Albert-Ludwigs-Universität Freiburg

(vorläufiges) Vorlesungsskript

Hidden-Markov-Modelle

Angelika Rohde

Wintersemester 2023/24

Inhaltsverzeichnis

1 Grundlagen **2**
1.1 Wiederholung . . . 2
1.2 Hidden-Markov-Modelle . . . 4
1.2.1 Bedingte Unabhängigkeit . . . 5
1.2.2 Nicht-Degeneriertheit . . . 7

2 Zustandsinferenz (State Inference) **8**
2.1 Filter-, Glättungs- und Vorhersagerekursionen . . . 8
2.2 Einfluss der Startbedingung . . . 12
2.2.1 Totalvariation und Dobrushin-Koeffizient . . . 12
2.2.2 Doeblin-Bedingung und gleichmäßige Ergodizität . . . 15
2.2.3 Vergessen der Startverteilung unter der Atar-Zeitouni-Bedingung . . . 17
2.3 Sequentielle Monte-Carlo-Approximationen . . . 20
2.3.1 Sequentielles Importance Sampling . . . 20
2.3.2 Interacting Particles – Importance Sampling Resampling . . . 21
2.3.3 Konvergenzanalysis des SISR-Algorithmus . . . 23

3 Parameterinferenz **28**
3.1 Grundlagen . . . 28
3.2 Der EM-Algorihmus . . . 29
3.3 Asymptotik des Maximum-Likelihood-Schätzers . . . 32
3.3.1 Konsistenz . . . 33
3.3.2 Ausblick: Asymptotische Normalität . . . 39

Literatur **40**

1 Grundlagen

1.1 Wiederholung

Wir wiederholen zunächst einige Grundlagen.

Definition 1.1. *Seien $(X, \mathcal{X}), (Y, \mathcal{Y})$ messbare Räume. Ein Kern von $(X, \mathcal{X})$ nach $(Y, \mathcal{Y})$ ist eine Abbildung $Q : X \times \mathcal{Y} \to [0, \infty]$ mit den Eigenschaften*

(i) $\forall\, x \in X$ *ist* $Q(x, .)$ *ein Maß auf* $(Y, \mathcal{Y})$,

(ii) $\forall\, A \in \mathcal{Y}$ *ist* $x \mapsto Q(x, A)$ *messbar.*

Er heißt Markovkern oder Wahrscheinlichkeitskern, falls alle $Q(x, .)$ in (i) W-Maße sind. Ist in diesem Falle $X = Y$, spricht man von einem Markovschen Übergangskern. Besitzt Letzterer eine Dichte bezüglich eines σ-endlichen Maßes μ, nennt man dieselbe auch μ-Übergangsdichte.

Markovkerne treten prominent in Form regulärer Versionen bedingter Verteilungen auf. Natürlich spielen diese beim Untersuchen stochastischer Prozesse eine wichtige Rolle.

Bemerkung 1.2 (Eigenschaften von Kernen)**.**

- *Sind Q und R (Markov-) Kerne von $(X, \mathcal{X})$ nach $(Y, \mathcal{Y})$ und von $(Y, \mathcal{Y})$ nach $(Z, \mathcal{Z})$, dann definiert*

$$QR(x, A) := \int R(y, A) Q(x, dy), \quad x \in X,\ A \in \mathcal{Z},$$

  einen (Markov-) Kern von $(X, \mathcal{X})$ nach $(Z, \mathcal{Z})$.

- *W-Maße operieren auf Markovkernen auf zwei Weisen. Sind μ ein W-Maß auf $(X, \mathcal{X})$ und Q ein Markovkern $(X, \mathcal{X})$ nach $(Y, \mathcal{Y})$, dann definiert*

$$\mu Q(A) := \int_X Q(x, A) \mu(dx), \quad A \in \mathcal{Y},$$

  ein W-Maß auf $(Y, \mathcal{Y})$. Darüberhinaus definiert

$$(\mu \otimes Q)(C) := \int_X \int_Y \mathbb{1}_C(x, y) Q(x, dy) \mu(dx), \quad C \in \mathcal{X} \otimes \mathcal{Y},$$

  ein W-Maß auf dem Produktraum $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$.

Für einen Markovschen Übergangskern Q auf $(X, \mathcal{X})$ definiert man induktiv

$$Q^0(x, .) := \delta_x \text{ für } x \in \mathcal{X} \text{ und } Q^k = QQ^{k-1} \text{ für } k \geq 1.$$

Hierfür gelten die **Chapman-Kolmogorov-Gleichungen** $Q^{n+m} = Q^n Q^m$ für alle $m, n \geq 0$ (Übungsblatt 1, Aufgabe 1(a))).

Definition 1.3 (Markovkette)**.** *Seien $(\Omega, \mathcal{F}, (\mathcal{F}_k)_{k\geq 0}, \mathbb{P})$ ein filtrierter W-Raum und Q ein Markovscher Übergangskern auf einem messbaren Raum $(X, \mathcal{X})$. Ein X-wertiger stochastischer Prozess $(X_k)_{k\geq 0}$ heißt (homogene) Markovkette unter $\mathbb{P}$ bezüglich der Filtration $(\mathcal{F}_k)_{k\geq 0}$ mit Übergangskern Q, falls er $(\mathcal{F}_k)$-adaptiert ist und für alle $k \geq 0$ und alle $A \in \mathcal{X}$*

$$P(X_{k+1} \in A | \mathcal{F}_k) = Q(X_k, A) \tag{1.1}$$

erfüllt. Die Verteilung von X_0 heißt Startverteilung der Kette und X wird Zustandsraum genannt.

Im Falle der natürlichen Filtration, d.h. $\mathcal{F}_k = \sigma(X_n : n \leq k)$ für alle $k \geq 0$, spricht man einfach von einer Markovkette (ohne Zusatz). Markovketten sind aus der Grundvorlesung Stochastik bekannt – Beispiele sind Gegenstand von Aufgaben 2 und 3 auf Übungsblatt 1. Gilt (1.1) sinngemäß mit Q_k statt Q für eine Folge von Markovschen Übergangskernen $(Q_k)_{k\geq 0}$, nennt man den Prozess $(X_k)_{k\geq 0}$ inhomogene Markovkette.

Eine zentrale Eigenschaft einer Markovkette ist, dass ihre Verteilung vollständig durch die Startverteilung und den Übergangskern bestimmt wird (die endlichdimensionalen Verteilungen bestimmen ein Maß auf dem unendlichen Produktraum eindeutig):

Proposition 1.4. *Sei $(X_k)_{k\geq 0}$ eine Markovkette mit Startverteilung ν und Übergangskern Q. Dann gilt $\forall\, k \in \mathbb{N}_0$ und jede $(\mathcal{X}^{k+1}, \mathcal{B}(\mathbb{R}))$-messbare Funktion $f : X^{k+1} \to \mathbb{R}$*

$$\mathbb{E}f(X_0, \ldots, X_k) = \int f(x_0, \ldots, x_k) Q(x_{k-1}, dx_k) \ldots Q(x_0, dx_1)\nu(dx_0). \tag{1.2}$$

Beweis. Wir zeigen die Aussage zunächst für $(\mathcal{X}^{k+1}, \mathcal{B}(\mathbb{R}))$-messbare, beschränkte Funktionen f der Form $f(x_0, \ldots, x_k) = \prod_{i=0}^k f_i(x_i)$. Hier gilt nach der Turmeigenschaft der bedingten Erwartung

$$\begin{aligned}
&\mathbb{E}\big(f_0(X_0) \ldots f_k(X_k)\big) \\
&\quad = \mathbb{E}\Big(f_0(X_0) \ldots f_{k-1}(X_{k-1}) \mathbb{E}\big(f(X_k) \,\big|\, X_0, \ldots, X_{k-1}\big)\Big) \\
&\quad = \mathbb{E}\Big(f_0(X_0) \ldots f_{k-1}(X_{k-1}) \int f_k(x_k) Q(X_{k-1}, dx_k)\Big) \\
&\quad = \mathbb{E}\Big(\mathbb{E}\Big(f_0(X_0) \ldots f_{k-1}(X_{k-1}) \int f_k(x_k) Q(X_{k-1}, dx_k) \,\Big|\, X_0, \ldots, X_{k-2}\Big)\Big) \\
&\quad = \mathbb{E}\Big(f_0(X_0) \ldots f_{k-1}(X_{k-2}) \mathbb{E}\Big(f_{k-1}(X_{k-1}) \int f_k(x_k) Q(X_{k-1}, dx_k) \,\Big|\, X_0, \ldots, X_{k-2}\Big)\Big) \\
&\quad = \mathbb{E}\Big(f_0(X_0) \ldots f_{k-1}(X_{k-2}) \int f_{k-1}(x_{k-1}) f_k(x_k) Q(x_{k-1}, dx_k) Q(X_{k-2}, dx_{k-1})\Big) \\
&\qquad \vdots \\
&\quad = \int f_0(x_0) \ldots f_k(x_k) Q(x_{k-1}, dx_k) \ldots Q(x_0, dx_1)\nu(dx_0).
\end{aligned}$$

Die Aussage für allgemeines f folgt dann aus dem Satz über monotone Klassen (oder Dynkin's $\pi - \lambda$-Theorem) → Übungsblatt 1, Aufgabe 1(b): Die Menge der beschränkten Funktionen f, für die (1.2) gilt, bilden einen Vektorraum, der die Konstanten enthält, stabil unter isotonen Limites gegen beschränkte Grenzfunktionen ist (nach dem Satz von der monotonen Konvergenz) und die multiplikative Klasse von messbaren, beschränkten Funktionen $(x_1, \ldots, x_n) \mapsto \prod_{i=0}^k f_i(x_i)$ beinhaltet. □

1.2 Hidden-Markov-Modelle

Ein Hidden-Markov-Modell (HMM) ist ein Markovprozess, der in zwei Komponenten aufgeteilt ist: Eine beobachtbare Komponente und eine unbeobachtbare oder "versteckte" Komponente. D.h. ein Hidden-Markov-Modell ist ein Markovprozess $(X_k, Y_k)_{k\geq 0}$ auf einem Zustandsraum $X \times Y$, wobei wir annehmen, dass nur die zweite Komponente Y_k beobachtet wird, nicht jedoch X_k. Wir nennen auch

(X_k) den Signalprozess, X den Signalzustandsraum

(Y_k) den beobachteten Prozess, Y den Beobachtungszustandsraum.

In einfachen Fällen wie vollkommen diskreten Modellen definiert man HMMs mithilfe bedingter Unabhängigkeiten im Sinne der nachfolgenden Proposition 1.8. Dieses Konzept ist aber insbesondere für nicht abzählbaren Zustandsraum weit schwieriger zu definieren. Deswegen werden HMMs hier als bivariate Markovketten $(X_k, Y_k)_{k\geq 0}$ definiert, deren Übergangskern eine spezielle Eigenschaft hat, und die üblichen bedingten Unabhängigkeitseigenschaften werden anschließend daraus abgeleitet. Eine zentrale Voraussetzung ist folgende:

Sowohl $(X_k, Y_k)_{k\geq 0}$ als auch $(X_k)_{k\geq 0}$ sollen die Markoveigenschaft besitzen. (1.3)

Aufgabe 4 (a) auf Übungsblatt 1 zeigt, dass dies für eine allgemeine bivariate Markovkette nicht zwangsläufig gilt.

Definition 1.5 (Hidden-Markov-Modell)**.** *Seien $(X, \mathcal{X}), (Y, \mathcal{Y})$ messbare Räume, Q ein Markovscher Übergangskern auf $(X, \mathcal{X})$ und G ein Markovkern von $(X, \mathcal{X})$ nach $(Y, \mathcal{Y})$. Dann definiert*

$$T\big((x,y), C\big) = \int_X \int_Y \mathbb{1}_C(x', y') G(x', dy') Q(x, dx'), \quad (x,y) \in X \times Y,\ C \in \mathcal{X} \otimes \mathcal{Y}, \tag{1.4}$$

einen Markovschen Übergangskern auf dem Produktraum $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ (Übungsblatt 1, Aufgabe 1(c)). Die Markovkette $(X_k, Y_k)_{k\geq 0}$ mit Übergangskern T und Startverteilung $\nu \otimes G$ für ein W-Maß ν auf $(X, \mathcal{X})$ heißt Hidden-Markov-Modell.

Aus der Definition von T ist ersichtlich, dass die Verteilung von (X_k, Y_k) für eine Markovkette $(X_k, Y_k)_{k\geq 0}$ mit Übergangskern T nur von der Marginalverteilung von X_0 abhängt, was auch immer die gemeinsame Verteilung von (X_0, Y_0) ist. Für ein HMM gemäß Definition 1.5 gelten

$$\mathbb{E}\big(f(X_{k+1}, Y_{k+1}) \,\big|\, X_0, Y_0, \ldots, X_k, Y_k\big) = \int f(x,y) G(x, dy) Q(X_k, dx)$$

sowie

$$\mathbb{E}(f(X_0, Y_0)) = \int f(x,y) G(x, dy) \nu(dx)$$

für jede beschränkte, messbare Funktion $f : X \times Y \to \mathbb{R}$. Der Prozess $(X_k)_{k\geq 0}$ ist ebenfalls eine Markovkette (auch bezüglich der Filtration $(\sigma(X_0, Y_0, \ldots, X_k, Y_k))_{k\geq 0}$) mit Übergangskern Q und Startverteilung ν, d.h. ein HMM besitzt insbesondere die Eigenschaft (1.3).

Bemerkung 1.6. *Aufgabe 2 auf Übungsblatt 2 zeigt, dass eine bivariate Markovkette mit Eigenschaft (1.3) nicht notwendig ein HMM sein muss, d.h. Eigenschaft (1.3) erzwingt nicht die Gestalt des Übergangskerns T aus Definition 1.5.*

Wir bezeichnen nachfolgend mit $\mathbb{P}_\nu$ die Verteilung von $(X_k, Y_k)_{k\geq 0}$ auf dem Produktraum $((X\times Y)^{\mathbb{N}_0}, (\mathcal{X}\otimes\mathcal{Y})^{\mathbb{N}_0})$ und mit $\mathbb{E}_\nu$ den assoziierten Erwartungswert. Den Markovkern G aus Definition 1.5 nennen wir auch Beobachtungskern. Sind sowohl X als auch Y abzählbar, heißt das HMM diskret.

Beispiel 1.7. *Seien $(\xi_k)_{k\geq 1}$ und $(\eta_k)_{k\geq 0}$, unabhängige iid-Folgen reellwertiger Zufallsvariablen mit Marginalverteilungen μ_ξ und μ_η. Für messbare Funktionen $f : X\times\mathbb{R}\to X$ und $h : X\times\mathbb{R}\to Y$ sowie $x_0\in X$ definieren wir $X_0 := x_0$, $Y_0 := h(X_0,\eta_0)$ sowie*

$$\begin{aligned} X_k &:= f(X_{k-1},\xi_k),\\ Y_k &:= h(X_k,\eta_k) \qquad \text{für } k\geq 1. \end{aligned}$$

Dann ist $(X_k, Y_k)_{k\geq 0}$ ein Hidden-Markov-Modell mit

$$Q(x,A) = \int \mathbb{1}_A\big(f(x,z)\big)\mu_\xi(dz)$$

und Beobachtungskern

$$G(x,B) = \int \mathbb{1}_B\big(h(x,z)\big)\mu_\eta(dz)$$

sowie $\nu = \delta_{x_0}$. Das sieht man folgendermaßen. Nach Aufgabe 2 auf Übungsblatt 1 ist $(X_k)_{k\geq 0}$ eine Markovkette (auch bezüglich $(\sigma(X_0,Y_0,\dots,X_k,Y_k))_{k\geq 0}$) mit Übergangskern Q. Weiter ist damit für jede beschränkte, messbare Funktion $g : X\times Y\to\mathbb{R}$ einerseits nach der Turmeigenschaft der bedingten Erwartung und der stochastischen Unabhängigkeit von η_{k+1} und $(X_0,\dots,X_{k+1},Y_0,\dots,Y_k)$

$$\begin{aligned} &\mathbb{E}\big(g(X_{k+1},Y_{k+1}) \,\big|\, X_0,Y_0,\dots,X_k,Y_k\big)\\ &\quad= \mathbb{E}\Big(g\big(X_{k+1},h(X_{k+1},\eta_{k+1})\big) \,\Big|\, X_0,Y_0,\dots,X_k,Y_k\Big)\\ &\quad= \mathbb{E}\Big[\mathbb{E}\Big(g\big(X_{k+1},h(X_{k+1},\eta_{k+1})\big) \,\Big|\, X_0,\dots,X_{k+1},Y_0,\dots,Y_k\Big) \,\Big|\, X_0,Y_0,\dots,X_k,Y_k\Big]\\ &\quad= \mathbb{E}\Big[\int g\big(X_{k+1},h(X_{k+1},z)\big)d\mu_\eta(z) \,\Big|\, X_0,Y_0,\dots,X_k,Y_k\Big]\\ &\quad= \mathbb{E}\Big[\int g(X_{k+1},y)G(X_{k+1},dy) \,\Big|\, X_0,Y_0,\dots,X_k,Y_k\Big]\\ &\quad= \int g(x,y)G(x,dy)Q(X_k,dx). \end{aligned}$$

Andererseits ist

$$\mathbb{E}g(X_0,Y_0) = \mathbb{E}g\big(x_0,h(x_0,\eta_0)\big) = \int g\big(x_0,h(x_0,z)\big)d\mu_\eta(z) = \int g(x_0,y)G(x_0,dy),$$

womit $\delta_{x_0}\otimes G$ als Startverteilung von (X_0,Y_0) identifiziert ist.

1.2.1 Bedingte Unabhängigkeit

In Beispiel 1.7 ist Y_k eine Funktion ausschließlich von X_k und einem unabhängigen Fehler η_k, der unabhängig ist von den Fehlern der übrigen Y_l, $l\neq k$. Ist $(Y_k)_{k\geq 0}$ die übertragene verrauschte Beobachtung eines Signalprozesses $(X_k)_{k\geq 0}$, entspricht dies einer Gedächtnislosigkeit des Übertragungsmechanismus'.

Proposition 1.8. *Über dem Produktraum $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ sei $(X_k, Y_k)_{k\geq 0}$ eine Markovkette mit Übergangskern T aus (1.4). Dann gilt für jedes $p \in \mathbb{N}$, jede geordnete Teilmenge $\{k_1, \dots, k_p\} \subset \mathbb{N}_0$, und alle beschränkten, messbaren Funktionen $f_1, \dots, f_p : Y \to \mathbb{R}$*

$$\mathbb{E}_\nu\left[\prod_{i=1}^{p} f_i(Y_{k_i}) \,\middle|\, X_{k_1}, \dots, X_{k_p}\right] = \prod_{i=1}^{p} \int f_i(y) G(X_{k_i}, dy).$$

Beweis. Für jede messbare und beschränkte Funktion $h : (X^p, \mathcal{X}^p) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gilt nach Proposition 1.4

$$\begin{aligned}
&\mathbb{E}_\nu\left[h(X_{k_1}, \dots X_{k_p}) \prod_{i=1}^{p} f_i(Y_{k_i})\right] \\
&= \int \cdots \int \left[\prod_{i=1}^{p} f_i(y_{k_i})\right] h(x_{k_1}, \dots x_{k_p}) G(x_0, dy_0) \nu(dx_0) \prod_{i=1}^{k_p} G(x_i, dy_i) Q(x_{i-1}, dx_i) \\
&= \int_X \cdots \int_X \left(\int_Y \cdots \int_Y G(x_0, dy_0) \left[\prod_{i=1}^{p} f_i(y_{k_i})\right] \left[\prod_{i=1}^{k_p} G(x_i, dy_i)\right]\right) \\
&\qquad\qquad h(x_{k_1}, \dots x_{k_p}) \nu(dx_0) \prod_{i=1}^{k_p} Q(x_{i-1}, dx_i) \\
&= \int_X \cdots \int_X \left(\int_Y G(x_0, dy_0) \prod_{i=1}^{p} \int_Y f_i(y_{k_i}) G(x_{k_i}, dy_{k_i})\right) \\
&\qquad\qquad h(x_{k_1}, \dots x_{k_p}) \nu(dx_0) \prod_{i=1}^{k_p} Q(x_{i-1}, dx_i),
\end{aligned}$$

wobei wir beim letzten Gleichheitszeichen $\int_Y G(x_i, dy_i) = 1$ verwendet haben. Also ist

$$\mathbb{E}_\nu\left[h(X_{k_1}, \dots X_{k_p}) \prod_{i=1}^{p} f_i(Y_{k_i})\right] = \mathbb{E}_\nu\left[h(X_{k_1}, \dots, X_{k_p}) \prod_{i=1}^{p} \int_Y f(y_{k_i}) G(X_{k_i}, dy_{k_i})\right].$$

Die Behauptung folgt nun aus der definierenden Eigenschaft der bedingten Erwartung. □

Korollar 1.9. *(i) Für jedes $p \in \mathbb{N}$ und jede Menge $\{k_1, \dots, k_p\} \subset \mathbb{N}_0$, $k_1 < \cdots < k_p$, sind die Zufallsvariablen $Y_{k_1}, \dots, Y_{k_p}$ bedingt unabhängig gegeben $X_{k_1}, \dots, X_{k_p}$.*

(ii) Für jedes $k \geq 0$ und $p \in \mathbb{N}$ und jede Menge $\{k_1, \dots, k_p\} \subset \mathbb{N}_0$, $k_1 < \cdots < k_p$, mit $k \notin \{k_1, \dots, k_p\}$ sind Y_k und $(X_{k_1}, \dots, X_{k_p})$ bedingt unabhängig gegeben X_k.

Beweis. (i) folgt unmittelbar aus Proposition 1.8. Seien $f : (Y, \mathcal{Y}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ und $h : (X^p, \mathcal{X}^{\otimes p}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ beschränkte messbare Funktionen. Dann gilt nach der Turmeigenschaft der bedingten Erwartung

$$\begin{aligned}
\mathbb{E}_\nu\big[f(Y_k) h(X_{k_1}, \dots, X_{k_p}) \,\big|\, X_k\big] &= \mathbb{E}_\nu\Big[\mathbb{E}_\nu\big[f(Y_k) \,\big|\, X_{k_1}, \dots, X_{k_p}, X_k\big] h(X_{k_1}, \dots, X_{k_p}) \,\Big|\, X_k\Big] \\
&\overset{\text{Prop. 1.8}}{=} \mathbb{E}_\nu\Big[\mathbb{E}_\nu\big[f(Y_k) \,\big|\, X_k\big] h(X_{k_1}, \dots, X_{k_p}) \,\Big|\, X_k\Big]
\end{aligned}$$

$$= \mathbb{E}_\nu\big[f(Y_k) \,\big|\, X_k\big]\mathbb{E}_\nu\big[h(X_{k_1},\dots,X_{k_p}) \,\big|\, X_k\big]$$

was (ii) beweist. □

Die bedingte Unabhängigkeit der Beobachtungen (Y_k) gegeben die zugrundeliegende Folge von Zuständen (X_k) impliziert nun Folgendes:

Seien $p, p' \in \mathbb{N}$, $k_1 < \cdots < k_p$, $k'_1 < \cdots < k'_{p'}$, so dass $\{k_1,\dots,k_p\} \cap \{k'_1,\dots,k'_{p'}\} = \emptyset$. Dann gilt für jede beschränkte, messbare Funktion $f : (Y^p, \mathcal{Y}^p) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$

$$\mathbb{E}_\nu\Big[f(Y_{k_1},\dots,Y_{k_p}) \,\Big|\, X_{k_1},\dots,X_{k_p},X_{k'_1},\dots X_{k'_{p'}},Y_{k'_1},\dots Y_{k'_{p'}}\Big] = \mathbb{E}_\nu\Big[f(Y_{k_1},\dots,Y_{k_p}) \,\Big|\, X_{k_1},\dots,X_{k_p}\Big].$$

Denn in Ausdrücken bedingter Unabhängigkeit (Notation $\perp\!\!\!\perp$) gilt einerseits

$$(Y_{k_1},\dots Y_{k_p}) \perp\!\!\!\perp (Y_{k'_1},\dots,Y_{k'_{p'}}) | (X_{k_1},\dots,X_{k_p},X_{k'_1},\dots,X_{k'_{p'}})$$

nach Proposition 1.8, andererseits erhält man wie in Korollar 1.9 (ii)

$$(Y_{k_1},\dots Y_{k_p}) \perp\!\!\!\perp (X_{k'_1},\dots,X_{k'_{p'}}) | (X_{k_1},\dots,X_{k_p})$$

und somit

$$(Y_{k_1},\dots Y_{k_p}) \perp\!\!\!\perp (X_{k'_1},\dots,X_{k'_{p'}},Y_{k'_1},\dots,Y_{k'_{p'}}) | (X_{k_1},\dots,X_{k_p}).$$

1.2.2 Nicht-Degeneriertheit

Beispiel 1.10. *Sei $X = Y = \mathbb{R}$. Seien $(\xi_k)_{k\geq 0}$ eine iid-Folge $\mathbb{Z}$-wertiger Zufallsvariablen und $(\eta_k)_{k\geq 0}$ eine iid-Folge $\mathbb{N}$-wertiger Zufallsvariablen. Definiert man nun rekursiv*

$$X_0 = Y_0 = 0, \quad X_k = X_{k-1} + \frac{\xi_k}{\eta_k}, \quad Y_k = X_k \quad (k \geq 1),$$

so ist die bivariate Markovkette $(X_k, Y_k)_{k\geq 0}$ ein HMM gemäß Definition 1.5. Der Beobachtungsprozess (Y_k) ist $\mathbb{Q}$-wertig. Ist der Signalprozess (X_k) aber auch nur minimal verrauscht, haben Beobachtungen $Y_0,\dots,Y_N$ aus der realen Welt diese Eigenschaft nicht länger. Ein statistisches Inferenzverfahren mag aber problematisch sein, wenn die Input-Variablen nach Modellannahme gar nicht möglich sind.

Natürlich ist das Beispiel sehr konstruiert. Wir werden aber Techniken entwickeln, die auf einem Abschnitt $Y_0, Y_1, \dots, Y_N$ der Beobachtungszeitreihe operieren, und diese sollen auch sinnvoll (und insbesondere definiert) sein, wenn die Zeitreihe nicht exakt dem angenommenen Modell genügt. Zusätzlich zu den Eigenschaften aus Definition 1.5 wird daher häufig folgende stärkere Voraussetzung an die Struktur des Beobachtungsprozesses $(Y_k)_{k\geq 0}$ gestellt, um die Problematik – wie wir später sehen werden – aus obigem Beispiel auszuräumen.

Definition 1.11. *Sei $(X_k, Y_k)_{k\geq 0}$ ein HMM auf $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ mit Beobachtungskern G aus Definition 1.5. Man sagt, das Modell habe nicht-degenerierte Beobachtungen, falls G von folgender Gestalt ist:*

$$G(x, B) = \int \mathbb{1}_B(y)\gamma(x,y)\phi(dy), \quad x \in X,\ B \in \mathcal{Y}, \tag{1.5}$$

mit einer strikt positiven, messbaren Funktion $\gamma : X \times Y \to \mathbb{R}$ und einem W-Maß ϕ auf $(Y, \mathcal{Y})$. γ wird Beobachtungsdichte genannt.

2 Zustandsinferenz (State Inference)

Sei $(X_k, Y_k)_{k\geq 0}$ ein vollständig charakterisiertes HMM, d.h. ν, Q und G sind bekannt (die statistisch deutlich relevantere Situation unbekannter Kerne wird später diskutiert). In diesem Kaptitel beschäftigen wir uns unter dieser Voraussetzung mit folgender Frage:

> Gegeben ein Beobachtungssegment $Y_0, \dots, Y_N$, was kann man über den Signalprozess $(X_k)_{k\geq 0}$ aussagen?

2.1 Filter-, Glättungs- und Vorhersagerekursionen

Für eine beschränkte, messbare Funktion $f : (X, \mathcal{X}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ differenziert man das damit insbesondere verbundene Problem, $\mathbb{E}(f(X_k)|Y_0, \dots, Y_N)$ zu bestimmen, in

- $k < N$ (Glättungsproblem),
- $k = N$ (Filterproblem) und
- $k > N$ (Vorhersageproblem).

Die Lösungen werden allesamt rekursiver Natur sein – eine Konsequenz aus der Markovstruktur des HMMs: Bedingt auf $Y_0, \dots, Y_N$ ist der Signalprozess zwar keine homogene, aber dennoch eine inhomogene Markovkette ($\to$ Übungsblatt 3, Aufgabe 1).

Zentral für die genannten Probleme ist die bedingte Verteilung $\pi_{k|N} := \mathbb{P}^{X_k|Y_0,\dots,Y_N}$. Ist das HMM nicht-degeneriert (Definition 1.11), folgt aus Proposition 1.4 mit T aus Definition 1.5

$$\begin{aligned}&\mathbb{E}f(X_0, Y_0, \dots, X_k, Y_k)\\ &= \int f(x_0, y_0 \dots, x_k, y_k)\gamma(x_0, y_0)\phi(dy_0)\nu(dx_0)\prod_{l=1}^{k}\gamma(x_l, y_l)Q(x_{l-1}, dx_l)\phi(dy_l).\end{aligned}$$

Hieraus lässt sich dann $\pi_{k|N}$ mithilfe der Bayes-Formel ermitteln, die wir zur Erinnerung noch einmal formulieren.

Lemma 2.1 (Satz von Bayes)**.** *Seien U eine $(X, \mathcal{X})$-wertige und V eine $(Y, \mathcal{Y})$-wertige Zufallsvariable auf einem W-Raum $(\Omega, \mathcal{A}, \mathbb{P})$. Angenommen, für W-Maße μ_U und μ_V auf $(X, \mathcal{X})$ bzw. $(Y, \mathcal{Y})$ sei $\mathbb{P}^{(U,V)} \ll \mu_U \otimes \mu_V$ mit einer Dichte $\gamma : X \times Y \to (0, \infty)$. Dann gilt*

$$\underbrace{\mathbb{P}^{U|V=y}(A)}_{=:k(y,A)} = \frac{\int \mathbb{1}_A(x)\gamma(x, y)\mu_U(dx)}{\int \gamma(x, y)\mu_U(dx)} \quad \forall\, A \in \mathcal{X}, y \in Y.$$

Beweis. Nach Definition der bedingten Erwartung ist zu zeigen, dass

$$\mathbb{E}\big(k(V, A)\mathbb{1}_C(V)\big) = \mathbb{E}\big(\mathbb{1}_A(U)\mathbb{1}_C(V)\big)$$

für alle $A \in \mathcal{X}$ und $C \in \mathcal{Y}$ gilt. Umformen der linken Seite ergibt

$$\mathbb{E}\big(k(V, A)\mathbb{1}_C(V)\big) = \mathbb{E}\left[\frac{\int \mathbb{1}_A(x)\mathbb{1}_C(V)\gamma(x, V)\mu_U(dx)}{\int \gamma(x, V)\mu_U(dx)}\right]$$

$$
\begin{aligned}
&= \int \frac{\int \mathbb{1}_A(x)\mathbb{1}_C(y)\gamma(x,y)\mu_U(dx)}{\int \gamma(x,y)\mu_U(dx)}\gamma(x',y)\mu_U(dx')\mu_V(dy)\\
&= \int \mathbb{1}_A(x)\mathbb{1}_C(y)\gamma(x,y)\mu_U(dx)\mu_V(dy)\\
&= \mathbb{E}\big(\mathbb{1}_A(U)\mathbb{1}_C(V)\big). \qquad \square
\end{aligned}
$$

Filterrekursion

Sei nun der Kern $\sigma_k : Y^{k+1} \times \mathcal{X} \to [0,\infty)$ gegeben durch

$$
\sigma_k((y_0,\dots,y_k),A) := \int \mathbb{1}_A(x_k)\gamma(x_0,y_0)\nu(dx_0)\prod_{l=1}^{k}\gamma(x_l,y_l)Q(x_{l-1},dx_l). \tag{2.1}
$$

Satz 2.2 (Filterrekursion)**.** *Die bedingte Verteilung $\pi_{k|k}$ besitzt die Darstellung*

$$
\pi_{k|k}((y_0,\dots,y_k),A) = \frac{\sigma_k((y_0,\dots,y_k),A)}{\sigma_k((y_0,\dots,y_k),X)} \tag{2.2}
$$

für alle $A \in \mathcal{X}$ und $y_0,\dots,y_k \in Y$. Darüberhinaus gilt die Rekursion

$$
\sigma_k((y_0,\dots,y_k),A) = \int \mathbb{1}_A(x)\gamma(x,y_k)Q(x',dx)\sigma_{k-1}((y_0,\dots,y_{k-1}),dx') \quad (k \geq 1)
$$

mit

$$
\sigma_0(y_0,A) = \int \mathbb{1}_A(x)\gamma(x,y_0)\nu(dx_0).
$$

Beweis. Wir definieren μ_Y auf $(Y^{k+1},\mathcal{Y}^{k+1})$ sowie μ_X auf $(X^{k+1},\mathcal{X}^{k+1})$ als

$$
\begin{aligned}
\mu_Y(dy_0,\dots,dy_k) &:= \phi(dy_0)\cdot\ldots\cdot\phi(dy_k),\\
\mu_X(dx_0,\dots,dx_k) &:= Q(x_{k-1},dx_k)\cdot\ldots\cdot Q(x_0,dx_1)\nu(dx_0);
\end{aligned}
$$

ferner $\Gamma(x_0,y_0,\dots,x_k,y_k) := \prod_{l=0}^{k}\gamma(x_l,y_l)$. Nach der Bayes-Formel ist

$$
\begin{aligned}
&\int f(x_0,\dots,x_k)d\mathbb{P}^{(X_0,\dots,X_k)|(Y_0,\dots,Y_k)=(y_0,\dots,y_k)}(x_0,\dots,x_k)\\
&\quad = \frac{\int f(x_0,\dots,x_k)\Gamma(x_0,y_0,\dots,x_k,y_k)d\mu_X(x_0,\dots,x_k)}{\int \Gamma(x_0,y_0,\dots,x_k,y_k)d\mu_X(x_0,\dots,x_k)}.
\end{aligned}
$$

Damit folgt für jede messbare und beschränkte Funktion $f : X \to \mathbb{R}$

$$
\begin{aligned}
&\int f(x_k)\pi_{k|k}((y_0,\dots,y_k),dx_k)\\
&\quad = \int f(x_k)d\mathbb{P}^{X_k|(Y_0,\dots,Y_k)=(y_0,\dots,y_k)}(x_k)\\
&\quad = \int f(x_k)d\mathbb{P}^{(X_0,\dots,X_k)|(Y_0,\dots,Y_k)=(y_0,\dots,y_k)}(x_0,\dots,x_k)\\
&\quad = \frac{\int f(x_k)\Gamma(x_0,y_0,\dots,x_k,y_k)d\mu_X(x_0,\dots,x_k)}{\int \Gamma(x_0,y_0,\dots,x_k,y_k)d\mu_X(x_0,\dots,x_k)}\\
&\quad = \frac{\int f(x_k)\sigma_k((y_0,\dots,y_k),dx_k)}{\int \sigma_k((y_0,\dots,y_k),dx_k)}.
\end{aligned}
$$

Die behauptete Rekursion ist unmittelbar aus der Definition von σ_k ersichtlich. $\square$

Bemerkung 2.3. *Mithilfe der rekursiven Darstellung von σ_k erhält man auch eine solche für $\pi_{k|k}$:*

$$\pi_{k|k}((y_0,\dots,y_k),A) = \frac{\int \mathbb{1}_A(x)\gamma(x,y_k)Q(x',dx)\pi_{k-1|k-1}((y_0,\dots,y_{k-1}),dx')}{\int \gamma(x,y_k)Q(x',dx)\pi_{k-1|k-1}((y_0,\dots,y_{k-1}),dx')} \tag{2.3}$$

($k \geq 1$) mit

$$\pi_{0|0}(y_0,A) = \frac{\int \mathbb{1}_A(x)\gamma(x,y_0)\nu(dx_0)}{\int \gamma(x,y_0)\nu(dx_0)}.$$

Die rekursive Gestalt ist für die Berechnung von $\pi_{k|k}$ vorteilhaft: Um $\pi_{k|k}$ zu ermitteln, benötigt man lediglich $\pi_{k-1|k-1}$ und Y_k.

Glättungsrekursion

Um $\pi_{k|N}$ für $0 \leq k < N$ zu bestimmen, verwenden wir erneut die Bayes-Formel. Dazu definieren wir die (unnormierte) *Glättungsdichte* $\beta_{k|N} : X \times Y^{N-k} \to (0,\infty)$ durch

$$\beta_{k|N}(x_k,(y_{k+1},\dots,y_N)) := \int_{X^{N-k}} \prod_{l=k+1}^{N} \gamma(x_l,y_l)Q(x_{l-1},dx_l). \tag{2.4}$$

Satz 2.4 (Glättungsrekursion)**.** *Die bedingte Verteilung $\pi_{k|N}$ ($k < N$) besitzt die Darstellung*

$$\pi_{k|N}((y_0,\dots,y_N),A) = \frac{\int \mathbb{1}_A(x)\beta_{k|N}(x,(y_{k+1},\dots,y_N))\sigma_k((y_0,\dots,y_k),dx)}{\int \beta_{k|N}(x,(y_{k+1},\dots,y_N))\sigma_k((y_0,\dots,y_k),dx)} \tag{2.5}$$

für alle $A \in \mathcal{X}$ und $y_0,\dots,y_N \in Y$. Darüberhinaus erfüllt $\beta_{k|N}$ die Rückwärtsrekursion

$$\beta_{k|N}(x,(y_{k+1},\dots,y_N)) = \int \beta_{k+1|N}(x',(y_{k+2},\dots,y_N))\gamma(x',y_{k+1})Q(x,dx') \tag{2.6}$$

mit der Endbedingung $\beta_{N|N} = 1$.

Die Beobachtungen $Y_0,\dots,Y_k$ und $Y_{k+1},\dots,Y_N$ treten in dieser Rekursion in unterschiedlicher Weise auf.

Beweis. Mit μ_X, μ_Y und Γ wie im Beweis von Satz 2.2 (mit N anstelle des dortigen k's) ist für jede messbare und beschränkte Funktion $f : (X,\mathcal{X}) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$

$$\begin{aligned}
&\int f(x_k)\pi_{k|N}((y_0,\dots,y_N),dx) \\
&= \int f(x_k)d\mathbb{P}^{X_k|(Y_0,\dots,Y_N)=(y_0,\dots,y_N)}(x_k) \\
&= \int f(x_k)d\mathbb{P}^{(X_0,\dots,X_N)|(Y_0,\dots,Y_N)=(y_0,\dots,y_N)}(x_0,\dots,x_N) \\
&= \frac{\int f(x_k)\Gamma(x_0,y_0,\dots,x_N,y_N)d\mu_X(x_0,\dots,x_N)}{\int \Gamma(x_0,y_0,\dots,x_N,y_N)d\mu_X(x_0,\dots,x_N)} \\
&= \frac{\int f(x_k)\beta_{k|N}(x_k,(y_{k+1},\dots,y_N))\sigma_k((y_0,\dots,y_k),dx_k)}{\int \beta_{k|N}(x_k,(y_{k+1},\dots,y_N))\sigma_k((y_0,\dots,y_k),dx_k)},
\end{aligned}$$

wobei beim vorletzten Gleichheitszeichen die Bayes-Formel verwendet wurde. Die behauptete Rückwärtsrekursion ergibt sich unmittelbar aus der Definition von $\beta_{k|N}$. □

Korollar 2.5. *Definieren wir für $k < N$ nun $\bar{\beta}_{k|N} : X \times Y^{N+1} \to (0, \infty)$ durch die Rückwärtsrekursion*

$$\bar{\beta}_{k|N}(x, (y_0, \dots, y_N)) := \frac{\int \bar{\beta}_{k+1|N}(x', (y_0, \dots, y_N))\gamma(x', y_{k+1})Q(x, dx')}{\int \gamma(x', y_{k+1})Q(x, dx')\pi_{k|k}((y_0, \dots, y_k), dx)} \tag{2.7}$$

$(1 \leq k \leq N)$ mit $\bar{\beta}_{N|N} := 1$, so gilt für alle $k < N$

$$\pi_{k|N}((y_0, \dots, y_N), A) = \int \mathbb{1}_A(x)\bar{\beta}_{k|N}(x, (y_0, \dots, y_N))\pi_{k|k}((y_0, \dots, y_k), dx)$$

Bemerkung 2.6. *Was die Berechnung anbelangt, müssen hier aber zuerst die Filterverteilungen $\pi_{k|k}$ mittels* Vorwärtsrekursion *aus (2.3) ermittelt werden, bevor $\pi_{k|N}$ für $k < N$ dann über die* Rückwärtsrekursion *von $\bar{\beta}_{k|N}$ aus (2.7) bestimmt werden kann. Dieses Vorgehen bezeichnet man als den Vorwärts-Rückwärts-Algorithmus.*

Beweis. (2.5) ergibt zusammen mit (2.2) die Identität

$$\begin{aligned}\pi_{k|N}((y_0, \dots, y_N), A) &= \frac{\int \mathbb{1}_A(x)\beta_{k|N}(x, (y_{k+1}, \dots, y_N))\pi_{k|k}((y_0, \dots, y_k), dx)}{\int \beta_{k|N}(x, (y_{k+1}, \dots, y_N))\pi_{k|k}((y_0, \dots, y_k), dx)} \\ &= \int \mathbb{1}_A(x)\tilde{\beta}_{k|N}(x, (y_0, \dots, y_N))\pi_{k|k}((y_0, \dots, y_k), dx)\end{aligned}$$

mit

$$\tilde{\beta}_{k|N}(x, (y_0, \dots, y_N)) := \frac{\beta_{k|N}(x, (y_{k+1}, \dots, y_N))}{\int \beta_{k|N}(x, (y_{k+1}, \dots, y_N))\pi_{k|k}((y_0, \dots, y_k), dx)} \tag{2.8}$$

Verwenden wir in (2.8) nun die Rückwärtsrekursion für $\beta_{k|N}$ aus Satz 2.4 und erweitern anschließend den Bruch mit $\int \beta_{k+1|N}(u, (y_{k+2}, \dots, y_N))\pi_{k-1|k-1}((y_0, \dots, y_{k+1}), du)$, so ergibt sich mit $\tilde{\beta}_{N|N} = 1$ für $\tilde{\beta}_{k|N}$ $(k < N)$:

$$\tilde{\beta}_{k|N}(x, (y_0, \dots, y_N)) = \frac{\int \tilde{\beta}_{k+1|N}(x', (y_0, \dots, y_N))\gamma(x', y_{k+1})Q(x, dx')}{\int \tilde{\beta}_{k+1|N}(x', (y_0, \dots, y_N))\gamma(x', y)Q(x, dx')\pi_{k|k}((y_0, \dots, y_k), dx)}.$$

Um $\tilde{\beta}_{k|N} = \bar{\beta}_{k|N}$ zu zeigen, bleibt nachzuweisen, dass

$$\begin{aligned}&\int \tilde{\beta}_{k+1|N}(x', (y_0, \dots, y_N))\gamma(x', y)Q(x, dx')\pi_{k|k}((y_0, \dots, y_k), dx) \\ &\qquad = \int \gamma(x', y)Q(x, dx')\pi_{k|k}((y_0, \dots, y_k), dx).\end{aligned}$$

Aber aus (2.3) folgt

$$\begin{aligned}&\frac{\int \tilde{\beta}_{k+1|N}(x', (y_0, \dots, y_N))\gamma(x', y)Q(x, dx')\pi_{k|k}((y_0, \dots, y_k), dx)}{\int \gamma(x', y)Q(x, dx')\pi_{k|k}((y_0, \dots, y_k), dx)} \\ &\qquad = \int \tilde{\beta}_{k+1|N}(x', (y_0, \dots, y_N))\pi_{k+1|k+1}((y_0, \dots, y_{k+1}), dx') = 1\end{aligned}$$

nach Konstruktion von $\tilde{\beta}_{k+1|N}$. □

Vorhersagerekursion

Auch für $k > N$ lässt sich die bedingte Verteilung $\pi_{k|N}$ mithilfe des Bayes-Formel ermitteln (Übungsblatt 4, Aufgabe 1) – wir führen hier aber einen anderen Beweis.

Satz 2.7 (Vorhersagerekursion)**.** *Für $k > N$ gilt die Rekursion*

$$\pi_{k|N}((y_0,\dots,y_N),A) = \int \mathbb{1}_A(x)Q(x',dx)\pi_{k-1|N}((y_0,\dots,y_N),dx')$$

für alle $A \in \mathcal{X}$ und $y_0,\dots,y_N \in Y$, wobei $\pi_{N|N}$ die Startbedingung ist.

Beweis. Sei $f : (X,\mathcal{X}) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$ eine beschränkte, messbare Funktion. Dann gilt nach der Turmeigenschaft der bedingten Erwartung sowie der Markoveigenschaft von $(X_k)_{k\geq 0}$

$$\begin{aligned}\mathbb{E}\big(f(X_k)\,\big|\,Y_0,\dots,Y_N\big) &= \mathbb{E}\Big(\mathbb{E}\big(f(X_k)\,\big|\,X_0,Y_0,\dots,X_N,Y_N\big)\,\Big|\,Y_0,\dots,Y_N\Big)\\ &= \mathbb{E}\bigg(\int f(u)Q^{k-N}(X_N,du)\,\bigg|\,Y_0,\dots,Y_N\bigg)\\ &= \int f(u)Q^{k-N}(x,du)\pi_{N|N}((Y_0,\dots,Y_N),dx),\end{aligned}$$

also nach der Chapman-Kolmogorov-Gleichung (Assoziativgesetz)

$$\pi_{k|N} = \pi_{N|N}Q^{k-N} = \pi_{N|N}(Q^{(k-1)-N}Q) = (\pi_{N|N}Q^{(k-1)-N})Q = \pi_{k-1|N}Q. \qquad \square$$

2.2 Einfluss der Startbedingung

In diesem Abschnitt untersuchen wir die Abhängigkeit der in Abschnitt 2.1 studierten bedingten Verteilung $\pi_{k|N} = \pi_{\nu,k|N}$ von der Startverteilung $\mathbb{P}^{X_0} = \nu$.

> Wie nah liegen bspw. $\pi_{\nu,k|N}$ und $\pi_{\nu',k|N}$ für zwei verschiedene Startbedingungen ν,ν' und für großes k zusammen?

2.2.1 Totalvariation und Dobrushin-Koeffizient

Seien $(X,\mathcal{X})$ und $(Y,\mathcal{Y})$ zwei messbare Räume. Wir bezeichnen die Menge der endlichen, signierten Maße auf $(X,\mathcal{X})$ mit $\mathcal{M}(X,\mathcal{X})$ und mit $\mathcal{F}_b(X,\mathcal{X})$ die Menge der beschränkten, messbaren reellwertigen Funktionen. Mit einem Markovkern K von $(X,\mathcal{X})$ nach $(Y,\mathcal{Y})$ werden zwei Abbildungen assoziiert:

- die Abbildung $\mathcal{M}(X,\mathcal{X}) \to \mathcal{M}(Y,\mathcal{Y})$, die $\xi \in \mathcal{M}(X,\mathcal{X})$ das endliche, signierte Maß $(\xi K)(\cdot) := \int K(x,\cdot)\xi(dx)$ auf $(Y,\mathcal{Y})$ (siehe Bemerkung 1.2) zuordnet und
- die Abbildung $\mathcal{F}_b(Y,\mathcal{Y}) \to \mathcal{F}_b(X,\mathcal{X})$, die einer beschränkten, messbaren Funktion $f : (Y,\mathcal{Y}) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$ die Funktion $(Kf)(\cdot) := \int f(y)K(\cdot,dy)$ zuordnet.

Wir bezeichnen mit $\xi = \xi_+ - \xi_-$ die Hahn-Jordan-Zerlegung von $\xi \in \mathcal{M}(X,\mathcal{X})$ und verwenden die Definition der Totalvariation

$$\|\xi\|_{TV} := \sup_{A\in\mathcal{X}} \big(|\xi(A)| + |\xi(A^c)|\big) = \xi_+(X) + \xi_-(X)$$

für ein endliches, signiertes Maß $\xi \in \mathcal{M}(X, \mathcal{X})$. An dieser Stelle sei darauf hingewiesen, dass diese Definition in der Literatur nicht einheitlich ist und gelegentlich einen zusätzlichen Faktor 1/2 enthält. Mithilfe der Hahn-Jordan-Zerlegung zeigt man folgenden Zusammenhang.

Lemma 2.8. *(i) Für $\xi \in \mathcal{M}(X, \mathcal{X})$ und $f \in \mathcal{F}_b(X, \mathcal{X})$ gilt*

$$\left| \int_X f d\xi \right| \leq \|\xi\|_{TV} \|f\|_{\sup}.$$

(ii) Für $\xi \in \mathcal{M}(X, \mathcal{X})$ ist

$$\|\xi\|_{TV} = \sup \left\{ \int_X f\xi(dx) : f \in \mathcal{F}_b(X, \mathcal{X}), \|f\|_{\sup} = 1 \right\}.$$

(iii) Für $f \in \mathcal{F}_b(X, \mathcal{X})$ ist

$$\|f\|_{\sup} = \sup \left\{ \int_X f\xi(dx) : \xi \in \mathcal{M}(X, \mathcal{X}), \|\xi\|_{TV} = 1 \right\}.$$

Beweis. Übungsblatt 4, Aufgabe 4. □

Definition 2.9 (Dobrushin-Koeffizient)**.** *Seien $(X, \mathcal{X})$ und $(Y, \mathcal{Y})$ messbare Räume und K ein Markovkern von $(X, \mathcal{X})$ nach $(Y, \mathcal{Y})$. Der Dobrushin-Koeffizient $\delta(K)$ von K wird definiert als*

$$\delta(K) := \frac{1}{2} \sup_{(x,x') \in X \times X} \|K(x, \cdot) - K(x', \cdot)\|_{TV} = \sup_{\substack{(x,x') \in X \times X, \\ x \neq x'}} \frac{\|K(x, \cdot) - K(x', \cdot)\|_{TV}}{\|\delta_x - \delta_{x'}\|_{TV}}.$$

Wegen $K(\cdot, X) = 1$ für alle $x \in X$ gilt $0 \leq \delta(K) \leq 1$.

Lemma 2.10. *Seien ξ ein endliches signiertes Maß auf $(X, \mathcal{X})$ und K ein Markovkern von $(X, \mathcal{X})$ nach $(Y, \mathcal{Y})$. Dann gilt*

$$\|\xi K\|_{TV} \leq \delta(K) \|\xi\|_{TV} + (1 - \delta(K)) |\xi(X)|. \tag{2.9}$$

Beweis. Sei $\xi \in \mathcal{M}(X, \mathcal{X})$ mit Hahn-Jordan-Zerlegung $\xi = \xi_+ - \xi_-$. Im Falle $\xi_-(X) = 0$ ist ξ ein Maß und es folgt wegen $0 \leq K(x, A) \leq 1$ für alle $x \in X$ und $A \in \mathcal{Y}$

$$\|\xi K\|_{TV} = \int K(x, X) \xi(dx) \leq \xi(X) = \|\xi\|_{TV}.$$

Dasselbe gilt für $\xi_+(X) = 0$. Seien also nachfolgend sowohl $\xi_+(X) > 0$ sowie $\xi_-(X) > 0$, ohne Einschränkung $\xi_+(X) \geq \xi_-(X)$ (andernfalls ersetze man ξ duch $-\xi$ – das ändert die zu beweisende Ungleichung (2.9) nicht). Der Kürze der Darstellung wegen setzen wir von nun an $\eta g := \int g d\eta$ für $g \in \mathcal{F}_b(X, \mathcal{X})$ und $\eta \in \mathcal{M}(X, \mathcal{X})$. Sei $f : Y \to \mathbb{R}$ eine $\mathcal{Y} - \mathcal{B}(\mathbb{R})$-messbare Funktion mit $\|f\|_{\sup} = 1$. Nach Lemma 2.8 (ii) reicht es,

$$|(\xi K) f| \leq \delta(K)(\xi_+(X) + \xi_-(X)) + (1 - \delta(K)) |\xi_+(X) - \xi_-(X)|$$

zu verifizieren, was wegen $|\xi_+(X) - \xi_-(X)| = \xi_+(X) - \xi_-(X)$ äquivalent ist zu

$$|(\xi K)f| \le 2\delta(K)\xi_-(X) + \xi_+(X) - \xi_-(X). \tag{2.10}$$

Mit

$$\begin{aligned}(\xi K)f &= \xi(Kf)\\ &= \int (Kf)(x)\xi_+(dx) - \int (Kf)(x)\xi_-(dx)\\ &= \frac{\iint (Kf)(x)\xi_+(dx)\xi_-(dx')}{\xi_-(X)} - \frac{\iint (Kf)(x')\xi_+(dx)\xi_-(dx')}{\xi_+(X)}\end{aligned}$$

folgt zunächst

$$\begin{aligned}|(\xi K)f| &\le \iint \Big| \frac{(Kf)(x)}{\xi_-(X)} - \frac{(Kf)(x')}{\xi_+(X)} \Big| \xi_+(dx)\xi_-(dx')\\ &\le \sup_{(x,x')\in X\times X} \Big| \frac{(Kf)(x)}{\xi_-(X)} - \frac{(Kf)(x')}{\xi_+(X)} \Big| \xi_+(X)\xi_-(X)\\ &= \sup_{(x,x')\in X\times X} |\xi_+(X)(Kf)(x) - \xi_-(X)(Kf)(x')|\\ &\le \sup_{(x,x')\in X\times X} \big\| \xi_+(X)K(x,\cdot) - \xi_-(X)K(x',\cdot) \big\|_{TV} \|f\|_{\sup},\end{aligned}$$

wobei in der letzten Ungleichung Lemma 2.8 (i) verwendet wurde. Schließlich ergibt sich (2.10) aus

$$\begin{aligned}&\big\| \xi_+(X)K(x,\cdot) - \xi_-(X)K(x',\cdot) \big\|_{TV}\\ &\qquad \le \xi_-(X)\|K(x,\cdot) - K(x',\cdot)\|_{TV} + |\xi_+(X) - \xi_-(X)|\|K(x,\cdot)\|_{TV}\\ &\qquad = \xi_-(X)\|K(x,\cdot) - K(x',\cdot)\|_{TV} + \xi_+(X) - \xi_-(X)\\ &\qquad \le 2\xi_-(X)\delta(K) + \xi_+(X) - \xi_-(X).\end{aligned}$$

□

Um schärfere Resultate zu erzielen, müssen wir K als einen Operator auf einer kleineren Menge als $\mathcal{M}(X,\mathcal{X})$ betrachten. Von speziellem Nutzen wird sich hier die Teilmenge

$$\mathcal{M}_0(X,\mathcal{X}) := \big\{\xi \in \mathcal{M}(X,\mathcal{X}) : \xi(X) = 0\big\}$$

der signierten Maße ξ mit $\xi(X) = 0$ erweisen.

Korollar 2.11. *Ist K ein Markovkern von $(X,\mathcal{X})$ nach $(Y,\mathcal{Y})$, so gilt*

$$\delta(K) = \sup\big\{\|\xi K\|_{TV} : \xi \in \mathcal{M}_0(X,\mathcal{X}), \|\xi\|_{TV} \le 1\big\}.$$

Beweis. Wegen $\xi(X) = 0$ folgt einerseits die Ungleichung

$$\delta(K) \ge \sup\big\{\|\xi K\|_{TV} : \xi \in \mathcal{M}_0(X,\mathcal{X}), \|\xi\|_{TV} \le 1\big\}$$

aus Lemma 2.10; die andere Richtung ist unmittelbare Konsequenz aus Definition 2.9:

$$\delta(K) = \sup_{(x,x')\in X\times X} \Big\| \frac{1}{2}(\delta_x - \delta_{x'})K \Big\|_{TV} \le \sup_{\substack{\xi\in\mathcal{M}(\mathcal{X},X):\\ \|\xi\|_{TV}\le 1}} \|\xi K\|_{TV}.$$

□

$\mathcal{M}_0(X,\mathcal{X})$ bildet einen reellen Vektorraum, der versehen mit der Totalvariation ein Banachraum ist. Das Korollar zeigt, dass für einen Markovkern K von $(X,\mathcal{X})$ nach $(Y,\mathcal{Y})$ der Dobrushin-Koeffizient $\delta(K)$ gerade die Operatornorm der linearen Abbildung

$$\begin{aligned}\mathcal{M}_0(X,\mathcal{X}) &\to \mathcal{M}_0(Y,\mathcal{Y}) \\ \xi &\mapsto \xi K\end{aligned} \tag{2.11}$$

ist. Dies wiederum impliziert die Submultiplikativität des Dobrushin-Koeffizienten.

Proposition 2.12. *Sind K ein Markovkern von $(X,\mathcal{X})$ nach $(Y,\mathcal{Y})$ sowie R ein Markovkern von $(Y,\mathcal{Y})$ nach $(Z,\mathcal{Z})$, so gilt*

$$\delta(KR) \le \delta(K)\cdot\delta(R).$$

Beweis. Für $\xi \in \mathcal{M}_0(X,\mathcal{X})$ folgt $\xi K(Y) = \int_X K(x,Y)\xi(dx) = \xi(X) = 0$, womit nach Korollar 2.11

$$\|\xi(KR)\|_{TV} = \|(\xi K)R\|_{TV} \le \delta(R)\|\xi K\|_{TV} \le \delta(R)\cdot\delta(K).$$

□

2.2.2 Doeblin-Bedingung und gleichmäßige Ergodizität

Wir betrachten nun Markovketten, deren Übergangskern Q eine der nachfolgenden Bedingungen erfüllt.

Annahme 2.13 (Doeblin-Bedingung)**.** *Es existieren $m \in \mathbb{N}$ und $\varepsilon \in (0,1)$ sowie ein Markovscher Übergangskern μ von $(X\times X, \mathcal{X}\otimes X)$ nach $(X,\mathcal{X})$, so dass für alle $x,x' \in X$ und $A \in \mathcal{X}$ gilt:*

$$\min\left\{Q^m(x,A), Q^m(x',A)\right\} \ge \varepsilon\mu\big((x,x'),A\big).$$

Annahme 2.14 (Starke Doeblin-Bedingung)**.** *Es existieren $m \in \mathbb{N}$ und $\varepsilon \in (0,1)$ sowie ein W-Maß μ auf $(X,\mathcal{X})$, so dass für alle $x \in X$ und $A \in \mathcal{X}$ gilt:*

$$Q^m(x,A) \ge \varepsilon\mu(A).$$

Der Zusammenhang mit dem Dobrushin-Koeffizienten ergibt sich nun wie folgt. Die Totalvariation der Differenz zweier W-Maße $\mathbb{P}$ und $\mathbb{Q}$ auf einem gemeinsamen messbaren Raum $(X,\mathcal{X})$ besitzt nach Aufgabe 1 auf Übungsblatt 5 die Darstellung

$$\|\mathbb{P}-\mathbb{Q}\|_{TV} = \int_X |p-q|dR = 2 - 2\int_X \min(p,q)dR = 2 - 2\inf\sum_i \min\big(\mathbb{P}(A_i),\mathbb{Q}(A_i)\big),$$

wobei R ein $\mathbb{P}$ und $\mathbb{Q}$ dominierendes σ-endliches Maß ist, $p = d\mathbb{P}/dR$, $q = d\mathbb{Q}/dR$ zugehörige Radon-Nikodym-Dichten sind und das Infimum über alle endlichen Zerlegungen von X in $\mathcal{X}$-messbare Mengen A_i läuft. Hiermit kann man den Dobrushin-Koeffizienten $\delta(Q^m)$ ausdrücken als

$$\delta(Q^m) = 1 - \inf_{x,x'\in X}\inf\sum_i \min\big(Q^m(x,A_i), Q^m(x',A_i)\big).$$

Gilt nun die Doeblin-Bedingung, ist die rechte Seite durch

$$\varepsilon\inf_{x,x'\in X}\inf\sum_i \mu((x,x'),A_i) = \varepsilon\inf_{x,x'\in X}\inf\mu\big((x,x'),\cup_i A_i\big) = \varepsilon$$

nach unten beschränkt, womit $\delta(Q^m) \le 1-\varepsilon$. Aber damit ist die Abbildung (2.11) für $K = Q^m$ eine (echte) Kontraktion.

Definition 2.15. *Seien Q ein Markovscher Übergangskern und π ein σ-endliches Maß auf $(X, \mathcal{X})$. Gilt $\pi = \pi Q$, so heißt π invariantes Maß zu Q.*

Ist X nicht endlich, muss ein invariantes Maß nicht existieren und im Falle seiner Existenz ist es nicht notwendig eindeutig ($\rightarrow$ Übungsblatt 5, Aufgabe 2). Ist ein invariantes Maß endlich, kann man es aber immer zu einem invarianten W-Maß normieren.

Ist nun $(X_k)_{k\geq 0}$ eine Markovkette mit Übergangskern Q, zu welchem ein invariantes W-Maß π existiert, und ist π zudem die Startverteilung, so gilt für alle $k \in \mathbb{N}_0$ und jede $\mathcal{X}^{k+1} - \mathcal{B}(\mathbb{R})$-messbare Funktion $f : X^{k+1} \to \mathbb{R}$ gemäß Proposition 1.4

$$\mathbb{E}_\pi f(X_0, \ldots, X_k) = \mathbb{E}_\pi f(X_1, \ldots, X_{k+1}).$$

Damit ist $\mathbb{P}_\pi^{X_0, X_1, \ldots} = \mathbb{P}_\pi^{X_1, X_2, \ldots}$, d.h. die Verteilung der Markovkette auf dem unendlichen Produktraum ist Shift-invariant. Stochastische Prozesse, deren Verteilung diese Shift-Invarianz besitzt, bezeichnet man als stationär; sie sind u.a. Gegenstand der Vorlesung Wahrscheinlichkeitstheorie II - Stochastische Prozesse. Startet die Markovkette in einer beliebigen Startverteilung ν und konvergiert die Folge der Marginalverteilungen $(\nu Q^n)_{n\geq 0}$ in Totalvariation gegen ein Maß μ, muss der Grenzwert μ zwangsläufig ein invariantes Maß sein. Denn für jedes $f \in \mathcal{F}_b(X, \mathcal{X})$ ist

$$\int f d\mu = \lim_{n\to\infty} \int f d(\nu Q^n) = \lim_{n\to\infty} \int Qf d(\nu Q^{n-1}) = \int Qf d\mu = \int f d(\mu Q). \tag{2.12}$$

Das nächste Resultat zeigt, dass ein Markovscher Übergangskern, der der Doeblin-Bedingung genügt, immer ein eindeutiges invariantes W-Maß besitzt.

Satz 2.16. *Seien $(X, \mathcal{X})$ ein messbarer Raum und Q ein Markovscher Übergangskern, der die Doeblin-Bedingung erfüllt. Dann besitzt Q ein eindeutiges invariantes W-Maß.*

Beweis. Wegen der Doeblin-Bedingung existiert ein $m \in \mathbb{N}$ mit $\delta(Q^m) < 1$, womit Abbildung (2.11) mit $K = Q^m$ eine Kontraktion ist. Nach Proposition 2.12 folgt für jedes W-Maß ν

$$\|\nu Q^{km} - \nu Q^{(k+l)m}\|_{TV} \leq \delta(Q^m)^k \|\nu - \nu Q^{lm}\|_{TV} \leq 2\delta(Q^m)^k,$$

womit $(\nu Q^{km})_{k\geq 1}$ eine Cauchyfolge in der abgeschlossenen Teilmenge der W-Maße im Banachraum $(\mathcal{M}(X, \mathcal{X}), \|\cdot\|_{TV})$ ist. Damit ist sie in der Totalvariationsnorm konvergent gegen ein W-Maß π. Wegen (2.12) (mit Q^m statt Q) muss dann $\pi = \pi Q^m$ gelten. Zudem ist der Grenzwert unabhängig von ν, denn $\|\nu Q^{km} - \nu' Q^{km}\|_{TV} \leq 2\delta(Q^m)^k \to 0$ für $k \to \infty$. Somit ist π eindeutiges invariantes W-Maß zu Q^m. Nach der Chapman-Kolmogorov-Gleichung ist aber auch

$$(\pi Q)Q^m = (\pi Q^m)Q = \pi Q,$$

also ist auch πQ invariantes Maß zu Q^m, womit $\pi = \pi Q$. □

Definition 2.17. *Seien $(X, \mathcal{X})$ ein messbarer Raum und Q ein Markovscher Übergangskern darauf, zu dem ein eindeutiges invariantes W-Maß π existiert. Die Markovkette mit Zustandsraum X und Übergangskern Q heißt ergodisch, falls eine Menge $A \in \mathcal{X}$ existiert mit $\pi(A) = 1$, so dass für alle $x \in A$ gilt: $\lim_{n\to\infty} \|Q^n(x, \cdot) - \pi\|_{TV} = 0$. Sie heißt gleichmäßig ergodisch, falls*

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|Q^n(x, \cdot) - \pi\|_{TV} = 0. \tag{2.13}$$

Eine Markovkette, deren Übergangskern Q die Doeblin-Bedingung erfüllt, ist gleichmäßig ergodisch. Denn einerseits existiert ein eindeutiges invariantes W-Maß π nach Satz 2.16, andererseits existiert ein $m \in \mathbb{N}$ mit $\delta(Q^m) < 1$, womit

$$\sup_{x \in X} \|Q^n(x, \cdot) - \pi\|_{TV} = \sup_{x \in X} \|Q^n(x, \cdot) - \pi Q^n\|_{TV} \leq 2\delta(Q^n) \to 0 \quad (n \to \infty)$$

nach der Submultiplikativität des Dobrushin-Koeffizienten.

Bemerkung 2.18. *Gleichmäßige Ergodizität impliziert gleichmäßig geometrische Ergodizität. Denn mit der Dreiecksungleichung folgt aus (2.13), dass $\delta(Q^m) < 1$ für ein $m \in \mathbb{N}$. Aber mit $M := 2\delta(Q^m)^{-1}$ und $\rho := \delta(Q^m)^{1/m}$ ist dann wegen $\pi = \pi Q^n$*

$$\sup_{x \in X} \|Q^n(x, \cdot) - \pi\|_{TV} \leq 2\delta(Q^n) \leq M\rho^n \quad \forall\, n \in \mathbb{N}.$$

2.2.3 Vergessen der Startverteilung unter der Atar-Zeitouni-Bedingung

Man könnte versucht sein zu denken, dass gleichmäßige Ergodizität des Signalprozesses $(X_k)_{k \geq 0}$ ausreicht, um $\|\pi_{\nu,k|N} - \pi_{\nu',k|N}\|_{TV} \to 0$ für $k, N \to \infty$ zu folgern. Dies ist allerdings mitnichten der Fall, wie folgendes Beispiel zeigt.

Beispiel 2.19. *Sei $(U_k)_{k \geq 1}$ eine iid-Folge von Bernoulli-verteilten Zufallsvariablen mit Parameter $p \in (0, 1)$. Sei $(X_k)_{k \geq 0}$ die Markovkette mit Zustandsraum $X = \{0, 1, 2, 3\}$, welche durch die Rekursion $X_k = (X_{k-1} + U_k) \mod 4$ sowie $X_0 \sim \nu$ für ein W-Maß ν auf der Potenzmenge von X gegeben ist. Sei weiter $Y_k := \mathbb{1}_{\{0,2\}}(X_k)$, $k \in \mathbb{N}_0$. Dann gilt:*

(i) $(X_k, Y_k)_{k \geq 0}$ *ist ein HMM,*

(ii) für den Übergangskern Q von $(X_k)_{k \geq 0}$ erfüllt der Dobrushin-Koeffizient $\delta(Q^4) < 1$ und $(X_k)_{k \geq 0}$ ist damit gleichmäßig ergodisch und

(iii) $\|\pi_{\nu,k|k} - \pi_{\nu',k|k}\|_{TV} = \|\pi_{\nu,0|0} - \pi_{\nu',0|0}\|_{TV}$ *mit*

$$\|\pi_{\nu,0|0} - \pi_{\nu',0|0}\|_{TV} = \begin{cases} y_0 \frac{\nu(\{j\})}{\nu(\{0\}) + \nu(\{2\})} & \text{falls } j = 0, 2 \\ (1 - y_0) \frac{\nu(\{j\})}{\nu(\{1\}) + \nu(\{3\})} & \text{falls } j = 1, 3. \end{cases}$$

Beweis: Übungsblatt 6, Aufgabe 2.

Der Sachverhalt ist also deutlich spannender. Es sei an die Definitionen von $\sigma_{k|N}$ in (2.1) und $\beta_{k|N}$ in (2.4) erinnert: $\sigma_{k|N}$ hängt von ν ab, $\beta_{k|N}$ hingegen nicht. In Aufgabe 1, Übungsblatt 3, wurde gezeigt, dass der bedingte Prozess $(X_k)_{k \geq 0} | Y_0, \ldots Y_N$ eine inhomogene Markovkette mit Übergangskern

$$\tilde{Q}_{k|N}(x, A) = \begin{cases} \frac{\int_A \beta_{k+1|N}(x', (y_{k+2}, \ldots, y_N)) \gamma(x', y_{k+1}) Q(x, dx')}{\beta_{k|N}(x, (y_{k+1}, \ldots, y_N)))} & \text{falls } k \leq N \\ Q(x, A) & \text{falls } k > N \end{cases} \tag{2.14}$$

ist, wobei die explizite Abhängigkeit von $(y_{k+1}, \ldots, y_N)$ in der Notation $\tilde{Q}_{k|N}(x, A)$ unterdrückt ist. Der wesentliche Punkt hier ist, dass die Startverteilung ν in dem Ausdruck nicht vorkommt, d.h. die einzige Abhängigkeit in $\pi_{\nu,k|N} = \pi_{\nu,0|N} \prod_{i=1}^{k} \tilde{Q}_{i-1|N}$ von

ν steckt in der bedingten Verteilung $\pi_{\nu,0|N} = \mathbb{P}^{X_0|Y_0,\dots,Y_n}$, die gegeben ist durch

$$\pi_{\nu,0|N}((y_0,\dots,y_N),A) = \frac{\int_A \gamma(x_0,y_0)\beta_{0|N}(x_0,(y_1,\dots,y_N))\nu(dx_0)}{\int_X \gamma(x_0,y_0)\beta_{0|N}(x_0,(y_1,\dots,y_N))\nu(dx_0)}.$$

Damit ergibt Lemma 2.10

$$\|\pi_{\nu,k|N} - \pi_{\nu',k|N}\|_{TV} \le \delta\Big(\prod_{i=1}^{k}\tilde{Q}_{i-1|N}\Big)\|\pi_{\nu,0|N} - \pi_{\nu',0|N}\|_{TV}. \tag{2.15}$$

Das zeigt bereits, dass für jedes $k \in \mathbb{N}$ der Abstand der beiden bedingten Verteilungen $\|\pi_{\nu,k|N} - \pi_{\nu',k|N}\|_{TV}$ immer durch $\|\pi_{\nu,0|N} - \pi_{\nu',0|N}\|_{TV}$ beschränkt bleibt.

Wir werden nachfolgend Bedingungen formulieren, welche erlauben, nicht-triviale obere Schranken an den Dobrushin-Koeffizienten $\delta(\prod_{i=1}^{k}\tilde{Q}_{i-1|N})$ abzuleiten. Wie gehabt bezeichnet Q den Übergangskern des Signalprozesses $(X_k)_{k\ge 0}$.

Annahme 2.20. *Es existieren $\mathcal{Y}-\mathcal{B}(\mathbb{R})$-messbare Funktionen $\zeta^-,\zeta^+ : Y \to (0,\infty)$ sowie ein Markovkern K von $(Y,\mathcal{Y})$ nach $(X,\mathcal{X})$, so dass*

$$\zeta^-(y)K(y,A) \le \int_A \gamma(x',y)Q(x,dx') \le \zeta^+(y)K(y,A) \quad \forall\, A\in\mathcal{X}, \forall x\in X \text{ und } \forall\, y\in Y.$$

Proposition 2.21. *Unter Annahme 2.20 gilt für alle $k,N\in\mathbb{N}$ und für alle W-Maße ν,ν' auf $(X,\mathcal{X})$*

$$\|\pi_{\nu,k|N}-\pi_{\nu',k|N}\|_{TV} \le \left[\Big(1-\int\zeta^-d\phi\Big)^{k-\min(k,N)}\prod_{j=1}^{\min(k,N)}\Big(1-\frac{\zeta^-(y_j)}{\zeta^+(y_j)}\Big)\right]\|\pi_{\nu,0|N}-\pi_{\nu',0|N}\|_{TV}.$$

Beweis. Wegen (2.15) reicht es zu zeigen, dass

$$\delta\Big(\prod_{i=1}^{k}\tilde{Q}_{i-1|N}\Big) \le \Big(1-\int\zeta^-d\phi\Big)^{k-\min(k,N)}\prod_{j=1}^{\min(k,N)}\Big(1-\frac{\zeta^-(y_j)}{\zeta^+(y_j)}\Big).$$

Sei zunächst $k<N$. Mit $\zeta^-(y_j) \le \int\gamma(x',y_j)Q(x,dx') \le \zeta^+(y_j)\ \forall\, x\in X$ folgt aus (2.4)

$$\prod_{j=k+1}^{N}\zeta^-(y_j) \le \beta_{k|N}(x,(y_{k+1},\dots,y_N)) \le \prod_{j=k+1}^{N}\zeta^+(y_j).$$

Rückwärtsrekursion (2.6) und Annahme 2.20 ergeben für jedes W-Maß κ auf $(X,\mathcal{X})$

$$\begin{aligned}
&\int_X \beta_{k|N}(x,(y_{k+1},\dots,y_N))\kappa(dx)\\
&\quad= \int_X\int_X \beta_{k+1|N}(x_{k+1},(y_{k+2},\dots,y_N))\gamma(x_{k+1},y_{k+1})Q(x,dx_{k+1})\kappa(dx)\\
&\quad\le \zeta^+(y_{k+1})\int_X \beta_{k+1|N}(x_{k+1},(y_{k+2},\dots,y_N))K(y_{k+1},dx_{k+1})
\end{aligned}$$

sowie die analoge untere Schranke

$$\geq \zeta^-(y_{k+1}) \int_X \beta_{k+1|N}(x_{k+1}, (y_{k+2}, \ldots, y_N)) K(y_{k+1}, dx_{k+1}).$$

Mithilfe der der Darstellung

$$\tilde{Q}_{k|N}(x, A) = \frac{\int_A \beta_{k+1|N}(x'), (y_{k+2} \ldots, y_N) \gamma(x', y_{k+1}) Q(x, dx')}{\int_X \beta_{k+1|N}(x', (y_{k+2}, \ldots, y_N)) \gamma(x', y_{k+1}) Q(x, dx')}$$

folgt daraus

$$\frac{\zeta^-(y_{k+1})}{\zeta^+(y_{k+1})} \lambda_{k,n}((y_{k+1}, \ldots, y_N), A) \leq \tilde{Q}_{k|N}(x, A) \leq \frac{\zeta^+(y_{k+1})}{\zeta^-(y_{k+1})} \lambda_{k,n}((y_{k+1}, \ldots, y_N), A)$$

für den Markovkern $\lambda_{k,N} : Y^{N-k} \times \mathcal{X} \to [0, 1]$, gegeben durch

$$\lambda_{k,n}((y_{k+1}, \ldots, y_N), A) := \frac{\int_A \beta_{k+1|N}(x_{k+1}, (y_{k+2}, \ldots, y_N)) K(y_{k+1}, dx_{k+1})}{\int_X \beta_{k+1|N}(x_{k+1}, (y_{k+2}, \ldots, y_N)) K(y_{k+1}, dx_{k+1})}$$

für $A \in \mathcal{X}$ und $y_{k+1}, \ldots, y_N \in Y$. Aber für festes $y_{k+1}, \ldots, y_N$ erfüllt $\tilde{Q}_{k|N}$ damit die starke Doeblin-Bedingung, womit

$$\delta(\tilde{Q}_{k|N}) \leq 1 - \frac{\zeta^-(y_{k+1})}{\zeta^+(y_{k+1})}.$$

Im Falle $k > N$ gilt $\tilde{Q}_k = Q$. Wegen $\int_Y \gamma(x, y) \phi(dy) = 1$ ist

$$Q(x, A) = \int_A \int_Y \gamma(x', y) \phi(dy) Q(x, dx') = \int_Y \int_A \gamma(x', y) Q(x, dx') \phi(dy)$$

nach dem Satz von Fubini, wobei ϕ das W-Maß aus der Nicht-Degeneriertheitsannahme ist. Dies ergibt mit mit Annahme 2.20 die untere Schranke

$$Q(x, A) \geq \int_Y \zeta^-(y) K(y, A) \phi(dy) = \int_Y \zeta^-(y) \phi(dy) \frac{\int_Y \zeta^-(y) K(y, A) \phi(dy)}{\int_Y \zeta^-(y) \phi(dy)}$$

und damit ebenfalls die starke Doeblin-Bedingung. Zusammen mit der Submultiplikativität des Dobrushin-Koeffizienten aus Proposition 2.12 folgt dann die Behauptung. □

Mitunter ist eine verschärfte Variante von Bedingung 2.20 gerechtfertigt.

Annahme 2.22 (Atar und Zeitouni 1997)**.** *Es existieren $\varepsilon > 0$ und ein W-Maß κ auf $(X, \mathcal{X})$, so dass*

$$\varepsilon \kappa(A) \leq Q(x, A) \leq \frac{1}{\varepsilon} \kappa(A) \quad \forall\, A \in \mathcal{X} \text{ und } \forall\, x \in X. \tag{2.16}$$

Annahme 2.22 impliziert Annahme 2.20 mit

$$\zeta^-(y) = \varepsilon \int_X \gamma(x, y) \kappa(dx), \ \zeta^+(y) = \frac{1}{\varepsilon} \int_X \gamma(x, y) \kappa(dx) \text{ und } K(y, A) = \frac{\int_A \gamma(x, y) \kappa(dx)}{\int_X \gamma(x, y) \kappa(dx)}.$$

Unter Annahme 2.22 verschärft sich die Aussage aus Proposition 2.21 zu

$$\|\pi_{\nu,k|N} - \pi_{\nu',k|N}\|_{TV} \leq (1-\varepsilon^2)^{\min(k,N)}(1-\varepsilon)^{k-\min(k,N)}\|\pi_{\nu,0|N} - \pi_{\nu',0|N}\|_{TV}. \quad (2.17)$$

Die bedingte Verteilung $\pi_{\nu,k|N}$ vergisst die Startbedingung $\mathbb{P}^{X_0} = \nu$ exponentiell schnell.

Vergleichen wir die Atar-Zeitouni-Bedingung mit der Doeblin-Bedingung, springt ins Auge, dass die linke Ungleichung in (2.16) gerade die starke Doeblin-Bedingung im Spezialfall $m = 1$ ist. Das legt nahe, dass die Atar-Zeitouni-Bedingung dahingehend abgeschwächt werden kann, dass lediglich für ein $m \in \mathbb{N}$ die beiden Ungleichungen (2.16) für Q^m anstelle von Q gefordert werden. Unter einer zusätzlichen Bedingung an die ϕ-Dichte γ des Beobachtungskerns G ist das auch tatsächlich der Fall (Übungsblatt 7, Aufgabe 1).

2.3 Sequentielle Monte-Carlo-Approximationen

In Abschnitt 2.1 wurde das Filterproblem grundsätzlich für nicht-degenerierte HMMs gelöst. Im Hinblick auf ihre Berechnung ist die (im Allgemeinen unendlichdimensionale) Filterrekursion allerdings mit beträchtlichen Schwierigkeiten verbunden. Das suggeriert, diese Rekursionen durch sogenannte Monte-Carlo-Verfahren zu approximieren.

2.3.1 Sequentielles Importance Sampling

Die Idee hinter Monte-Carlo-Approximationen ist einfach – wir entwickeln sie exemplarisch für den Fall der Filterrekursion. Per definitionem gilt für jede beschränkte, messbare Funktion $f : X \to \mathbb{R}$ mit dem Kern σ_k aus (2.1)

$$\int f(x)\sigma_k((y_0,\dots,y_k),dx) = \mathbb{E}\big(f(X_k)\gamma(X_0,y_0)\cdot\ldots\cdot\gamma(X_k,y_k)\big).$$

Angenommen, wir können eine iid-Stichprobe $X^{(1)},\dots,X^{(n)}$ von $\mathbb{P}^{(X_0,\dots,X_k)}$ generieren, d.h. den Signalprozess simulieren. Dann gilt nach Satz 2.2 sowie dem starken Gesetz der großen Zahlen

$$\frac{\sum_{i=1}^n f(X_k)\gamma(X_0^{(i)},y_0)\cdot\ldots\cdot\gamma(X_k^{(i)},y_k)}{\sum_{i=1}^n \gamma(X_0^{(i)},y_0)\cdot\ldots\cdot\gamma(X_k^{(i)},y_k)} \longrightarrow \int f(x)\pi_{k|k}((y_0,\dots,y_k),dx) \quad \text{f.s.}$$

für $n \to \infty$. Also ist das Mittel $\sum_{i=1}^n w_k^{(i)} f(X_k^{(i)})$ mit den Gewichten

$$w_k^{(i)} = \frac{\gamma(X_0^{(i)},y_0)\cdot\ldots\cdot\gamma(X_k^{(i)},y_k)}{\sum_{j=1}^n \gamma(X_0^{(j)},y_0)\cdot\ldots\cdot\gamma(X_k^{(j)},y_k)} \quad (2.18)$$

eine Approximation an $\int f(x)\pi_{k|k}((y_0,\dots,y_k),dx)$.

Bemerkung 2.23. *Die Gewichte summieren sich zu* 1 *und können daher als W-Gewichte interpretiert werden. Wenngleich jeder Stichprobenpfad*

$$x^{(i)} = (x_0^{(i)},\dots,x_k^{(i)})$$

nach Konstruktion gleichwahrscheinlich ist mit W't $1/n$*, wird er bei der Berechnung des Filters durch die beobachtungsabhängigen Gewichte* (2.18) *neu gewichtet. Also modifizieren die Beobachtungen* $Y_0,\dots,Y_k$ *die relative Wichtigkeit für jeden der Pfade – der Grund, warum man von 'Importance Sampling' spricht.*

Für die Berechnung ist nun attraktiv, dass, genau wie die Filterverteilungen, sowohl die Stichprobe als auch die Gewichte rekursiv ermittelt werden können, was zum sogenannten sequentiellen Importance Sampling (kurz SIS) führt:

- Erzeuge $X_0^{(i)}$, $i = 1, \dots, n$, iid nach der Startverteilung ν.
- Berechne $w_0^{(i)} = \frac{\gamma(X_0^{(i)}, y_0)}{\sum_{j=1}^n \gamma(X_0^{(j)}, y_0)}$.
- Für $l = 1, \dots, k$
  - erzeuge $X_l^{(i)}$, $i = 1, \dots, n$, nach $Q(X_{l-1}^{(i)}, \cdot)$ und
  - berechne $w_l^{(i)} = \frac{w_{l-1}^{(i)} \gamma(X_l^{(i)}, y_l)}{\sum_{j=1}^n w_{l-1}^{(j)} \gamma(X_l^{(j)}, y_l)}$.
- Berechne die Filterapproximation $\sum_{i=1}^n w_k^{(i)} f(X_k^{(i)})$.

Das soeben beschriebene sequentielle Importance Sampling hat allerdings einen beachtlichen Haken, wenn die beiden Verteilungen $\pi_{k|k}$ und $\mathbb{P}^{X_k}$ nicht nahe aneinander liegen.

> Sind beispielsweise $Y_0, \dots, Y_k$ eine nahezu unverrauschte Wiedergabe des Signals $X_0, \dots, X_k$, so konzentriert sich die bedingte Verteilung $\pi_{k|k}((y_0, \dots, y_k), \cdot)$ sehr stark um das tatsächlich realisierte Signal $(X_0, \dots X_k) = (x_0, \dots, x_k)$. Erzeugt man nun eine iid-Stichprobe $X^{(1)}, \dots, X^{(n)} \sim \mathbb{P}^{(X_0, \dots, X_k)}$, liegt aber nur ein kleiner Anteil der erzeugten Pfade nahe an $(x_0, \dots, x_k)$.

Das führt dann dazu, dass die Gewichte $w_k^{(i)}$ für alle übrigen Pfadrealisierungen $x^{(i)}$ nahezu Null sind und praktisch nicht zur Filterapproximation beitragen. Der erwartete Approximationsfehler

$$\mathbb{E}\left| \int f(x) \pi_{k|k}((y_0, \dots, y_k), dx) - \sum_{i=1}^n w_k^{(i)} f(X_k^{(i)}) \right|$$

wächst in k entsprechend sehr schnell. Um dieser Problematik entgegenzuwirken, muss man die Strategie zum Erzeugen der Stichprobe $X^{(1)}, \dots, X^{(n)}$ ändern.

2.3.2 Interacting Particles – Importance Sampling Resampling

Nach Satz 2.7 ist $\pi_{l|l} Q = \pi_{l+1|l}$, womit die Filterrekursion aus Bemerkung 2.3 geschrieben werden kann als

$$\begin{aligned} \pi_{l|l}((y_0, \dots, y_l), A) &= \frac{\int \mathbb{1}_A(x) \gamma(x, y_l) Q(x', dx) \pi_{l-1|l-1}((y_0, \dots, y_{l-1}), dx')}{\int \gamma(x, y_l) Q(x', dx) \pi_{l-1|l-1}((y_0, \dots, y_{l-1}), dx')} \\ &= \frac{\int \mathbb{1}_A(x) \gamma(x, y_l) (\pi_{l-1|l-1} Q)((y_0, \dots, y_{l-1}), dx)}{\int \gamma(x, y_l) (\pi_{l-1|l-1} Q)((y_0, \dots, y_{l-1}), dx)} \\ &= \frac{\int \mathbb{1}_A(x) \gamma(x, y_l) \pi_{l|l-1}((y_0, \dots, y_{l-1}), dx)}{\int \gamma(x, y_l) \pi_{l|l-1}((y_0, \dots, y_{l-1}), dx)}. \end{aligned} \tag{2.19}$$

Damit lässt sich die Filterrekursion in zwei aufeinanderfolgenden Schritten realisieren:

$$\pi_{l-1|l-1} \overset{\text{Vorhersageschritt}}{\longrightarrow} \pi_{l|l-1} \overset{\text{“Korrektur”}}{\longrightarrow} \pi_{l|l}.$$

Angenommen, wir können eine iid-Stichprobe

$$X_{l|l}^{(1)}, \dots, X_{l|l}^{(n)} \overset{iid}{\sim} \pi_{l|l}((y_0, \dots, y_l), \cdot) \tag{2.20}$$

erzeugen und verfahren dann wie beim SIS-Algorithmus:

> Wir generieren eine iid-Stichprobe $X_{l+1|l}^{(i)} \sim Q(X_{l|l}^{(i)}, \cdot)$, $i = 1, \dots, n$. Nach Satz 2.7 ist dies gerade eine iid-Stichprobe gemäß der Verteilung $\pi_{l+1|l}((y_0, \dots, y_l), \cdot)$.
>
> Anschließend berechnen wir die Gewichte $w_{l+1}^{(i)} = \frac{\gamma(X_{l+1|l}^{(i)}, y_{l+1})}{\sum_{j=1}^n \gamma(X_{l+1|l}^{(j)}, y_{l+1})}$.

Dann gilt für jede beschränkte, messbare Funktion $f : X \to \mathbb{R}$ nach (2.19) sowie dem starken Gesetz der großen Zahlen

$$\sum_{i=1}^n w_{l+1}^{(i)} f\big(X_{l+1|k}^{(i)}\big) \longrightarrow \int f(x) \pi_{l+1|l+1}((y_0, \dots, y_{l+1}), dx) \quad \text{f.s.}$$

für $n \to \infty$, d.h. die Filterverteilung $\pi_{l+1|l+1}$ wird hier approximiert durch das empirische Maß

$$\widetilde{\pi}_{l+1|l+1} := \sum_{i=1}^n w_{l+1}^{(i)} \delta_{X_{l+1|l}^{(i)}}. \tag{2.21}$$

Im SIS-Algorithmus aus Abschnitt 2.3.1 würden wir nun unabhängige $Q(X_{l+1|l}^{(i)}, \cdot)$-verteilte Zufallsvariablen erzeugen, $i = 1, \dots, n$, und der Rekursion entsprechend die Gewichte aktualisieren. Wegen (2.20) ist unsere Stichprobe aber nicht nach $\mathbb{P}^{X_{l+1}}$-verteilt, sondern gemäß Satz 2.7 nach $\pi_{l+1|l}$. Entsprechend kann man eine neue Iteration nun dadurch beginnen, dass man eine iid-Stichprobe aus der Approximation $\widetilde{\pi}_{l+1|l+1}$ in (2.21) der Filterverteilung $\pi_{l+1|l+1}$ generiert. Diese Idee liefert das sequentielle Importance-Sampling-Resampling-Schema (kurz SISR):

- Erzeuge $\tilde{X}_0^{(i)}$, $i = 1, \dots, n$, iid nach der Startverteilung ν.
- Berechne $w_0^{(i)} = \frac{\gamma(\tilde{X}_0^{(i)}, y_0)}{\sum_{j=1}^n \gamma(\tilde{X}_0^{(j)}, y_0)}$.
- Erzeuge $X_0^{(i)}$, $i = 1, \dots, n$, iid nach $\sum_{j=1}^n w_0^{(j)} \delta_{\tilde{X}_0^{(j)}}$ (gegeben $\tilde{X}_0^{(1)}, \dots, \tilde{X}_0^{(n)}$).
- Für $l = 1, \dots, k$
  - erzeuge $\tilde{X}_l^{(i)}$, $i = 1, \dots, n$, nach $Q(X_{l-1}^{(i)}, \cdot)$,
  - berechne $w_l^{(i)} = \frac{w_{l-1}^{(i)} \gamma(\tilde{X}_l^{(i)}, y_l)}{\sum_{j=1}^n w_{l-1}^{(j)} \gamma(\tilde{X}_l^{(j)}, y_l)}$ und
  - erzeuge $X_l^{(i)}$, $i = 1, \dots, n$, iid nach $\sum_{j=1}^n w_l^{(j)} \delta_{\tilde{X}_l^{(j)}}$ (gegeben $\tilde{X}_l^{(1)}, \dots, \tilde{X}_l^{(n)}$).
- Berechne die Filterapproximation $\frac{1}{n} \sum_{i=1}^n f(X_k^{(i)})$.

Gegeben $Y_0, \dots, Y_k$ taucht im "Resample" $X_l^{(1)}, \dots, X_l^{(n)}$ eine Realisierung von $\tilde{X}_l^{(i)}$ mit geringem Gewicht seltener auf als eine mit großem Gewicht – letztere kann natürlich beim Resampling auch mehrfach realisiert werden. Das am Ende von Abschnitt 2.3.1 aufgeführte Problem des SIS-Algorithmus wäre mit diesem neuen SISR erst einmal behoben.

Bemerkung 2.24. *Wohingegen die Konvergenz des SIS eine unmittelbare Anwendung des starken Gesetzes der großen Zahlen ist, ist beim SISR zu konstatieren, dass die Pfade*

$$(X_0^{(i)}, \ldots, X_k^{(i)}), \quad i = 1, \ldots, n,$$

nicht mehr stochastisch unabhängig sind. Sie "interagieren" miteinander im Resampling-Schritt, weshalb man bei der Filterapproximation vom 'Interacting Particle Filter' spricht.

Die mathematische Analyse des SISR-Algorithmus gestaltet sich entsprechend etwas anspruchsvoller und ist Gegenstand des nächsten Abschnitts.

2.3.3 Konvergenzanalyse des SISR-Algorithmus

Satz 2.25. *Sei $(Y_0, \ldots, Y_k) = (y_0, \ldots, y_k)$ eine Realisierung des Beobachtungssegments. Angenommen,*

$$\sup_{x \in X} \gamma(x, y_l) < \infty$$

für alle $l = 1, \ldots, k$. Sei $X_k^{(1)}, \ldots, X_k^{(n)}$ die im SISR erzeugte Stichprobe nach der k-ten Iteration. Dann gilt für eine nicht-negative Funktion $C_k : \mathbb{R}^k \to \mathbb{R}$

$$\sup_{f \in \mathcal{F}_1} \mathbb{E}\left[\left(\int f(x)\pi_{k|k}((y_0, \ldots, y_n), dx) - \frac{1}{n}\sum_{i=1}^{n} f(X_k^{(i)})\right)^2\right] \leq \frac{C_k(y_0, \ldots, y_k)}{n}.$$

wobei $\mathcal{F}_1$ die Menge der messbaren Funktionen $f : (X, \mathcal{X}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit $\|f\|_{\sup} \leq 1$ bezeichnet.

Sei

$$\widehat{\pi}_{l|l} := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_l^{(i)}}, \quad l \in \{1, \ldots, k\},$$

das empirische Maß aus den in der l-ten SISR-Iteration erzeugten $X_l^{(i)}$, $i = 1, \ldots, n$. Mit $\widetilde{\pi}_{l|l}$ aus (2.21) besteht jede einzelne SISR-Iteration dann aus folgender Sequenz:

$$\widehat{\pi}_{l|l} \overset{\text{Vorhersage}}{\longrightarrow} \widehat{\pi}_{l+1|l} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\tilde{X}_{l+1}^{(i)}} \overset{\text{Korrektur}}{\longrightarrow} \widetilde{\pi}_{l+1|l+1} = \sum_{i=1}^{n} w_{l+1}^{(i)} \delta_{\tilde{X}_{l+1}^{(i)}} \overset{\text{Resampling}}{\longrightarrow} \widehat{\pi}_{l+1|l+1}.$$

Der Beweis des Satzes besteht nun darin, den Approximationsfehler eines jeden einzelnen dieser Schritte separat abzuschätzen.

Beweis. Nachfolgend sei $(y_0, \ldots, y_k)$ fest; entsprechend sind alle Erwartungswerte bedingt auf $(Y_0, \ldots, Y_k) = (y_0, \ldots, y_k)$ zu verstehen und explizite Abhängigkeiten von $(y_0, \ldots, y_k)$ werden in der Notation unterdrückt. Insbesondere schreiben wir $\gamma_k(x)$ für $\gamma(x, y_k)$.

(i) Resampling-Fehler: Da $X_{l+1}^{(1)}, \ldots, X_{l+1}^{(n)}$ gegeben $\tilde{X}_{l+1}^{(1)}, \ldots, \tilde{X}_{l+1}^{(n)}$ bedingt unabhängig identisch nach $\widetilde{\pi}_{l+1|l+1}$ verteilt sind, ist für $f \in \mathcal{F}_1$

$$\mathbb{E}\left[\left(\int f(x)\widetilde{\pi}_{l+1|l+1}(dx) - \frac{1}{n}\sum_{i=1}^{n} f(X_{l+1}^{(i)})\right)^2 \,\middle|\, \tilde{X}_{l+1}^{(1)}, \ldots, \tilde{X}_{l+1}^{n)}\right]$$
$$= \frac{1}{n}\left(\int f(x)^2 \pi_{l+1|l+1}(dx) - \left(\int f(x)\widetilde{\pi}_{l+1|l+1}(dx)\right)^2\right) \leq \frac{\|f\|_{\sup}^2}{n}.$$

Folglich ist nach der Dreiecksungleichung für die $L_2(\mathbb{P})$-Norm

$$\sup_{f\in\mathcal{F}_1}\mathbb{E}^{1/2}\Big[\Big(\int f(x)\pi_{l+1|l+1}(dx)-\int f(x)\widehat{\pi}_{l+1|l+1}(dx)\Big)^2\Big]$$
$$\leq \frac{1}{\sqrt{n}}+\sup_{f\in\mathcal{F}_1}\mathbb{E}^{1/2}\Big[\Big(\int f(x)\widetilde{\pi}_{l+1|l+1}(dx)-\int f(x)\pi_{l+1|l+1}(dx)\Big)^2\Big].$$

(ii) Korrektur-Fehler: Nach (2.19) und Definition des SISR-Schemas gelten

$$\pi_{l+1|l+1}(dx)=\frac{\gamma_{l+1}(x)\pi_{l+1|l}(dx)}{\int_X\gamma_{l+1}(x)\pi_{l+1|l}(dx)}\quad\text{sowie}\quad\widetilde{\pi}_{l+1|l+1}(dx)=\frac{\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)}{\int_X\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)}.$$

Nach Nullergänzung mit $\int f(x)\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)|/\int\gamma_{l+1}(x)\pi_{l+1|l}(dx)$ folgt

$$\Big|\int f(x)\widetilde{\pi}_{l+1|l+1}(dx)-\int f(x)\pi_{l+1|l+1}(dx)\Big|$$
$$\leq\frac{|\int f(x)\gamma_{l+1}(x)\pi_{l+1|l}(dx)-\int f(x)\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)|}{\int\gamma_{l+1}(x)\pi_{l+1|l}(dx)}$$
$$+\Big|\frac{\int f(x)\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)}{\int\gamma_{l+1}(x)\pi_{l+1|l}(dx)}-\frac{\int f(x)\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)}{\int\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)}\Big|$$
$$\leq\frac{|\int f(x)\gamma_{l+1}(x)\pi_{l+1|l}(dx)-\int f(x)\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)|}{\int\gamma_{l+1}(x)\pi_{l+1|l}(dx)}$$
$$+\frac{|\int f(x)\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)|}{\int\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)}\frac{\int\gamma_{l+1}(x)\widehat{\pi}_{l+1|l}(dx)-\int\gamma_{l+1}(x)\pi_{l+1|l}(dx)}{\int\gamma_{l+1}(x)\pi_{l+1|l}(dx)}\Big|.$$

Mit $f_1(x):=f(x)\gamma_{l+1}(x)/\|f\gamma_{l+1}\|_{\sup}$ und $f_2(x):=\gamma_{l+1}(x)/\|\gamma_{l+1}\|_{\sup}$ ist der letzte Ausdruck aber nach oben durch

$$\frac{\|f\|_{\sup}\|\gamma_{l+1}\|_{\sup}}{\int\gamma_{l+1}(x)\pi_{l+1|l}(dx)}\Big|\int f_1(x)\pi_{l+1|l}(dx)-\int f_1(x)\widehat{\pi}_{l+1|l}(dx)\Big|$$
$$+\frac{\|f\|_{\sup}\|\gamma_{l+1}\|_{\sup}}{\int\gamma_{l+1}(x)\pi_{l+1|l}(dx)}\Big|\int f_2(x)\pi_{l+1|l}(dx)-\int f_2(x)\widehat{\pi}_{l+1|l}(dx)\Big|$$

beschränkt. Nach Konstruktion gilt $\|f_j\|_{\sup}\leq 1$ für $j=1,2$, womit schließlich

$$\sup_{f\in\mathcal{F}_1}\mathbb{E}^{1/2}\Big[\Big(\int f(x)\widetilde{\pi}_{l+1|l+1}(dx)-\int f(x)\pi_{l+1|l+1}(dx)\Big)^2\Big]$$
$$\leq\frac{2\|\gamma_{l+1}\|_{\sup}}{\int\gamma_{l+1}(x)\pi_{l+1|l}(dx)}\sup_{f\in\mathcal{F}_1}\mathbb{E}^{1/2}\Big[\Big(\int f(x)\pi_{l+1|l}(dx)-\int f(x)\widehat{\pi}_{l+1|l}(dx)\Big)^2\Big].$$

(iii) Vorhersage-Fehler: Bedingt auf $X_l^{(1)},\ldots,X_l^{(n)}$ sind $\tilde{X}_{l+1}^{(i)}\sim Q(X_l^{(i)},.)$ unabhängig, womit

$$\mathbb{E}\Big[\Big|\int f(x)\widehat{\pi}_{l+1|l}(dx)-\int f(x)(\widehat{\pi}_{l|l}Q)(dx)\Big|^2\Big]$$
$$=\mathbb{E}\mathbb{E}\Big(\Big|\int f(x)\Big(\frac{1}{n}\sum_{i=1}^n\delta_{\tilde{X}_{l+1}^{(i)}}\Big)(dx)-\int f(x)(\widehat{\pi}_{l|l}Q)(dx)\Big|^2\Big|X_l^{(1)},\ldots,X_l^{(n)}\Big)\leq\frac{\|f\|_{\sup}^2}{n}$$

für $f \in \mathcal{F}_1$. Nach der Dreiecksungleichung, der Identität $\pi_{l+1|l} = \pi_{l|l}Q$, dem verallgemeinerten Satz von Fubini für Markovkerne sowie $\|Qf\|_{\sup} \leq \|f\|_{\sup}$ folgt schließlich

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_1} \mathbb{E}^{1/2}\left[\left|\int f(x)\widehat{\pi}_{l+1|l}(dx) - \int f(x)\pi_{l+1|l}(dx)\right|^2\right] \\
&\quad\leq \frac{1}{\sqrt{n}} + \sup_{f\in\mathcal{F}_1} \mathbb{E}^{1/2}\left[\left|\int f(x)(\widehat{\pi}_{l|l}Q)(dx) - \int f(x)\pi_{l+1|l}(dx)\right|^2\right] \\
&\quad= \frac{1}{\sqrt{n}} + \sup_{f\in\mathcal{F}_1} \mathbb{E}^{1/2}\left[\left|\int (Qf)(x)\widehat{\pi}_{l|l}(dx) - \int (Qf)(x)\pi_{l|l}(dx)\right|^2\right] \\
&\quad\leq \frac{1}{\sqrt{n}} + \sup_{f\in\mathcal{F}_1} \mathbb{E}^{1/2}\left[\left|\int f(x)\widehat{\pi}_{l|l}(dx) - \int f(x)\pi_{l|l}(dx)\right|^2\right].
\end{aligned}$$

Zusammensetzen der drei Abschätzungsschritte aus (i) – (iii) ergibt für den Fehler einer einzelnen SISR-Iteration

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_1} \mathbb{E}^{1/2}\left[\left(\int f(x)\pi_{l+1|l+1}(dx) - \int f(x)\widehat{\pi}_{l+1|l+1}(dx)\right)^2\right] \\
&\quad\leq \frac{1+D_{l+1}}{\sqrt{n}} + D_{l+1} \sup_{f\in\mathcal{F}_1} \mathbb{E}^{1/2}\left[\left|\int f(x)\widehat{\pi}_{l|l}(dx) - \int f(x)\pi_{l|l}(dx)\right|^2\right]
\end{aligned}$$

mit

$$D_{l+1} = \frac{2\|\gamma_{l+1}\|_{\sup}}{\int \gamma_{l+1}(x)\pi_{l+1|l}(dx)}.$$

Durch Iteration ergibt sich entsprechend

$$\sup_{f\in\mathcal{F}_1} \mathbb{E}^{1/2}\left[\left|\int f(x)\widehat{\pi}_{k|k}(dx) - \int f(x)\pi_{k|k}(dx)\right|^2\right] \leq \frac{1}{\sqrt{n}} \sum_{l=0}^{k} \left((1+D_l)\prod_{j=l+1}^{k} D_j\right),$$

sofern noch gezeigt wird, dass

$$\sup_{f\in\mathcal{F}_1} \mathbb{E}^{1/2}\left[\left|\int f(x)\widehat{\pi}_{0|0}(dx) - \int f(x)\pi_{0|0}(dx)\right|^2\right] \leq \frac{1+D_0}{\sqrt{n}}$$

ist, wobei

$$D_0 = \frac{2\|\gamma_0\|_{\sup}}{\int \gamma_0(x)\nu(dx)}.$$

Aber das folgt analog zur obigen Abschätzung eines einzelnen Iterationsschrittes (Übungsblatt 9, Aufgabe 1). □

Gleichmäßige SISR-Konvergenz unter Mischungsbedingungen

In vielen Anwendungen wie beispielsweise der Zielverfolgung (target tracking) möchte man eine Filterapproximation $\int f d\widehat{\pi}_{k|k}$ basierend auf Beobachtungen $Y_0, \ldots, Y_k$ nicht nur zu einem festen Zeitpunkt k: Verfolgt man ein sich bewegendes Objekt und repräsentiert das Signal X_k dessen Ort zum Zeitpunkt k, ist es wünschenswert, bei neu eingehenden Beobachtungen $Y_{k+1}, Y_{k+2}, \ldots$ die Lokation kontinuierlich zu aktualisieren.

Die Fehlerschranke aus Satz 2.25 rechtfertigt ein solches Vorgehen allerdings noch nicht, denn wegen

$$D_{l+1} \geq \frac{2\|\gamma_{l+1}\|_{\sup}}{\int \|\gamma_{l+1}\|_{\sup} \pi_{l+1|l}(dx)} = 2$$

ist

$$C_k(y_0, \dots, y_k) = \sum_{l=0}^{k} \left((1 + D_l) \prod_{j=l+1}^{k} D_j \right) \geq \sum_{l=0}^{k} 3 \cdot 2^{k-l} \geq 2^k.$$

Wir zeigen schließlich noch eine in k gleichmäßige Schranke unter der in Abschnitt 2.2.3 eingeführten Atar-Zeitouni-Bedingung 2.22 und einer analogen Voraussetzung an die Beobachtungsdichte γ, die in Aufgabe 1 auf Blatt 7 schon einmal verwendet wurde. Diese Bedingungen werden auch kurz als Mischungsbedingungen bezeichnet.

Satz 2.26. *Angenommen, Q erfülle die Atar-Zeitouni-Bedingung 2.22 und γ sei beschränkt sowie gleichmäßig von Null weg beschränkt. Seien $(y_0, y_1, \dots)$ eine Realisierung des Beobachtungsprozesses $(Y_k)_{k\geq 0}$ und*

$$\left(X_k^{(1)}, \dots, X_k^{(n)}\right)_{k\geq 0},$$

die jeweils im SISR erzeugte Stichprobe. Dann gilt für alle $n \in \mathbb{N}$

$$\sup_{k\geq 0} \sup_{f \in \mathcal{F}_1} \mathbb{E}\left[\left(\int f(x) \pi_{k|k}((y_0, \dots, y_k), dx) - \frac{1}{n} \sum_{i=1}^{n} f(X_k^{(i)}) \right)^2 \right] \leq \frac{C}{n}$$

für eine Konstante $C > 0$, die weder von n noch von $(y_0, y_1, \dots)$ abhängt.

Beweis. Wie im Beweis des Satzes zuvor sei $(y_0, y_1, \dots)$ fest und nachfolgend in der Notation unterdrückt. Für beliebige W-Maße μ auf $(X, \mathcal{X})$ definieren wir

$$F_l \mu(A) := \frac{\int \mathbb{1}_A(x) \gamma_l(x) Q(x', dx) \mu(dx')}{\int \gamma_l(x) Q(x', dx) \mu(dx')} \quad \forall\, A \in \mathcal{X}. \tag{2.22}$$

Nach Bemerkung 2.3 gilt die Filterrekursion $\pi_{l|l} = F_l \pi_{l-1|l-1}$, und der Approximationsfehler $\pi_{k|k} - \widehat{\pi}_{k|k}$ besitzt die Zerlegung

$$\pi_{k|k} - \widehat{\pi}_{k|k} = F_k \pi_{k-1|k-1} - F_k \widehat{\pi}_{k-1|k-1} + F_k \widehat{\pi}_{k-1|k-1} - \widehat{\pi}_{k|k}.$$

Mit $F_0 \widehat{\pi}_{-1|-1} := \pi_{0|0}$ erhalten wir damit die Teleskopsummendarstellung

$$\pi_{k|k} - \widehat{\pi}_{k|k} = \sum_{l=0}^{k-1} \left(F_k \dots F_{l+1} F_l \widehat{\pi}_{l-1|l-1} - F_k \dots F_{l+1} \widehat{\pi}_{l|l} \right) + F_k \widehat{\pi}_{k-1|k-1} - \widehat{\pi}_{k|k}. \tag{2.23}$$

Es ist nun naheliegend, für jeden der Summanden sukzessive die Operationen $F_l, \dots, F_k$ abzuschätzen. Allerdings definiert F_l keinen Markovkern, denn die Zuordnung $\mu \mapsto F_l \mu$ ist nicht linear. Mit den Übergangskernen $\tilde{Q}_{l|k}$ aus (2.14) des inhomogenen Markovprozesses $(X_l)_{0\leq i\leq k} | Y_0, \dots, Y_k$ sowie

$$\mu_{l|k}(A) = \frac{\int \mathbb{1}_A(x) \beta_{l|k}(x, (y_{l+1}, \dots, y_k) \mu(dx)}{\int \beta_{l|k}(x, (y_{l+1}, \dots, y_k) \mu(dx)}$$

ist aber für jedes W-Maß μ auf $(X, \mathcal{X})$

$$F_k \dots F_{l+1}\mu = \mu_{l|k}\tilde{Q}_{l|k} \dots \tilde{Q}_{k-1|k}, \tag{2.24}$$

denn für alle $x \in X$ und $A \in \mathcal{X}$ folgt mit der Rückwärtsrekursion (2.6), dem verallgemeinerten Satz von Fubini für Markovkerne sowie $\beta_{k|k} = 1$

$$\begin{aligned}
(\mu_{k-1|k}\tilde{Q}_{k-1|k})(A) &= \int \tilde{Q}_{k-1|k}(x, A)\mu_{k-1|k}(dx) \\
&= \int \frac{\int_A \beta_{k|k}(x')\gamma_k(x')Q(x, dx')}{\beta_{k-1|k}(x)} \cdot \frac{\beta_{k-1|k}(x)\mu(dx)}{\int \beta_{k-1|k}(x)\mu(dx)} \\
&= \frac{\iint \mathbb{1}_A(x')\beta_{k|k}(x')\gamma_k(x')Q(x, dx')\mu(dx)}{\iint \beta_{k|k}(\tilde{x})\gamma_k(\tilde{x})Q(x, d\tilde{x})\mu(dx)} \\
&= \frac{\int \mathbb{1}_A(x')\gamma_k(x')(\mu Q)(dx')}{\int \gamma_k(x')(\mu Q)(dx')} = F_k\mu(A),
\end{aligned}$$

und Iteration ergibt (2.24). Damit kann man (2.23) aber schreiben als

$$\pi_{k|k} - \widehat{\pi}_{k|k} = \sum_{l=0}^{k-1} \big((\widehat{\pi}_{l-1|l-1})_{l|k}\tilde{Q}_{l-1|k} \dots \tilde{Q}_{k-1|k} - (\widehat{\pi}_{l|l})_{l|k}\tilde{Q}_{l|k} \dots \tilde{Q}_{k-1|k}\big) + F_k\widehat{\pi}_{k-1|k-1} - \widehat{\pi}_{k|k}. \tag{2.25}$$

Erfüllt Q nun die Atar-Zeitouni-Bedingung 2.22 mit einem Markovkern κ und $\varepsilon > 0$, so erfüllt $\tilde{Q}_{l|k}$ für $l < k$ nach dem Beweis von Proposition 2.21 die starke Doeblin-Bedingung

$$\tilde{Q}_{l|k}(x, A) \geq \varepsilon^2 \frac{\int_A \beta_{l+1|k}(x)\gamma_{l+1}(x)\kappa(dx)}{\int \beta_{l+1|k}(x)\gamma_{l+1}(x)\kappa(dx)} =: \varepsilon^2 K_l(A) \quad \forall\, x \in X,\ A \in \mathcal{X}.$$

Aber dann wird durch

$$\tilde{K}_l := \frac{1}{1-\varepsilon^2}(\tilde{Q}_{l|k} - \varepsilon^2 K_l)$$

ein Markovscher Übergangskern definiert. Wegen $\nu\tilde{Q}_{l|k} - \nu'\tilde{Q}_{l|k} = (1-\varepsilon^2)(\nu\tilde{K}_l - \nu'\tilde{K}_l)$ für beliebige W-Maße ν, ν' auf $(X, \mathcal{X})$ und $\|\tilde{K}_l f\|_{\sup} \leq \|f\|_{\sup}$ gilt

$$\begin{aligned}
&\sup_{f \in \mathcal{F}_1} \mathbb{E}^{1/2}\left[\left(\int f(x)(\mu_{l|k}\tilde{Q}_{l|k})(dx) - \int f(x)(\mu'_{l|k}\tilde{Q}_{l|k})(dx)\right)^2\right] \\
&\quad = (1-\varepsilon^2) \sup_{f \in \mathcal{F}_1} \mathbb{E}^{1/2}\left[\left(\int f(x)(\mu_{l|k}\tilde{K}_l)(dx) - \int f(x)(\mu'_{l|k}\tilde{K}_l)(dx)\right)^2\right] \\
&\quad \leq (1-\varepsilon^2) \sup_{f \in \mathcal{F}_1} \mathbb{E}^{1/2}\left[\left(\int f(x)\mu_{l|k}(dx) - \int f(x)\mu'_{l|k}(dx)\right)^2\right] \\
&\quad \leq (1-\varepsilon^2)\frac{\|\beta_{l|k}\|_{\sup}}{\inf_{x \in X} \beta_{l|k}(x)} \sup_{f \in \mathcal{F}_1} \mathbb{E}^{1/2}\left[\left(\int f(x)\mu(dx) - \int f(x)\mu'(dx)\right)^2\right],
\end{aligned}$$

wobei die letzte Ungleichung wie in Schritt (ii) aus dem Beweis von Satz 2.25 folgt. Nach Definition (2.4) von $\beta_{l|k}$ und der Atar-Zeitouni-Bedingung ist

$$\varepsilon C_{k,l} \leq \beta_{l|k}(x) \leq \frac{1}{\varepsilon} C_{k,l}$$

mit $C_{k,l} := \int_{X^{k-l}} (\prod_{j=l+2}^{k} \gamma_j(x_j) Q(x_{j-1}, dx_j)) \gamma_{l+1}(x_{l+1}) \kappa(dx_{l+1}) \in (0, \infty)$ wegen $\eta \leq \gamma \leq 1/\eta$. Sukzessives Anwenden dieser Abschätzung auf die Summanden in (2.25) ergibt

$$\sup_{f \in F_1} \mathbb{E}^{1/2}\left[\left(\int f(x)\pi_{k|k}(dx) - \int f(x)\widehat{\pi}_{k|k}(dx)\right)^2\right]$$
$$\leq \sum_{l=0}^{k} \varepsilon^{-2}(1-\varepsilon^2)^{k-l} \sup_{f \in \mathcal{F}_1} \mathbb{E}^{1/2}\left[\left(\int f(x) F_l \widehat{\pi}_{l-1|l-1}(dx) - \int f(x)\widehat{\pi}_{l|l}(dx)\right)^2\right].$$

Nun ist aber

$$F_l \widehat{\pi}_{l-1|l-1}(dx) = \frac{\gamma_l(x)(\widehat{\pi}_{l-1|l-1} Q)(dx)}{\int_X \gamma_l(x)(\widehat{\pi}_{l-1|l-1} Q)(dx)},$$

und die sinngemäß ausgeführten Schritte aus dem Beweis von Satz 2.25 (mit der oberen Schranke $2/\eta^2$ an die Konstante im Korrekturschritt) ergeben

$$\sup_{f \in \mathcal{F}_1} \mathbb{E}^{1/2}\left[\left(\int f(x) F_l \widehat{\pi}_{l-1|l-1}(dx) - \int f(x)\widehat{\pi}_{l|l}(dx)\right)^2\right] \leq \frac{1 + 2\eta^{-2}}{\sqrt{n}}.$$

Die Behauptung folgt schließlich durch Supremumsbildung über $k \in \mathbb{N}$. □

Bemerkung 2.27. *Unter den Mischungsbedingungen aus Satz 2.26 kumuliert der Fehler nicht – genau wie $(\pi_{k|k})_{k \geq 0}$ die Startbedingung "vergisst" der Algorithmus vorherige Approximationsfehler.*

3 Parameterinferenz

Ein HMM ist vollständig durch die Startverteilung ν von X_0, den Übergangskern Q sowie den Beobachtungskern G spezifiziert. Wohingegen wir uns im vorherigen Kapitel damit beschäftigt haben, wie man auf Basis von Beobachtungen $Y_0, \ldots, Y_N$ bei bekanntem (ν, Q, G) etwas über den unbeobachteten Signalprozess $(X_k)_{k \geq 0}$ aussagen kann, lösen wir uns nun von dieser statistisch oft nicht realistischen Situation des vollständig spezifizierten HMMs.

> Basierend auf Beobachtungen $Y_0, \ldots, Y_N$ studieren wir Parameterschätzung in der Situation, wo die Verteilung der bivariaten Markovkette $(X_k, Y_k)_{k \geq 0}$ des HMMs, also das Triplett (ν, Q, G), bis auf einen endlichdimensionalen und a priori unbekannten Parameter gegeben ist.

Es gibt zwei sehr prominente Ansätze zur Parameterschätzung – Bayes-Schätzer sowie den Maximum-Likelihood-Schätzer (MLE). In HMMs hat sich der MLE als sehr erfolgreich erwiesen und kann wesentlich effizienter berechnet werden als Bayes-Schätzer.

3.1 Grundlagen

Sei Θ eine Parametermenge. Jedem $\theta \in \Theta$ sei ein HMM mit Triplett $(\nu^\theta, Q^\theta, G^\theta)$ zugeordnet. $\mathbb{P}_\theta$ bezeichne die Verteilung der zugehörigen bivariaten Markovkette $(X_k, Y_k)_{k \geq 0}$, $\mathbb{P}_{\theta,N}$ die Verteilung des Abschnittes $(X_k, Y_k)_{0 \leq k \leq N}$. Entsprechend wird Größen aus dem vorherigen Kapitel oben oder unten ein zusätzlicher Index θ hinzugefügt, also bspw.

$\pi^\theta_{k|N}$, wenn sie unter dem HMM mit Triplett $(\nu^\theta, Q^\theta, G^\theta)$ gebildet werden. Wir setzen voraus, dass Θ mit einer σ-Algebra $\mathcal{H}$ versehen werden kann, so dass für alle $x \in X$, $A \in \mathcal{X}$ und $B \in \mathcal{Y}$ die Abbildungen $\theta \mapsto Q^\theta(x, A)$ und $\theta \mapsto G^\theta(x, B)$ $\mathcal{H} - \mathcal{B}(\mathbb{R})$-messbar sind.

Definition 3.1. *Die Familie $(\mathbb{P}_\theta : \theta \in \Theta)$ von HMMs heißt nicht-degeneriert, falls für ein W-Maß ϕ auf $(Y, \mathcal{Y})$ und jedes $\theta \in \Theta$ die Beobachtungskerne G^θ die Darstellung*

$$G^\theta(x, B) = \int \mathbb{1}_B(y)\gamma^\theta(x, y)\phi(dy), \quad x \in X,\ B \in \mathcal{Y}, \tag{3.1}$$

mit einer strikt positiven, messbaren Funktion $\gamma^\theta : X \times Y \to \mathbb{R}$ besitzen.

Im Falle der Nicht-Degeneriertheit sind die Marginalverteilungen $\mathbb{P}_\theta^{Y_0,\dots,Y_N}$ für jedes $\theta \in \Theta$ stetig bezüglich des Produktmaßes $\phi^{\otimes(N+1)}$. Der MLE basierend auf Beobachtungen $(Y_0, \dots, Y_N) = (y_0, \dots, y_N)$ ist dann definiert als

$$\hat{\theta}_N := \operatorname*{argmax}_{\theta\in\Theta} \frac{d\mathbb{P}_\theta^{Y_0,\dots,Y_N}}{d\phi^{\otimes(N+1)}}(y_0, \dots, y_N). \tag{3.2}$$

Lemma 3.2. *Für eine nicht-degenerierte Familie $(\mathbb{P}_\theta : \theta \in \Theta)$ von HMMs ist*

$$\begin{aligned} L_N(\theta) &:= \frac{d\mathbb{P}_\theta^{Y_0,\dots,Y_N}}{d\phi^{\otimes(N+1)}}(y_0, \dots, y_N) \\ &= \sigma_N^\theta((y_0, \dots, y_N), X) \\ &= \sigma_0^\theta(y_0, X) \prod_{k=1}^N \iint \gamma^\theta(x, y_k) Q^\theta(x', dx) \pi^\theta_{k-1|k-1}((y_0, \dots, y_{k-1}), dx'). \end{aligned} \tag{3.3}$$

Beweis. Für jede messbare, beschränkte Funktion $f : (Y^{N+1}, \mathcal{Y}^{N+1}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ist

$$\begin{aligned} &\mathbb{E}^\theta f(Y_0, \dots, Y_N) \\ &= \int f(y_0, \dots, y_N)\gamma^\theta(x_0, y_0)\nu(dx_0)\phi(dy_0) \prod_{k=1}^N \gamma^\theta(x_k, y_k)Q^\theta(x_{k-1}, dx_k)\phi(dy_k) \\ &= f(y_0, \dots, y_N)\sigma_N^\theta((y_0, \dots, y_N), X)\phi(dy_0) \dots \phi(dy_N) \end{aligned}$$

nach dem Satz von Fubini und der Definition (2.1) des Kerns σ_N^θ. Das beweist $L_N(\theta) = \sigma_N^\theta((y_0, \dots, y_N), X)$. Die zweite behauptete Gleichheit ist eine unmittelbare Konsequenz aus Identität (2.2) und der Filterrekursion. □

Für jedes $\theta \in \Theta$ kann man die Likelihood-Funktion leicht mithilfe der Filterrekursion berechnen. Um den MLE zu berechnen, müssen wir den Filter aber nicht nur für ein festes θ sondern eigentlich simultan für alle bestimmen, um anschließend das Maximum zu ermitteln. Ein solches Vorgehen ist allerdings extrem rechenintensiv. Einen Ausweg aus dieser Misere liefert unter geeigneten Voraussetzungen der sogenannte EM-Algorithmus.

3.2 Der EM-Algorihmus

Der Expectation-Maximization-Algorithmus (kurz EM-Algorithmus) ist ein Verfahren, eine Maximalstelle von $L_N(\theta)$ zu bestimmen, ohne diese Funktion für alle $\theta \in \Theta$ auswerten zu müssen – vorausgesetzt, die Parameterabhängigkeit ist von geeigneter Natur.

Annahme 3.3 (EM-Annahme)**.** *Es existieren ein Übergangskern Q auf $(X, \mathcal{X})$ sowie W-Maße ν auf $(X, \mathcal{X})$ und ϕ auf $(Y, \mathcal{Y})$, so dass Q^θ, G^θ und ν^θ Dichten der Gestalt von Exponentialfamilien haben:*

$$Q^\theta(x, dx') = q^\theta(x, x')Q(x, dx') \quad mit \quad q^\theta(x, x') = \exp\left(\sum_{l=1}^{d_q} c_l(\theta)q_l(x, x')\right),$$

$$G^\theta(x, dy) = \gamma^\theta(x, y)\phi(dy) \quad mit \quad \gamma^\theta(x, y) = \exp\left(\sum_{l=1}^{d_\gamma} \kappa_l(\theta)\gamma_l(x, y)\right),$$

$$\nu^\theta(dx) = u^\theta(x)\nu(dx) \quad mit \quad u^\theta(x) = \exp\left(\sum_{l=1}^{d_u} \eta_l(\theta)u_l(x)\right).$$

Diese Voraussetzungen sind durchaus nicht unrealistisch, wie folgende Beispiele zeigen.

Beispiel 3.4 (Endliche Zustandsräume)**.** *Ist der Signalzustandsraum $X = \{1, \dots, d\}$ endlich, kann man den Übergangskern Q^θ als stochastische Matrix $\mathbf{Q}^\theta = (\mathbf{Q}^\theta_{ij})_{1\le i,j\le d}$ darstellen. Gilt dann*

$$Q^\theta(i, \{j\}) = \mathbf{Q}^\theta_{ij} > 0$$

für alle $1 \le i, j \le d$ und alle $\theta \in \Theta$, so erfüllt Q^θ die entsprechende Voraussetzung aus Annahme 3.3. Für den Übergangskern Q mit $Q(i, \{j\}) = 1/d$ für alle $1 \le i, j \le d$ folgt nämlich

$$Q^\theta(i, \{j\}) = \mathbf{Q}^\theta_{ij} = \exp\left(\sum_{k,l=1}^{d} \underbrace{\log\left(\mathbf{Q}^\theta_{kl} d\right)}_{=:c_{kl}(\theta)} \underbrace{\mathbb{1}_k(i)\mathbb{1}_l(j)}_{=:q_{kl}(i,j)}\right) Q(i, \{j\}).$$

Ähnlich verifiziert man die Voraussetzungen an G^θ und ν^θ aus Annahme 3.3, wenn der Beobachtungszustandsraum Y endlich ist.

Beispiel 3.5 (Gaußsche Beobachtungen)**.** *Es seien wieder $X = \{1, \dots, d\}$ endlich und $Q^\theta(i, \{j\}) > 0$ für alle $1 \le i, j \le d$ und alle $\theta \in \Theta$. Nach Beispiel 3.4 erfüllt die Familie $(Q^\theta : \theta \in \Theta)$ die entsprechende Voraussetzung aus Annahme 3.3. Ist nun $G^\theta(i, dy)$ eine Gaußverteilung für alle $i \in \{1, \dots, d\}$ und $\theta \in \Theta$, so besitzt auch die Familie $(G^\theta : \theta \in \Theta)$ die in Annahme 3.3 geforderte Darstellung. Denn sind $m_i(\theta)$ und $v_i(\theta)$ Erwartungswert und Varianz der Normalverteilung $G^\theta(i, \cdot)$ sowie $\phi = \mathcal{N}(0, 1)$, dann folgt*

$$\begin{aligned} G^\theta(i, dy) &= \exp\left(\frac{1}{2}y^2 - \frac{(y - m_i(\theta))^2}{2v_i(\theta)} - \log\left(\sqrt{v_i(\theta)}\right)\right)\phi(dy) \\ &= \exp\left(\sum_{k=1}^{3}\sum_{l=1}^{d} \kappa_{kl}(\theta)\gamma_{kl}(i, y)\right)\phi(dy) \end{aligned}$$

mit $\gamma_{1l}(i, y) = y^2$, $\gamma_{2l}(i, y) = \mathbb{1}_l(i)y$, $\gamma_{3l}(i, y) = \mathbb{1}_l(i)$ und

$$\kappa_{1l} = \frac{1 - v_l(\theta)^{-1}}{2}, \quad \kappa_{2l}(\theta) = \frac{m_l(\theta)}{v_l(\theta)}, \quad \kappa_{3l}(\theta) = -\frac{m_l(\theta)^2}{2v_l(\theta)} - \log\left(\sqrt{v_l(\theta)}\right).$$

Unter der EM-Annahme ist $d\mathbb{P}_\theta^{Y_0,\dots,Y_N} \ll d\mathbb{P}_{\theta'}^{Y_0,\dots,Y_N}$ für $\theta, \theta' \in \Theta$ (Aufgabe 1(ii), Blatt 11). Wegen

$$\frac{d\mathbb{P}_\theta^{Y_0,\dots,Y_N}}{d\mathbb{P}_{\theta'}^{Y_0,\dots,Y_N}} \cdot \underbrace{\frac{d\mathbb{P}_{\theta'}^{Y_0,\dots,Y_N}}{d\phi^{\otimes(N+1)}}}_{\text{unabh. von } \theta} = \frac{d\mathbb{P}_\theta^{Y_0,\dots,Y_N}}{d\phi^{\otimes(N+1)}}$$

sowie Monotonie des Logarithmus ist dann für jedes feste $\theta' \in \Theta$ der MLE gleichermaßen gegeben durch

$$\hat{\theta}_N = \operatorname*{argmax}_{\theta\in\Theta} \log\left(\frac{d\mathbb{P}_\theta^{Y_0,\dots,Y_N}}{d\mathbb{P}_{\theta'}^{Y_0,\dots,Y_N}}(y_0,\dots,y_N)\right).$$

Nun ist aber nach Aufgabe 1(i) auf Blatt 11

$$\log\left(\frac{d\mathbb{P}_\theta^{Y_0,\dots,Y_N}}{d\mathbb{P}_{\theta'}^{Y_0,\dots,Y_N}}\right) = \log\left(\mathbb{E}^{\theta'}\left[\frac{d\mathbb{P}_{\theta,N}}{d\mathbb{P}_{\theta',N}}\,\middle|\, Y_0,\dots,Y_N\right]\right). \tag{3.4}$$

Das Maximum dieses Ausdrucks über θ ist a priori schwer zu berechnen. Betrachten wir statt dessen aber Erwartungswert und Logarithmus in vertauschter Reihenfolge, also die Größe

$$R_N(\theta,\theta') = \mathbb{E}^{\theta'}\left[\log\left(\frac{d\mathbb{P}_{\theta,N}}{d\mathbb{P}_{\theta',N}}\right)\,\middle|\, Y_0,\dots,Y_N\right],$$

lässt sich das Maximum leicht bestimmen, wenn die EM-Annahme gilt. Denn dann ist

$$\begin{aligned}
&\log\left(\frac{d\mathbb{P}_{\theta,N}}{d\mathbb{P}_{\theta',N}}(x_0,y_0,\dots,x_N,y_N)\right)\\
&\quad= \sum_{k=0}^{N}\sum_{l=1}^{d_\gamma}\big(\kappa_l(\theta)-\kappa_l(\theta')\big)\gamma_l(x_k,y_k) \;+\; \sum_{k=0}^{N}\sum_{l=1}^{d_q}\big(c_l(\theta)-c_l(\theta')\big)q_l(x_{k-1},x_k)\\
&\qquad+\; \sum_{l=1}^{d_u}\big(\eta_l(\theta)-\eta_l(\theta')\big)u_l(x_0),
\end{aligned}$$

womit

$$\begin{aligned}
R_N(\theta,\theta') &= \sum_{k=0}^{N}\sum_{l=1}^{d_\gamma}\big(\kappa_l(\theta)-\kappa_l(\theta')\big)\int \gamma_l(x,y_k)\pi_{k|N}^{\theta'}(dx)\\
&\quad+\; \sum_{k=0}^{N}\sum_{l=1}^{d_q}\big(c_l(\theta)-c_l(\theta')\big)\int q_l(x,x')\mathbb{P}_{\theta'}^{X_{k-1},X_k|Y_0,\dots,Y_N}(dx,dx')\\
&\quad+\; \sum_{l=1}^{d_u}\big(\eta_l(\theta)-\eta_l(\theta')\big)\int u_l(x)\pi_{0|N}^{\theta'}(dx).
\end{aligned}$$

Nach Blatt 11, Aufgabe 2 erfüllt auch $\pi_{k-1,k|N} := \mathbb{P}^{X_{k-1},X_k|Y_0,\dots,Y_N}$ eine Filterrekursion. Damit kann die Maximalstelle von $R_N(\theta,\theta')$ in folgenden zwei Schritten ermittelt werden:

(E) Bestimme die univariaten und bivariaten Glättungsverteilungen $\pi_{k|N}^{\theta'}$ und $\pi_{k-1,k|N}^{\theta'}$ (beispielsweise mithilfe des SISR-Algorithmus).

(M) Löse das deterministische Optimierungsproblem, R_N über θ zu maximieren.

Im Vergleich zum ursprünglichen Problem ist das deutlich einfacher, denn die Maximierung über θ ist separiert von der Berechnung der bedingten Erwartungswerte.

Was wir gerade exemplarisch im HMM beschrieben haben, ist die Kernidee des EM-Algorithmus. Unsere Überlegung hat allerdings eine gravierende Schwachstelle: Natürlich können wir nicht einfach den Erwartungswert mit dem Logarithmus vertauschen, wie wir es getan haben, um R_N zu erhalten – entsprechend ist

$$\operatorname{argmax}_{\theta\in\Theta} R_N(\theta,\theta')$$

gar nicht der MLE! Bemerkenswerterweise gilt allerdings Folgendes:

Lemma 3.6 (EM-Lemma)**.** *Ist* $\theta^* = \operatorname{argmax}_{\theta\in\Theta} R_N(\theta,\theta')$*, dann gilt mit der Likelihood-Funktion* L_N *aus* (3.3)

$$L_N(\theta^*) \geq L_N(\theta').$$

Beweis. Mit der Identität (3.4) gilt nach der Jensen-Ungleichung

$$\log L_N(\theta^*) - \log L_N(\theta') = \log\left(\mathbb{E}^{\theta'}\left[\frac{d\mathbb{P}_{\theta^*,N}}{d\mathbb{P}_{\theta',N}}\,\middle|\, Y_0,\dots,Y_N\right]\right) \geq R_N(\theta^*,\theta'). \tag{3.5}$$

Da θ^* Maximierer von $\theta \mapsto R_N(\theta,\theta')$ ist und $R_N(\theta',\theta') = 0$ gilt, folgt $R_N(\theta^*,\theta') \geq 0$. □

Im Übrigen ist obige Ungleichung (3.5) sogar strikt, sofern

$$\mathbb{P}_\theta^{Y_0,\dots,Y_N} \neq \mathbb{P}_{\theta'}^{Y_0,\dots,Y_N} \quad \forall\,\theta\in\Theta\setminus\{\theta'\}, \tag{3.6}$$

denn der Logarithmus ist streng konkav. Das Lemma suggeriert jetzt den sogenannten EM-Algorithmus:

- Wähle einen beliebigen Startparameter $\hat\theta_N^{(0)}$.
- Konstruiere eine Folge von Schätzern

  $$\hat\theta_N^{(n)} \in \operatorname{argmax}_{\theta\in\Theta} R_N\big(\theta,\hat\theta_N^{(n-1)}\big), \quad n\in\mathbb{N},$$

  durch iterierte Anwendung der (E)- und (M)-Schritte.

Die Likelihood L_N der EM-Schätzer $\hat\theta_N^{(n)}$ wächst in n, und sofern (3.6) gilt für alle $\theta,\theta'\in\Theta$ mit $\theta\neq\theta'$, wächst sie sogar strikt. Insofern besteht Anlass zur Hoffnung, dass die Folge $\big(\hat\theta_N^{(n)}\big)_{n\in\mathbb{N}}$ gegen den MLE konvergiert, was sie unter geeigneten Voraussetzungen auch tatsächlich tut. Um globale Konvergenzresultate und Konvergenzraten für den EM-Algorithmus in zufriedenstellender Allgemeinheit zu beweisen, müssten wir an dieser Stelle allerdings recht weit ausholen. Da unser Fokus aber statistische Eigenschaften des MLEs sind, gehen wir deswegen hier nicht weiter darauf ein.

3.3 Asymptotik des Maximum-Likelihood-Schätzers

Die klassische Asymptotik des MLEs basierend auf N iid-Beobachtungen für $N\to\infty$ hängt im Falle eines kompakten Parameterraumes $\Theta\subset\mathbb{R}^d$ und zweimal stetig differenzierbarer log-Likelihood-Funktion $\mathcal{L}_N$ mit dem den Beobachtungen zugrundeliegenden wahren Parameter θ_0 an drei grundlegenden Resultaten:

(i) einem Gesetz der großen Zahlen vom Glivenko-Cantelli-Typ:

$$\sup_{\theta\in\Theta}\Big|\frac{1}{N}\mathcal{L}_N(\theta)-\mathcal{L}(\theta)\Big| \to_{\mathbb{P}_{\theta_0}} 0 \quad \text{für } N\to\infty,$$

(ii) einem zentralen Grenzwertsatz für die Score-Funktion $\nabla_\theta\mathcal{L}_N(\theta_0)$ sowie

(iii) einem in θ_0 lokal gleichmäßigen Gesetz der großen Zahlen für $N^{-1}\nabla^2_\theta\mathcal{L}_N(\theta)$ gegen die Informationsmatrix $\mathcal{I}(\theta_0)$:

$$\lim_{\delta\searrow 0}\lim_{N\to\infty}\mathbb{P}_{\theta_0}\Big(\sup_{\|\theta-\theta_0\|_2\le\delta}\Big\|\frac{1}{N}\nabla^2_\theta\mathcal{L}_N(\theta)-\mathcal{I}(\theta_0)\Big\|>\varepsilon\Big)=0 \quad \forall\varepsilon>0.$$

Ist θ_0 eindeutiges $\operatorname{argmax}_{\theta\in\Theta}\mathcal{L}(\theta)$, garantiert Bedingung (i) Konsistenz des MLEs (gelegentlich formuliert man (i) mit einer fast sicheren Konvergenzaussage und schlussfolgert entsprechend die sogenannte starke Konsistenz). Gegeben die Konsistenz liefern (ii) und (iii) dann asymptotische Normalität, sofern θ_0 innerer Punkt von Θ ist und $\mathcal{I}(\theta_0)$ nichtsingulär ist.

3.3.1 Konsistenz

Natürlich ist es im Hinblick auf obiges (i) naheliegend, einer ähnlichen Strategie für Konsistenzresultate des MLEs in HMMs zu folgen. Nach Lemma 3.2 und Satz 2.7 besitzt $\ell_N := N^{-1}\log L_N$ mit der Konvention $\pi^\theta_{0|-1}=\nu^\theta$ folgende Darstellung:

$$\ell_N(\theta)=\frac{1}{N}\sum_{k=0}^{N}\underbrace{\log\Big(\int\gamma^\theta(x,Y_k)\pi^\theta_{k|k-1}(Y_0,\dots,Y_{k-1},dx)\Big)}_{=:D^\theta_k}=\frac{1}{N}\sum_{k=0}^{N}D^\theta_k. \tag{3.7}$$

Hierfür möchten wir also ein Gesetz der großen Zahlen beweisen – die Zufallsvariablen D^θ_k sind allerdings weder identisch verteilt noch stochastisch unabhängig.

Proposition 3.7. *Seien $(\mathcal{F}_k)_{k\ge0}$ eine Filtration und $(Z_k)_{k\ge0}$ eine adaptierte Folge von Zufallsvariablen mit*

$$\big|\,\mathbb{E}(Z_k|\mathcal{F}_l)\,\big|\le C\rho^{k-l}\ \text{f.s.}$$

für alle $1\le l\le k$ und eine Konstante $C>0$, $0<\rho<1$. Dann gilt

$$\frac{1}{n}\sum_{k=0}^{n}Z_k\longrightarrow 0 \quad \text{f.s. für } n\to\infty.$$

Beweis. Wir beweisen zunächst die Konvergenz im quadratischen Mittel. Mit $S_n := n^{-1}\sum_{k=0}^n Z_k$ gilt

$$\mathbb{E}\big(S_n^2\big)=\frac{1}{n^2}\sum_{k=0}^{n}\mathbb{E}\big(Z_k^2\big)+\frac{2}{n^2}\sum_{k=1}^{n}\sum_{l=0}^{k-1}\mathbb{E}\big(Z_kZ_l\big).$$

Nach Voraussetzung sind $Z_k^2\le C^2$ sowie

$$\big|\,\mathbb{E}\big(Z_kZ_l\big)\,\big| \;=\; \big|\,\mathbb{E}\big(\mathbb{E}\big(Z_k\,\big|\,\mathcal{F}_l\big)Z_l\big)\,\big|\le C^2\rho^{k-l},$$

womit nach Summierbarkeit der geometrischen Reihe

$$\mathbb{E}\big(S_n^2\big) \le \frac{C^2}{n} + \frac{2C^2}{n}\sum_{j=0}^{\infty}\rho^j = \frac{K}{n} \tag{3.8}$$

für eine Konstante $K = K(C,\rho) > 0$. Wir zeigen noch die fast sichere Konvergenz. Für jedes $\alpha > 1$ und $\varepsilon > 0$ gilt nach der Chebychev-Ungleichung und (3.8)

$$\sum_{k=1}^{\infty}\mathbb{P}\big(\,|\,S_{\lceil\alpha^k\rceil}\,|\,>\varepsilon\big) \le \sum_{k=1}^{\infty}\frac{\mathbb{E}\big(S^2_{\lceil\alpha^k\rceil}\big)}{\varepsilon^2} \le \frac{K}{\varepsilon^2}\sum_{k=1}^{\infty}\alpha^{-k} < \infty.$$

Das Borel-Cantelli-Lemma ergibt dann die fast-sicher-Konvergenz der Teilfolge $S_{\lceil\alpha^k\rceil} \to 0$ für $k \to \infty$. Mit

$$k_+^\alpha(n) = \inf\{j \in \mathbb{Z}|\, n \le \alpha^j\} \text{ sowie } k_-^\alpha(n) = \sup\{j \in \mathbb{Z}|\, \alpha^j < n\},$$

folgt weiter wegen der Nicht-Negativität von $Z_l + C$ fast sicher ($l \ge 1$) für alle $n \in \mathbb{N}$

$$\frac{\alpha^{k_-^\alpha(n)}}{\alpha^{k_+^\alpha(n)}}\frac{1}{\alpha^{k_-^\alpha(n)}}\sum_{l=0}^{\alpha^{k_-^\alpha(n)}}\big(Z_l+C\big) \le \frac{1}{n}\sum_{l=0}^{n}\big(Z_l+C\big) \le \frac{\alpha^{k_+^\alpha(n)}}{\alpha^{k_-^\alpha(n)}}\frac{1}{\alpha^{k_+^\alpha(n)}}\sum_{l=0}^{\alpha^{k_+^\alpha(n)}}\big(Z_l+C\big).$$

Für n hinreichend groß muss aber immer $k_+^\alpha(n) = k_-^\alpha(n) + 1$ gelten, womit

$$\frac{C}{\alpha} \le \liminf_{n\to\infty}\frac{1}{n}\sum_{l=0}^{n}\big(Z_l+C\big) \le \limsup_{n\to\infty}\frac{1}{n}\sum_{l=0}^{n}\big(Z_l+C\big) \le C\alpha \quad \text{f.s.}$$

Da $\alpha > 1$ beliebig war, folgt schließlich die Behauptung. □

Wir beweisen die Konsistenzresultate unter den folgenden Annahmen.

Annahme 3.8. *(i) Θ ist eine kompakte Teilmenge von $\mathbb{R}^d$.*

(ii) Es existieren $\varepsilon \in (0,1)$ sowie eine Familie $(\rho^\theta)_{\theta\in\Theta}$ von W-Maßen, so dass

$$\varepsilon\rho^\theta(A) \le Q^\theta(x,A) \le \frac{1}{\varepsilon}\rho^\theta(A) \quad \textit{für alle } x \in X,\ A \in \mathcal{X},\ \theta \in \Theta.$$

(iii) Es existiert eine Konstante $\kappa \in (0,1)$, so dass

$$\kappa \le \gamma^\theta(x,y) \le \frac{1}{\kappa} \quad \textit{für alle } x \in X,\ y \in Y,\ \theta \in \Theta.$$

(iv) Es gelten folgende Lipschitz-Eigenschaften:

$$\sup_{x\in X}\sup_{A\in\mathcal{X}}\big|\,Q^\theta(x,A) - Q^{\theta'}(x,A)\,\big| \le c_1\|\theta-\theta'\|_2$$
$$\sup_{x\in X}\sup_{y\in Y}|\gamma^\theta(x,y) - \gamma^{\theta'}(x,y)| \le c_2\|\theta-\theta'\|_2$$

für Kontanten $c_1, c_2 > 0$.

(v) Die Startverteilungen ν^θ sind jeweils invariante W-Maße von Q^θ, d.h. $\nu^\theta = \nu^\theta Q^\theta$.

Voraussetzung (ii) ist gerade die Atar-Zeitouni-Bedingung für die gesamte Familie $(Q^\theta)_{\theta\in\Theta}$ zu ein und demselben $\varepsilon > 0$, unter (ii) + (iii) haben wir im vorangehenden Kapitel die gleichmäßige Schranke für den SISR-Algorithmus bewiesen. Die Stationaritätsvoraussetzung aus (v) ist nicht notwendig – sie dient hier lediglich dazu, den technischen Aufwand nachfolgend gering zu halten.

Die nächste Proposition hat das entscheidende Glivenko-Cantelli-Resultat der reskalierten log-Likelihood-Funktion ℓ_N zum Gegenstand.

Proposition 3.9. *Angenommen, Annahme 3.8 sei erfüllt. Dann existiert für alle $\theta \in \Theta$ der Grenzwert $\ell(\theta) = \lim_{N\to\infty} \mathbb{E}_{\theta_0}\big(\ell_N(\theta)\big)$ und*

$$\sup_{\theta\in\Theta} \big|\, \ell_N(\theta) - \ell(\theta) \,\big| \longrightarrow 0 \quad \mathbb{P}_{\theta_0}\text{-f.s.} \quad (N \to \infty).$$

Bevor wir zum eigentlichen Beweis kommen, werden ihm noch zwei Hilfslemmata vorausgeschickt. Das erste zeigt, dass die Zufallsvariablen D_k^θ unter Annahme 3.8 gleichmäßig in k approximiert werden können durch die Größen

$$D_{k,l}^\theta := \log\left(\int \gamma^\theta(x, Y_k)\pi_{l|l-1}^\theta((Y_{k-l},\dots,Y_{k-1}), dx)\right),$$

welche nur von den jeweils letzten $l+1$ Beobachtungen $Y_{k-l},\dots,Y_k$ abhängen.

Lemma 3.10. *Angenommen, Annahme 3.8 sei erfüllt. Dann gilt für alle $l \in \mathbb{N}$*

$$\sup_{k\geq l} \big|\, D_{k,l}^\theta - D_k^\theta \,\big| \leq \frac{2}{\kappa^2}(1-\varepsilon^2)^l.$$

Beweis. Nach Annahme 3.8 (iii) ist $\gamma^\theta(x,y) \in [\kappa, \kappa^{-1}]$. Für $z \geq \kappa$ ist wegen $\log' z = z^{-1} \leq \kappa^{-1}$ nach dem Mittelwertsatz $|\log x - \log x'| \leq \kappa^{-1}|x - x'|$ für $x, x' \in [\kappa, \kappa^{-1}]$, womit

$$\begin{aligned}\big|\, D_{k,l}^\theta - D_k^\theta \,\big| \leq \frac{1}{\kappa}\Bigg| &\int \gamma^\theta(x, Y_k)\pi_{l|l-1}^\theta((Y_{k-l},\dots,Y_{k-1}), dx) \\ &- \int \gamma^\theta(x, Y_k)\pi_{k|k-1}^\theta((Y_0,\dots,Y_{k-1}), dx)\Bigg|.\end{aligned}$$

Die beiden Integrale im letzten Ausdruck kann man nach der Vorhersagerekursion aus Satz 2.7 mit $\tilde\gamma^\theta(x,y) = \int \gamma^\theta(x',y)Q^\theta(x,dx')$ in Ausdrücke der Filterverteilung umschreiben, d.h.

$$\begin{aligned}\big|\, D_{k,l}^\theta - D_k^\theta \,\big| \leq \frac{1}{\kappa}\Bigg| &\int \tilde\gamma^\theta(x, Y_k)\pi_{l-1|l-1}^\theta((Y_{k-l},\dots,Y_{k-1}), dx) \\ &- \int \tilde\gamma^\theta(x, Y_k)\pi_{k-1|k-1}^\theta((Y_0,\dots,Y_{k-1}), dx)\Bigg|.\end{aligned}$$

Für W-Maße μ auf $(X, \mathcal{X})$ definieren wir wie im Beweis von Satz 2.26

$$\mu_{l|k}(A) = \frac{\int \mathbb{1}_A(x)\beta_{l|k}^\theta(x,(y_{l+1},\dots,y_k)\mu(dx)}{\int \beta_{l|k}^\theta(x,(y_{l+1},\dots,y_k)\mu(dx)}.$$

Mit den Übergangskernen $\tilde Q_{l|k-1}^\theta$ der inhomogenen Markovkette $(X_{k'})_{k'\geq 0}|Y_0,\dots Y_{k-1}$ aus (2.14) (die Abhängigkeit von $Y_{l+1},\dots,Y_{k-1}$ ist in der Notation von $\tilde Q_{l|k-1}^\theta$ unterdrückt) gilt nun einerseits mit F_l aus (2.22) nach (2.24)

$$\begin{aligned}\pi_{k-1|k-1}^\theta((Y_0,\dots,Y_{k-1}),\cdot) &= F_{k-1}\dots F_{k-l}\pi_{k-l-1|k-l-1}^\theta \\ &= \big(\pi_{k-l-1|k-l-1}^\theta\big)_{k-l|k-1}\tilde Q_{k-l-1|k}^\theta\dots\tilde Q_{k-1|k-1}^\theta,\end{aligned}$$

andererseits ist

$$\pi_{l-1|l-1}((Y_{k-l},\dots,Y_{k-1}),\cdot) = \nu^\theta_{k-l|k-1}\tilde{Q}^\theta_{k-l|k-1}\dots\tilde{Q}^\theta_{k-1|k-1}. \tag{3.9}$$

Annahme 3.8 (ii) ist aber gerade die Atar-Zeitouni-Bedingung zur Konstante $\varepsilon \in (0,1)$, womit

$$\tilde{Q}^\theta_{j|k-1} \quad \text{für } j < k$$

nach dem Beweis von Proposition 2.21 die starke Doeblin-Bedingung zur Konstante ε^2 erfüllt. Wegen $\tilde{\gamma}^\theta \le \kappa^{-1}$ ist damit nach Lemma 2.8 (ii) und der Submultiplikativität des Dobrushin-Koeffizienten

$$\begin{aligned}
|\, D^\theta_{k,l} - D^\theta_k \,| &\le \frac{1}{\kappa^2}\left\| \left(\pi^\theta_{l-1|l-1}\right)_{l-1|k-1}\tilde{Q}^\theta_{l|k-1}\dots\tilde{Q}^\theta_{k-1|k-1} - \nu^\theta_{l-1|k-1}\tilde{Q}^\theta_{l|k-1}\dots\tilde{Q}^\theta_{k-1|k-1}\right\|_{TV} \\
&\le \frac{2}{\kappa^2}\delta\Big(\prod_{i=1}^{l}\tilde{Q}^\theta_{k-i|k-1}\Big) \le \frac{2}{\kappa^2}(1-\varepsilon^2)^l.
\end{aligned}$$

□

Das zweite Hilfslemma zeigt, dass die Familie (in k) der Abbildungen $\theta \mapsto D^\theta_k$ gleichgradig Lipschitz-stetig ist.

Lemma 3.11. *Angenommen, Annahme 3.8 sei erfüllt. Dann existiert eine Konstante $K > 0$, so dass für alle $\theta, \theta' \in \Theta$*

$$\sup_{k\in\mathbb{N}} |\, D^\theta_k - D^{\theta'}_k \,| \le K\|\theta - \theta'\|_2.$$

Beweis. Vollkommen analog zum obigen Beweis von Lemma 3.10 erhält man mit (3.9)

$$\begin{aligned}
&|\, D^\theta_k - D^{\theta'}_k \,| \\
&\quad\le \frac{1}{\kappa^2}\left\| \pi^\theta_{k-1|k-1}((Y_0,\dots,Y_{k-1}),\cdot) - \pi^{\theta'}_{k-1|k-1}((Y_0,\dots,Y_{k-1}),\cdot)\right\|_{TV} \\
&\quad= \frac{1}{\kappa^2}\left\| F^\theta_{k-1}\dots F^\theta_1\nu^\theta - F^{\theta'}_{k-1}\dots F^{\theta'}_1\nu^{\theta'}\right\|_{TV} \\
&\quad\le \frac{1}{\kappa^2}\sum_{j=1}^{k-2}\left\| F^{\theta'}_{k-1}\dots F^{\theta'}_{k-j}F^\theta_{k-j-1}\dots F^\theta_1\nu^\theta - F^{\theta'}_{k-1}\dots F^{\theta'}_{k-(j-1)}F^\theta_{k-j}\dots F^\theta_1\nu^\theta\right\|_{TV} \\
&\qquad + \frac{1}{\kappa^2}\left\| F^{\theta'}_{k-1}\dots F^{\theta'}_1\nu^\theta - F^{\theta'}_{k-1}\dots F^{\theta'}_1\nu^{\theta'}\right\|_{TV} \\
&\quad\le \frac{1}{\kappa^2}\sum_{j=1}^{k-2}\varepsilon^{-2}(1-\varepsilon^2)^j\left\| F^{\theta'}_{k-j}F^\theta_{k-j-1}\dots F^\theta_1\nu^\theta - F^\theta_{k-j}\dots F^\theta_1\nu^\theta\right\|_{TV} \\
&\qquad + \varepsilon^{-2}(1-\varepsilon^2)^{k-1}\frac{1}{\kappa^2}\left\| \nu^\theta - \nu^{\theta'}\right\|_{TV},
\end{aligned}$$

mit F_j aus (2.22), wobei beim letzten Ungleichheitszeichen Aufgabe 1 auf Übungsblatt 13 verwendet wurde. Nun ergibt einerseits die aufeinanderfolgende Anwendung von Annahme 3.8 (v), der Dreiecksungleichung für die Totalvariationsnorm nach Nullergänzung mit $\pm\nu^{\theta'}Q^\theta$ sowie Annahme 3.8 (ii) und (iv)

$$\left\| \nu^\theta - \nu^{\theta'}\right\|_{TV} = \left\| \nu^\theta Q^\theta - \nu^{\theta'}Q^{\theta'}\right\|_{TV} \le c_1\|\theta-\theta'\|_2 + (1-\varepsilon)\left\| \nu^\theta - \nu^{\theta'}\right\|_{TV},$$

womit $\left\| \nu^\theta - \nu^{\theta'} \right\|_{TV} \leq \varepsilon^{-1} c_1 \|\theta - \theta'\|_2$. Mit $\tilde{\gamma}^\theta(x,y) = \int \gamma^\theta(x',y) Q^\theta(x,dx')$ wie oben gilt andererseits für jedes W-Maß μ auf $(X, \mathcal{X})$

$$\begin{aligned}
&\left\| F_j^\theta \mu - F_j^{\theta'} \mu \right\|_{TV} \\
&\quad \leq \sup_{x \in \mathcal{X}} \left| \frac{\tilde{\gamma}^\theta(x, Y_j)}{\int \tilde{\gamma}^\theta(x, Y_j) \mu(dx)} - \frac{\tilde{\gamma}^{\theta'}(x, Y_j)}{\int \tilde{\gamma}^{\theta'}(x, Y_j) \mu(dx)} \right| \\
&\quad \leq \frac{|\tilde{\gamma}^\theta(x, Y_j) - \tilde{\gamma}^{\theta'}(x, Y_j)|}{\int \tilde{\gamma}^\theta(x, Y_j) \mu(dx)} + \frac{\int \tilde{\gamma}^\theta(x, Y_j) \mu(dx) \int \left| \tilde{\gamma}^\theta(x, Y_j) - \tilde{\gamma}^{\theta'}(x, Y_j) \right| \mu(dx)}{\int \tilde{\gamma}^\theta(x, Y_j) \mu(dx) \int \tilde{\gamma}^{\theta'}(x, Y_j) \mu(dx)} \\
&\quad \leq \left(\frac{1}{\kappa} + \frac{1}{\kappa^3} \right) \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \left| \tilde{\gamma}^\theta(x,y) - \tilde{\gamma}^{\theta'}(x,y) \right| \\
&\quad \leq \left(\frac{1}{\kappa} + \frac{1}{\kappa^3} \right) \left(c_2 + \frac{c_1}{\kappa} \right) \|\theta - \theta'\|_2.
\end{aligned}$$

□

Der Beweis von Proposition 3.9 untergliedert sich in drei Schritte:

(1) Als Erstes wird gezeigt, dass mit D_k^θ aus (3.7) der Grenzwert

$$\ell(\theta) = \lim_{k \to \infty} \mathbb{E}_{\theta_0} D_k^\theta$$

für jedes $\theta \in \Theta$ existiert.

(2) Anschließend beweisen wir, dass Konstanten $C > 0$ und $0 < \rho < 1$ existieren, so dass

$$\left| \mathbb{E}_{\theta_0} \left(D_k^\theta - \mathbb{E}_{\theta_0}(D_k^\theta) \,\middle|\, Y_0, \ldots, Y_l \right) \right| \leq C \rho^{k-l}$$

für alle $0 \leq l \leq k$, also dass die Zufallsvariablen $Z_k := D_k^\theta - \mathbb{E}_{\theta_0} D_k^\theta$ die Vorausetzung aus Proposition 3.7 erfüllen. Letztere ergibt dann zusammen mit dem ersten Schritt die Konvergenz

$$\ell_N(\theta) = \frac{1}{N} \sum_{k=0}^{N} \left(D_k^\theta - \mathbb{E}_{\theta_0} D_k^\theta \right) + \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{\theta_0} D_k^\theta \longrightarrow \ell(\theta) \quad \mathbb{P}_{\theta_0}\text{-f.s.} \quad (N \to \infty).$$

(3) Schließlich wird gezeigt, dass die Konvergenz sogar gleichmäßig in $\theta \in \Theta$ gilt.

Beweis von Proposition 3.9. (1) Sei $\Delta_l(\theta) := \mathbb{E}_{\theta_0} D_l^\theta$. Da der Prozess $(Y_k)_{k \geq 0}$ wegen Annahme 3.8 (v) unter $\mathbb{P}_{\theta_0}$ stationär ist, gilt $\Delta_l(\theta) = \mathbb{E}_{\theta_0} D_{k,l}^\theta$ für jedes $k > l$. Es folgt mit Lemma 3.10

$$|\Delta_{m+n} - \Delta_m| = \left| \mathbb{E}_{\theta_0} D_{m+n}^\theta - \mathbb{E}_{\theta_0} D_{m+n,m}^\theta \right| \leq \frac{2}{\kappa^2} (1 - \varepsilon^2)^m,$$

womit insbesondere $\sup_{n \in \mathbb{N}} |\Delta_{m+n} - \Delta_m| \to 0$ für $m \to \infty$. Also ist $(\Delta_m)_{m \geq 0}$ eine Cauchyfolge und somit konvergent, was wiederum die Konvergenz von

$$\frac{1}{n} \sum_{m=0}^{n} \Delta_m = \mathbb{E}_{\theta_0}\left(\ell_N(\theta) \right)$$

für jedes $\theta \in \Theta$ impliziert. Der Grenzwert sei mit $\ell(\theta)$ bezeichnet.

(2) Für jedes feste $n > 1$ ist $D^\theta_{k+n,n-1} = f^\theta_n(Y_{k+1}, \dots, Y_{k+n})$ für alle $k \geq 0$ mit einer messbaren Funktion $f^\theta_n : (Y^n, \mathcal{Y}^n) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Nach Definition 1.5 des Hidden-Markov-Modells ist

$$\mathbb{E}_{\theta_0}\big(D^\theta_{l+n,n-1} \,\big|\, X_0, Y_0, \dots, X_l, Y_l\big) = g^\theta_n(X_l)$$

für alle $l \geq 0$ mit einer messbaren Funktion $g^\theta_n : (X, \mathcal{X}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Damit gilt für jedes $n > 1$ und alle $l, m \geq 0$ nach der Turmeigenschaft der bedingten Erwartung, der Markov-Eigenschaft des Signalprozesses $(X_k)_{k\geq 0}$ sowie Annahme 3.8 (v)

$$\begin{aligned}
&\Big|\, \mathbb{E}_{\theta_0}\big(D^\theta_{l+m+n,n-1} \,\big|\, X_0, Y_0, \dots, X_l, Y_l\big) - \mathbb{E}_{\theta_0} D^\theta_{l+m+n,n-1} \,\Big| \\
&\quad = \Big|\, \mathbb{E}_{\theta_0}\big(g^\theta_n(X_{l+m}) \,\big|\, