hierzu auch Remark 1 auf Seite 162 in [Li/Wo/Ko1]).

Als triviale Folgerung aus Definition 4.1 halten wir fest, dass die χ^2-Funktion stets nichtnegativ ist. Nimmt sie den Wert 0 an, so ist dies bereits äquivalent zur Gleichheit der Dichten π und p. Außerdem gilt die prägnante Formel:

$$d_\pi^2(\pi, p) = \mathrm{Var}_\pi\left(\frac{p(X)}{\pi(X)}\right).$$

Der Gibbs Sampler mit systematischer Abtastung erzeugt eine Markov-Kette, deren Übergangsdichte von $X^{(n)}$ nach $X^{(n+1)}$ folgende Gestalt besitzt:

$$P(X^{(n)}, X^{(n+1)}) = \prod_1^d \pi(x_i^{(n+1)} | x_1^{(n+1)}, x_2^{(n+1)}, \ldots, x_{i-1}^{(n+1)}, x_{i+1}^{(n)}, \ldots x_d^{(n)}). \tag{4.2}$$

Wenn $X^{(0)}$ die Anfangsverteilung $p_0(x)$ besitzt, erhalten wir die Dichte $p_n(x)$, welche die Verteilung von $X^{(n)}$ beschreibt, per Rekursion:

$$p_n(x) = \int P(y,x) p_{n-1}(y)\, dy, \quad n = 1, 2, \ldots. \tag{4.3}$$

Die Gibbs-Sequenz konvergiert bezüglich der χ^2-Funktion gegen eine Zielverteilung π, wenn

$$\lim_{n\to\infty} d_\pi(\pi, p_n) = 0.$$

Die Anfangsverteilung p_0 muss hierfür natürlich so gewählt werden, dass die χ^2-Funktion $d_\pi(\pi, p_n)$ nach endlich vielen Schritten endliche Werte annimmt. Analog zu Definition 1.26 definieren wir:

Definition 4.2 *Die kleinste Konstante $\rho^* \in [0,1)$, sodass*

$$\lim_{n\to\infty} \rho^{-n} d_\pi(\pi, p_n) = 0 \quad \textit{für alle } \rho > \rho^*$$

nennen wir Konvergenzrate der Gibbs-Sequenz bezüglich der χ^2-Funktion.

Die Gibbs-Sequenz $X^{(0)}, X^{(1)}, \ldots$ stellt bekanntlich eine zeithomogene Markov-Kette mit stationärer Verteilung $\pi(x)$ und Übergangsdichte $P(x,y)$ dar.

Betrachten wir nun den Hilbertraum quadratisch integrierbarer, komplexwertiger Funktionen $t : \mathbb{R}^d \to \mathbb{C}$, deren Mittelwert verschwindet:

$$H_0^2(\pi) = \{t(x) | E_\pi(t(X)) = 0, \quad E_\pi(|t(X)|^2) < +\infty\}$$

mit dem Skalarprodukt

$$< t(x), s(x) >= E_\pi(t(X)\overline{s(X)},$$

hierbei sei $|t(X)|$ der Betrag einer komplexen Zahl und $\overline{s(X)}$ die komplex konjugierte Abbildung der Funktion $s(X)$. Die Varianz der komplexen Zufallsvariablen $t(X)$ ist schließlich definiert durch $\mathrm{Var}_\pi(t(X)) =< t(x), t(x) >=\| t(x) \|^2$, was nichts anderes als

die Summe der Varianzen von Real- und Imaginärteil der Funktion t ist. Die Norm eines Operators A auf $H_0^2(\pi)$ wird folgendermaßen erklärt:

$$\| A \| = \sup_{t \in H_0^2(\pi),\ \|t\|=1} \| At \| = \sup_{t \in H_0^2(\pi)} \frac{\| At \|}{\| t \|}.$$

Die Spektraltheorie beschäftigt sich mit der Frage, ob komplexe Werte $\lambda \in \mathbb{C}$ existieren, sodass für einen linearen Operator A aus der Menge $L(H)$ der linearen Operatoren auf einem Hilbertraum H, der Operator $\lambda Id - A$ injektiv ist. Sie verallgemeinert die Eigenwerttheorie für Matrizen auf Operatoren auf unendlichdimensionalen Banachräumen.

Definition 4.3 *Sei $A \in L(H)$. Die Resolventenmenge von A ist*

$$\rho(A) = \{\lambda \in \mathbb{C} : (\lambda Id - A)^{-1} \text{ existiert in } L(H)\}.$$

Das Spektrum von A ist dann $\sigma(A) = \mathbb{C} \backslash \rho(A)$.

Für beschränkte Operatoren, solche mit endlicher Norm, existiert der Limes

$$\lim_{n \to \infty} \| A^n \|^{\frac{1}{n}} = r(A) \tag{4.4}$$

und wird Spektralradius von A genannt. Es gilt stets die Beziehung

$$r(A) \leq \| A \| . \tag{4.5}$$

Mehr noch, für einen beschränkten linearen Operator $A \in L(H)$ haben wir

$$r(A) = \sup_{\lambda \in \sigma(A)} |\lambda|, \tag{4.6}$$

was nichts anderes bedeutet, als dass der größte Eigenwert des Operators A betraglich nicht größer als sein Spektralradius sein kann (vgl. Abschnitt VIII.2, Theorem 3 und 4 in [Yos]). Existiert zu einem Operator A ein weiterer Operator A^*, sodass die Beziehung

$$< At_1, t_2 > = < t_1, A^* t_2 >$$

für alle $t_1, t_2 \in H$ erfüllt ist, so bezeichnen wir A^* als den zu A adjungierten Operator. Im Fall $A = A^*$ nennen wir A selbstadjungiert. Letztere Eigenschaft beschert uns sofort eine Verschärfung der Aussage aus Gleichung (4.5), nämlich

Lemma 4.4 *Ist A ein beschränkter und selbstadjungierter Operator auf $L(H)$, dann ist*

$$r(A) = \| A \| .$$

Beweis: Ein selbstadjungierter Operator A ist offensichtlich normal, das heißt er erfüllt die Beziehung $AA^* = A^*A$. Satz VI.1.7 in [Wer] liefert damit das gewünschte Resultat. □

Ein linearer Operator $A \in L(H)$ heißt kompakt, wenn er die Einheitskugel $B = \{t \in H : \| t \| \leq 1\}$ so abbildet, dass $A(B)$ relativkompakt ist. Mit dieser Eigenschaft ergibt sich folgendes Lemma:

Lemma 4.5 *Ist A kompakt, beschränkt und selbstadjungiert, so sind die Eigenwerte von A reell und die korrespondierenden Eigenvektoren existieren.*

Beweis: Der Beweis ergibt sich unmittelbar aus dem Spektralsatz für kompakte und normale Operatoren (Theorem VI.3.2 in [Wer] in Kombination mit Lemma VI.3.1 ebendort). □

Mit diesem funktionalanalytischen Rüstzeug sind wir ausreichend präpariert für unser Vorhaben, die Konvergenzrate der Gibbs-Sequenz als Spektralradius eines linearen Operators auf dem Hilbertraum $H_0^2(\pi)$ zu identifizieren.

Definition 4.6 *Für $t(x) \in H_0^2(\pi)$ definieren wir den Vorwärts-Operator F (forward) und den Rückwärts-Operator B (backward):*

$$F(t(x)) = E(t(X^{(1)})|X^{(0)} = x) = \int t(y)P(x,y)\, dy.$$
$$B(t(y)) = E(t(X^{(0)})|X^{(1)} = y) = \int t(x)\pi(x)\frac{P(x,y)}{\pi(y)}\, dx.$$

Es lässt sich elementar nachrechnen, dass F und B zueinander adjungierte Operatoren auf dem Hilbertraum $H_0^2(\pi)$ sind. Ebenso ist einzusehen, dass ihre Normen durch den Wert 1 beschränkt werden, das heißt $\| F \| = \| B \| \leq 1$. Legen wir einen endlichen Zustandsraum zu Grunde, so kann man F als Übergangsmatrix einer Markov-Kette identifizieren. B ist dann einfach ihre konjugiert Transponierte. Mit den Kolmogorov-Chapman-Gleichungen erhält man

$$F^n(t(x)) = E(t(X^{(n)})|X^{(0)} = x) \quad \text{und} \quad B^n(t(y)) = E(t(X^{(0)})|X^{(n)} = y).$$

Bemerkung: Wir könnnen unsere Betrachtungen natürlich auch auf den Raum $H^2(\pi) = \{t(x)|\ E_\pi(|t(X)|^2) < +\infty\}$ beziehen, das heißt, wir lassen auch Funktionen t zu, deren Mittelwert ungleich 0 ist. $H_0^2(\pi)$ ist ein Teilraum von $H^2(\pi)$. Die Operatoren F und B existieren natürlich auch auf $H^2(\pi)$. Wenn wir uns aber auf $H_0^2(\pi)$ zurückziehen, eliminieren wir die konstante Funktion als Eigenwert des Vorwärts-Operators F. Der größte Eigenwert von F, eingeschränkt auf $H_0^2(\pi)$, entspricht nämlich gerade dem zweitgrößten Eigenwert von F, sofern wir $H^2(\pi)$ zu Grunde legen (vgl. die Bemerkung auf Seite 30 in [Li/Wo/Ko2]).

4.2 Die Konvergenzrate ρ^* als Spektralradius des Vorwärts-Operators F

Bevor wir nun zeigen, dass die Konvergenzrate ρ^* gerade dem Spektralradius $r(F)$ des Vorwärts-Operators F entspricht, haben wir drei Bedingungen und zwei Lemmata zu formulieren, die in dem Beweis des Resultats zum Tragen kommen:

1. Für die Übergangsdichte $P(x,y)$ des Gibbs-Sampler mit Zielverteilung π gilt:

$$\int \left(\frac{P(x,y)}{\pi(y)}\right)^2 \pi(y)\pi(x)\, dxdy < +\infty. \tag{4.7}$$

2. In $H_0^2(\pi)$ existiert keine nicht-konstante Funktion $t(x)$, die der Gleichung

$$E_\pi(t(X)|X_1, X_2, \ldots, X_{(i-1)}, X_{(i+1)}, \ldots, X_d) = t(X) \quad \pi\text{-f.s.} \tag{4.8}$$

   für alle i genügt.

3. Die detaillierte Gleichgewichtsgleichung ist erfüllt, das heißt:

$$P(x,y)\pi(x) = P(y,x)\pi(y) \quad \text{für alle } x,y \in \mathbb{R}^d. \tag{4.9}$$

   Letzteres impliziert die Reversibilität der Markov-Kette.

Diese Bedingungen, insbesondere die ersten beiden, bedürfen einer kurzen Erklärung. Die 1. Bedingung stellt die Endlichkeit der χ^2-Funktion von $p_n(x)$ und π nach endlich vielen Iterationsschritten sicher. Außerdem liefert sie zusammen mit der 2. Bedingung die π-Irreduzibilität der Markov-Kette. Eine stärkere Bedingung stellt die Positivitätsbedingung aus Definition 2.6 dar (vgl. die Ausführungen auf Seite 160 und 161 in [Li/Wo/Ko1]). Nun zu den bereits angekündigten Lemmata.

Lemma 4.7 *Bedingung 1 impliziert die Kompaktheit des Vorwärts-Operators F*

Beweis: Für beliebiges $t(x) \in H_0^2(\pi)$ erweist sich

$$Ft(x) = \int t(y)P(x,y)\, dy = \int t(y)\pi(y)\frac{P(x,y)}{\pi(y)}\, dy$$

als Integraloperator mit Kern $G(x,y) = P(x,y)/\pi(y)$. Bedingung 2 liefert:

$$\int P^2(x,y)\pi(y)\pi(x)\, dxdy < +\infty \tag{4.10}$$

und damit unter Beachtung von example 2 des Abschnitts X.2 in [Yos] die gewünschte Kompaktheit des Operators F. □

Dieses Resultat können wir sofort im nächsten Lemma aufgreifen.

Lemma 4.8 *Die Bedingungen 1 und 2 implizieren, dass der Spektralradius des Vorwärts-Operators F echt kleiner als* 1 *ist.*

Mit anderen Worten: Der größte Eigenwert von F ist kleiner als 1. Die Aussage entspricht Lemma 2 in [Li/Wo/Ko1].

Beweis: Wie wir bereits festgestellt haben, ist die Norm und damit auch der Spektralradius von F nicht größer als 1 (siehe Gleichung (4.5)). Nach Lemma 4.7 ist F kompakt.

Daher ist das Spektrum abzählbar mit 0 als einzigem möglichen Akkumulationspunkt. Folglich existiert zum betraglich größten Eigenwert λ_1 eine Eigenfunktion t. Angenommen $|\lambda_1| = 1$, dann gilt:

$$F(tX) = \lambda_1 t(X) \quad \pi\text{-fast sicher.}$$

Dies impliziert natürlich Gleichheit für die Varianzen

$$\mathrm{Var}(Ft(X)) = \mathrm{Var}(t(X)). \tag{4.11}$$

Mit der Bezeichnung $X^{-[i]} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$ und der Gestalt von $P(X,Y)$ aus Gleichung (4.2) erhalten wir für $Ft(X)$ auch:

$$Ft(X) = E(\ldots(E(E(t(X)|X^{-[1]})|X^{-[2]})\ldots)|X^{-[d]}).$$

Des Weiteren besteht zwischen zwei komplexwertigen Zufallsvariablen W und V die Beziehung

$$\mathrm{Var}(W) = \mathrm{Var}(E(W|Z)) + \mathrm{Var}(W - E(W|Z)), \tag{4.12}$$

welche sich genauso beweisen lässt, wie ihr reelles Pendant. Es ergeben sich folgende Ungleichungen:

$$\begin{aligned}\mathrm{Var}(t(X)) &\geq \mathrm{Var}(E(t(X)|X^{-[1]})) \geq \ldots \\ &\geq \mathrm{Var}\left(E(\ldots(E(E(t(X)|X^{-[1]})|X^{-[2]})\ldots)|X^{-[d]})\right) = \mathrm{Var}(Ft(X)).\end{aligned}$$

Formel (4.11) erzwingt die Gleichheit sämtlicher Varianzen. Insbesondere erhält man aus $\mathrm{Var}(t(X)) \geq \mathrm{Var}(E(t(X)|X^{-[1]}))$ zusammen mit Gleichung (4.12) die Beziehung $t(X) = E(t(X)|X^{-[1]})$ π-fast sicher. Induktiv ergibt sich $t(X) = E(t(X)|X^{-[i]})$ π-fast sicher für alle i. Letzteres widerspricht Bedingung 2, sodass wir unsere Annahme verwerfen müssen. □

Nach all diesen Vorbereitungen formulieren wir nun das Hauptergebnis dieses Kapitels:

Satz 4.9 *Für einen Gibbs-Sampler, der die Bedingungen 1-3 erfüllt, gilt:*

(i) Es liegt geometrische Konvergenz gegen die Zielverteilung π bezüglich der χ^2-Funktion vor, das heißt

$$d_\pi(\pi, p_n) \leq r^n d_\pi(\pi, p_0) \text{ wenn } d_\pi(\pi, p_0) < \infty,$$

wobei r den Spektralradius $r(F)$ des Vorwärts-Operators F bezeichnet.

(ii) r ist die gesuchte Konvergenzrate des Gibbs-Samplers mit Zielverteilung π und r entspricht der Norm des Vorwärts-Operators, das heißt $r = \| F \|$.

Beweis: Der Vorwärts-Operator F aus Definition 4.6 ist offensichtlich durch den Wert 1 beschränkt und selbstadjungiert, letzteres folgt unmittelbar aus der detaillierten Gleichgewichtsgleichung (Bedingung 3). Damit liefert Lemma 4.4 die Beziehung $r(F) = \| F \| \leq 1$. Nach Lemma 4.7 ist F außerdem kompakt, sodass sämtliche Eigenwerte reellwertig sein müssen und die korrespondierenden Eigenfunktionen existieren.

Für den betraglich größsten Eigenwert gilt $|\lambda_1| \leq r(F) = \| F \| < 1$ (Gleichung (4.6) und Lemma 4.8).

Widmen wir uns zunächst der Aussage (i). Bekanntlich lässt sich zur Übergangsdichte P(x,y) die n-Schritt Übergangsdichte erklären durch:

$$P^{(n)}(x,y) = \int P(x',y)P^{(n-1)}(x,x')\,dx', \quad n = 2,3,\dots\ .$$

Wähle eine geeignete Anfangsverteilung, sodass $d_\pi(\pi, p_0) < \infty$. Setze $t_0(x) = \frac{p_0(x)}{\pi(x)} - 1$. Dann erhalten wir für alle $t(x) \in H_0^2(\pi)$

$$\begin{aligned} \left| E_{p_n}(t(X)) - \overbrace{E_\pi(t(X))}^{=0} \right| &= \left| \int t(y)P^{(n)}(x,y) \left(\frac{p_0(x)}{\pi(x)} - 1 \right) \pi(x)\,dxdy \right| \\ &= |< F^n t, t_0 >| \ \leq \| F^n \| \, \| t \| \, \| t_0 \| . \end{aligned} \tag{4.13}$$

Wenn wir $t(x) = \frac{p_n(x)}{\pi(x)} - 1$ setzen, ist

$$E_{p_n}(t(X)) = E_{p_n}\left(\frac{p_n(X)}{\pi(X)} - 1 \right) = d_\pi^2(\pi, p_n)$$

und

$$\| t(x) \| = \sqrt{\int \left(\frac{p_n(x)}{\pi(x)} - 1 \right)^2 \pi(x)\,dx} = d_\pi(\pi, p_n).$$

Entsprechend ist $t_0(x) = d_\pi(\pi, p_0)$. Mit Gleichung (4.13) führt dies zu

$$d_\pi(\pi, p_n) \leq \| F^n \| \ d_\pi(\pi, p_0).$$

Unter Beachtung der Definition des Spektralradius

$$r = \lim_{n\to\infty} \sqrt[n]{\| F^n \|}$$

folgt

$$d_\pi(\pi, p_n) \leq r^n d_\pi(\pi, p_0).$$

Lemma 4.8 liefert die geometrische Konvergenz bezüglich der χ^2-Funktion, das heißt $\lim_{n\to\infty} d_\pi(\pi, p_n) = 0$.

Im Folgenden ist nun zu zeigen, dass der Spektralradius tatsächlich die gesuchte Konvergenzrate des Gibbs Samplers unter der χ^2-Funktion ist. Sei $t^{**}(x)$ der Eigenvektor des größten Eigenwertes λ. Ohne Beschränkung der Allgemeinheit sei $t^{**}(x)$ eine reelle Funktion. Sei $t^*(x)$ nämlich ein beliebiger Eigenvektor des Eigenwerts λ, so definieren wir einfach

$$t^{**}(x) = \frac{t^*(x) + \overline{t^*(x)}}{2}.$$

$t^{**}(x)$ ist damit ein reeller Eigenvektor zum Eigenwert λ. Sicherlich ist $t^{**}(x) \in H_0^2(,\pi)$. Es sei bemerkt, dass durch die Bedingung an p_0 und $t^{**}(x) \in H_0^2(\pi)$, sowohl $\| t_0 \| < \infty$ als

auch $\| t^{**} \| < \infty$. Da $|\lambda| = r < 1$, befinden wir uns in einer Situation, in der für alle $\varepsilon > 0$ ein $m \in \mathbb{N}$ existiert, sodass

$$2r^m | < t^{**}, t_0 > | < \frac{\varepsilon}{2}, \qquad 2r^{2m} \| t^{**} \| < \frac{\varepsilon}{2}. \tag{4.14}$$

Definiere $\widetilde{t(x)} = \lambda^m t^{**}(x)$ und $p_0^*(x) = \pi(x)\widetilde{t(x)} + p(x)$. Dann ist

$$\int \frac{(\pi(x) - p_0^*(x))^2}{\pi(x)} \, dx = \lambda^{2m} \| t^{**}(x) \|^2 < \infty.$$

Auch wenn es sich hierbei nicht um eine Wahrscheinlichkeitsdichte handelt, definieren wir ferner analog zu Gleichung (4.3)

$$p_n^*(x) = \int P(y,x) p_{(n-1)}^*(y) \, dy,$$

Wir haben

$$p_1^*(y) = \int \widetilde{t(x)} P(x,y) \pi(x) \, dx + \pi(y) = \lambda \pi(y) \widetilde{t(y)} + \pi(y)$$

und bei wiederholter Anwendung

$$p_n^*(y) = \lambda^{n-1} \int \widetilde{t(x)} P(x,y) \pi(x) \, dx + \pi(y) = \lambda^n \pi(y) \widetilde{t(y)} + \pi(y).$$

Sei $t_0^*(x) = \frac{p_0^*(x)}{\pi(x)} - 1$. Ähnlich wie im Beweis von Gleichung (4.13), ergeben sich für alle $t(x) \in H_0^2(\pi)$ die Beziehung

$$\begin{aligned} &\left| \int t(x) p_n^*(x) \, dx - E_\pi((t(X)) \right| \\ =& \left| \int t(y) P^{(n)}(x,y) \left(\frac{p_0^*(x)}{\pi(x)} - 1 \right) \pi(x) \, dx dy \right| \\ =& |< F^n t, t_0^* >| . \end{aligned}$$

Ersetze $t(x)$ in der obigen Gleichung durch $\frac{p_n^*(x)}{\pi(x)} - 1$. Nutze ferner die Darstellung $t(x) = \lambda^n \widetilde{t(x)}$ und beachte, dass es sich bei F um einen selbstadjungierten Operator handelt. Es folgt:

$$\int \frac{(\pi(x) - p_n^*(x))^2}{\pi(x)} \, dx = |< t, F^n t_0^* >| = \left| < \lambda^n \widetilde{t(x)}, \lambda^n \widetilde{t(x)} > \right| = r^{2n} \| \widetilde{t(x)} \|^2 . \tag{4.15}$$

Andererseits haben wir mit $\int t^{**}(x) f(x) \, dx = 0$:

$$\begin{aligned} d_f^2(\pi, p_n) &= \int \frac{(p_n(x) - \pi(x))^2}{\pi(x)} \, dx \\ &= \int \frac{[(p_n(x) - p_n^*(x)) + (p_n^*(x) - \pi(x))]^2}{\pi(x)} \, dx \\ &= \int \frac{(p_n^*(x) - \pi(x))^2}{\pi(x)} \, dx + \int \frac{(p_n(x) - p_n^*(x))^2}{\pi(x)} \, dx \\ &\quad + \left[2\lambda^m \lambda^n \int t^{**}(x) p_n(x) \, dx - 2\lambda^{2m} \lambda^{2n} \int (t^{**}(x))^2 \pi(x) \, dx \right]. \end{aligned}$$

Mittels Gleichung (4.13) ergibt sich:

$$\left| \int t^{**}(x) p_n(x)\, dx \right| = r^n \,|< t^{**}, t_0 >|\,.$$

Gleichung (4.14) liefert:

$$\begin{aligned} & \left| \left[2\lambda^m \lambda^n \int t^{**}(x) p_n(x)\, dx - 2\lambda^{2m} \lambda^{2n} \int (t^{**}(x))^2 \pi(x)\, dx \right] \right| \\ < \quad & 2r^m \,|< t^{**}, t_0 >| + 2r^{2m} \parallel t^{**} \parallel^2 < \varepsilon. \end{aligned}$$

Da ε beliebig gewählt war, erhalten wir schlussendlich

$$d_\pi^2(\pi, p_n) \geq \int \frac{(p_n^*(x) - \pi(x))^2}{\pi(x)}\, dx.$$

Wenn nun r nicht die Konvergenzrate aus dem ersten Teil des Beweises dieses Satzes ist, existiert ein $r_1 < r$, sodass

$$\lim_{n\to\infty} r_1^{-n} d_\pi(\pi, p_n) = 0$$

und damit natürlich auch

$$\lim_{n\to\infty} r_1^{-2n} d_\pi^2(\pi, p_n) = 0.$$

Mit Gleichung (4.15) erhalten wir einen Widerspruch:

$$\begin{aligned} \lim_{n\to\infty} r_1^{-2n} d_\pi(\pi, p_n) &\geq \lim_{n\to\infty} r_1^{-2n} \int \frac{(p_n^*(x) - \pi(x))^2}{\pi(x)}\, dx \\ &= \lim_{n\to\infty} \Bigg(\underbrace{\frac{r}{r_1}}_{>0} \Bigg)^{2n} \lambda^m \parallel t^{**} \parallel = \infty. \end{aligned}$$

Folglich erweist sich r als die gesuchte Konvergenzrate. □

Es gibt Fälle, in denen eine oder gar mehrere Komponenten aus der d-dimensionalen Zielverteilung π durch Integration eliminiert werden können. Ist etwa eine Integration bezüglich der letzten Komponente d möglich, erhalten wir die Marginaldichte

$$\pi(x_-) = \int \pi(x_1, \ldots, x_d)\, dx_d$$

in Abhängigkeit von $x_- = (x_1, \ldots, x_{d-1}) \in \mathbb{R}^{d-1}$. Auch diese Verteilung lässt sich natürlich mit Hilfe des Gibbs Samplers simulieren. Den dazugehörigen Vorwärts-Operator bezeichnen wir mit F_c (c steht für „collapsed“ und soll an den um eine Dimension geschrumpften Raum $(\mathbb{R}^{d-1}, \mathfrak{B}^{d-1})$ erinnern, auf dem $\pi(x_-)$ definiert ist). Als einfache Folgerung aus Satz 4.9 halten wir fest:

Lemma 4.10 *Unter den Bedingungen aus Satz 4.9 liefert der Übergang von der Zielverteilung π zu ihrem reduzierten Pendant $\pi(x_-)$ für den Gibbs Sampler mit systematischer Abtastung eine bessere Konvergenzrate, das heißt*

$$\parallel F_c \parallel \leq \parallel F \parallel . \tag{4.16}$$

Beweis: Dass es sich bei $\| F_c \|$ und $\| F \|$ um die Konvergenzraten der entsprechenden Gibbs Sampler handelt, ist gerade der Aussage von Satz 4.9 (ii) geschuldet. Die Beziehung (4.16) beweist Liu mit Hilfe der maximalen Korrelation zwischen zwei Zufallsvariablen X und Y:

$$\gamma(X,Y) = \sup \mathrm{Cov}(t(X), s(Y)) = \sup \sqrt{\mathrm{Var}(E(t(X)|Y))},$$

wobei das Supremum über alle Funktionen $t(X)$ und $s(Y)$ mit $\mathrm{Var}(t(X)) = \mathrm{Var}(s(Y)) = 1$ gebildet wird. Für Details siehe Theorem 1 in [Liu] oder Theorem 5.1 in [Li/Wo/Ko2]. □

4.3 Anwendung auf eine normalverteilte Zielfunktion

Handelt es sich bei der Zielverteilung um eine Normalverteilung, können wir die Konvergenzrate des Gibbs Samplers als Spektralradius einer korrespondierenden Matrix darstellen.

Lemma 4.11 *Sei $X \sim \mathcal{N}_d(0,\Sigma)$ und Σ bezeichnet dabei eine positive $d \times d$-Matrix. Sei $Q = \Sigma^{-1} = (q_{ij})$. Mit $\mathcal{N}_d(0,\Sigma)$ als Zielverteilung lässt sich der Spektralradius des Vorwärts-Operators F, welcher sich aus der Übergangsfunktion gemäß Gleichung* (4.2) *ergibt, als Spektralradius der Matrix $A = \prod_{i=1}^{d}(I - D_i Q)$ identifizieren. (I bezeichnet die $(d \times d)$ Einheitsmatrix, D_i ist eine $(d \times d)$ Diagonalmatrix, in der die Elemente an der Position (i,i) den Wert q_{ii}^{-1} annehmen und die anderen Einträge verschwinden.)*

Beweis: Zum Beweis siehe [Ami]. □

Folgende Beziehung besteht zwischen der Matrix A und den Mittelwerten der univariaten bedingten Verteilungen:

$$E(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) = \sum_{j=1}^{d} A_{ij} x_j.$$

Korollar 4.12 *Sei F der Vorwärts-Operator, welcher durch die Übergangsfunktion aus* (4.2) *mit $\mathcal{N}_d(0,\Sigma)$ als Zielverteilung bestimmt ist. Dann ist die Konvergenzrate des Gibbs Samplers gleich dem Spektralradius der Matrix $H_L^{-1} H_U$ mit Σ und Q wie in Lemma 4.11, nämlich*

$$H_L = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1d} \\ 0 & q_{22} & \cdots & q_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_{dd} \end{bmatrix}, \tag{4.17}$$

und $H_U = H_L - Q$.

Der Beweis besteht im Wesentlichen aus elementaren Matrizenumformungen.

Beweis: Es kann gezeigt werden, dass die Übergangsfunktion, welche durch die Normalverteilung $\mathcal{N}_d(0,\Sigma)$ bestimmt ist, den Bedingungen in Satz 4.9 genügt. Nach Satz 4.9 und Lemma 4.11 entspricht die Konvergenzrate des Gibbs Samplers mit der Zielverteilung $\mathcal{N}_d(0,\Sigma)$ gerade dem Spektralradius der Matrix $\prod_{i=1}^{d}(I-D_iQ)$. Aus diesem Grunde haben wir im Folgenden nur noch $H_L^{-1}H_U = \prod_{i=1}^{d}(I-D_iQ)$ zu zeigen. Wir führen einen Induktionsbeweis. Der Gibbs Sampler liefert stets einen Zufallsvektor. Für $n=2$ gilt $H_L^{-1}H_U = \prod_{i=1}^{n}(I-D_iQ)$. Angenommen die Induktionsannahme gilt für $n=k-1$. Für $n=k$ sei

$$Q = \begin{bmatrix} Q_1 & h \\ h' & q_{dd} \end{bmatrix},$$

Q_1 ist eine $(k-1)\times(k-1)$-Matrix, h ein $(k-1)\times 1$-Vektor. H_{L1} und H_{U1} sind definiert wie in Gleichung (4.17) mit Matrix Q_1, das heißt

$$H_L = \begin{bmatrix} H_{L1} & h \\ 0 & q_{dd} \end{bmatrix}.$$

Folglich haben wir

$$H_L^{-1}H_U = I - H_L^{-1}Q = \begin{bmatrix} H_{L1}^{-1}H_{U1} + H_{L1}^{-1}hh'q_{dd}^{-1} & 0 \\ -q_{dd}^{-1}h' & 0 \end{bmatrix}.$$

Laut Induktionsannahme gilt $H_{L1}^{-1}H_{U1} = \prod_{i=1}^{k-1}(I-D_iQ_1)$ und damit

$$I - H_L^{-1}Q = \begin{bmatrix} \prod_{i=1}^{k-1}(I-D_iQ_1) & -H_{L1}^{-1}h \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & 0 \\ -q_{dd}^{-1}h' & 0 \end{bmatrix}.$$

Offensichtlich gilt

$$\begin{bmatrix} I & 0 \\ -q_{dd}^{-1}h' & 0 \end{bmatrix} = (I - D_dQ).$$

Beachte hierbei, dass die Einheitsmatrizen auf der linken und rechten Seite unterschiedliche Ordnungen aufweisen. Ebenso haben wir D_i mit unterschiedlichen Ordnungen verwandt. Nun haben wir nur noch zu zeigen, dass

$$\begin{bmatrix} \prod_{i=1}^{k-1}(I-D_iQ_1) & -H_{L1}^{-1}h \\ 0 & 1 \end{bmatrix} = \prod_{i=1}^{k-1}(I-D_iQ_1).$$

Sei u_i ein $(k-1)\times 1$-Vektor, das heißt

$$u_i = [0,\ldots,0,-\frac{q_{id}}{q_{ii}},0,\ldots,0]',$$

dann gilt:

$$\begin{aligned} &\begin{bmatrix} \prod_{i=1}^{k-1}(I-D_iQ_1) & -H_{L1}^{-1}h \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} (I-D_1Q_1) & u_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} (I-D_2Q_1) & u_2 \\ 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} (I-D_{(k-1)}Q_1) & u_{(k-1)} \\ 0 & 1 \end{bmatrix}. \end{aligned} \tag{4.18}$$

Die rechte Seite von Gleichung (4.18) kann geschrieben werden als

$$\begin{bmatrix} \prod_{i=1}^{k-1}(I-D_iQ_1) & T \\ 0 & 1 \end{bmatrix}$$

mit

$$T=\prod_{i=1}^{k-2}(I-D_iQ_1)u_{(k-1)}+\prod_{i=1}^{k-3}(I-D_iQ_1)u_{(k-2)}+\ldots+(I-D_iQ_1)u_2+u_1.$$

Um Gleichung (4.18) zu beweisen, haben wir zu zeigen, dass T die Lösung der Gleichung $H_{L1}x=-h$ ist, das heißt $H_{L1}T=-h$. Es ist offensichtlich, dass

$$H_{L1}u_1=\begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1(k-1)} \\ 0 & q_{22} & \cdots & q_{2(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & q_{(k-1)(k-1)} \end{bmatrix}\begin{bmatrix} -\frac{q_{1d}}{q_{11}} \\ 0 \\ \vdots \\ 0 \end{bmatrix}=\begin{bmatrix} -q_{1d} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$H_{L1}(I-D_1Q_1)u_2=\begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & q_{22} & \cdots & q_{2(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & q_{(k-1)(k-1)} \end{bmatrix}\begin{bmatrix} 0 \\ -\frac{q_{2d}}{q_{22}} \\ \vdots \\ 0 \end{bmatrix}=\begin{bmatrix} 0 \\ -q_{2d} \\ \vdots \\ 0 \end{bmatrix}.$$

So erhalten wir

$$H_{L1}\prod_{i=1}^{k-2}(I-D_iQ_1)u_{(k-1)}=[0,\ldots,0,-q_{(k-1)d}]',$$

$H_{L1}T=-h$. Mit Gleichung (4.18) und der Bemerkung, dass

$$\begin{bmatrix} (I-D_iQ_1) & u_i \\ 0 & 1 \end{bmatrix}=I-D_iQ,\quad i=1,2,\ldots,k-1,$$

haben wir das Resultat bewiesen. □

Theorem 1 in [Ro/Sa] ist gerade das obige Korollar als Spezialfall von Satz 4.9.

Literaturverzeichnis

[Ami] Amit, Y., *On Rates of Convergence of Stochastic Relaxation for Gaussian and Non-Gaussian Distributions*. Journal of Multivariate Analysis, 38, 82-99 (1991).

[Als1] Alsmeyer, G., *Wahrscheinlichkeitstheorie* (5. Auflage). Skripten zur Mathematischen Statistik, Nr. 30, Universität Münster (2007).

[Als2] Alsmeyer, G., *Stochastische Prozesse* (3. Auflage). Skripten zur Mathematischen Statistik, Nr. 33, Universität Münster (2005).

[Bes] Besag, J., *Spatial Interactions and the Statistical Analysis of Lattice Systems*. Journal of the Royal Statistical Society, Ser. B, 36, 192-236 (1974).

[Ca/Ge] Casella, G. und George, E., *Explaining the Gibbs Sampler*. The American Statistician 46, 167-174 (1992).

[For] Forster, O., *Analysis 3. Integralrechnung im $\mathbb{R}^n$ mit Anwendungen* (3.Auflage). Vieweg (1999).

[Gel/Sm] Gelfand, A. E. und Smith, A. F. M., *Sampling-Based Approaches to Calculating Marginal Densities*. Journal of the American Statistical Association 85, 398-409 (1990).

[Ge/Ge] Geman, S. und Geman, D., *Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images*. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721-741 (1984).

[Gh/Gr] Ghib, S. und Greenberg, E., *Understanding the Metropolis-Hastings Algorithm*. The American Statistician 49, 327-335 (1995).

[Gi/Ri/Sp] Gilks, W. R., Richardson, S., Spiegelhalter, D. J., *Introduction Markov chain Monte Carlo*. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J., *Markov Chain Monte Carlo in Practice*. Chapman and Hall (1996), 1-17.

[Hw/Sh] Hwang, C.-R., Sheu, S.-J., *On the Geometrical Convergence of the Gibbs Sampler in $\mathbb{R}^d$*. Journal of Multivariate Analysis 66, 22-37 (1998).

[Li/Ge] Li, K. und Geng, Z., *Convergence rate of Gibbs sampler and its applications*. Science in China, Ser. A Mathematics 48, 1430-1439 (2005).

[Liu] Liu, J. S., *The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem*. Journal of the American Statistical Society 89, 958-965 (1994).

[Li/Wo/Ko1] Liu, J. S., Wong, W. H., Kong, A., *Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans*. Journal of the Royal Statistical Society 57, 157-169 (1995).

[Li/Wo/Ko2] Liu, J. S., Wong, W. H., Kong, A., *Covariance structure of the Gibbs Sampler with applications to the comparisons of estimators and augmentation schemes*. Biometrika 81, 27-40 (1994).

[Mac] MacKay, D. J. C., *Information Theory, Inference, Learning Algorithms* (4. Auflage). Cambridge University Press (2005). Online im Internet erhältlich unter URL: http://www.inference.phy.cam.ac.uk/itprnn/book.pdf

[Me] Metropolis et al., *Equation of state calculations by fast computing machines*. Journal of Chem. Phys. 21, 1087-1091 (1953).

[Me/Tw] Meyn, S., Tweedie, R., *Markov Chains and Stochastic Stability*. London (1993).

[Num] Nummelin, E., *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press (1984).

[Ro/Ca] Robert, C. P. und Casella, G., *Monte Carlo Statistical Methods*. New York (2000).

[Rob] Roberts, G. O., *Markov chain concepts related to sampling algorithms*. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J., *Markov Chain Monte Carlo in Practice*. Chapman and Hall (1996), 45-57.

[Ro/Sa] Roberts, G. O. und Sahu, S. K., *Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler*. Journal of the Royal Statistical Society B 59, 291-317 (1997).

[Str] Strüber, F.-J., *Der Random-Walk-basierte Metropolis-Hastings Algorithmus: Konvergenzraten und eine Anwendung auf das Traveling-Salesman Problem*. Diplomarbeit am Institut für Mathematische Statistik der Westfälischen Wilhelms-Universität Münster (2002).

[Tie] Tierney, L., *Introduction to general state-space Markov chain theory*. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J., *Markov Chain Monte Carlo in Practice*. Chapman and Hall (1996), 59-73.

[Var] Varadhan, S. R. S., *Lectures on Diffusion Problems and Partial Differential Equations*. Tata Institute (1980).

[Wer] Werner, D., *Funktionalanalysis* (5. Auflage). New York (2005).

[Yos] Yosida, K., *Functional Analysis* (6. Auflage). New York (1980).

Hiermit versichere ich, dass ich die vorliegende Diplomarbeit selbständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder Sinn nach entnommen wurden, habe ich in jedem Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht.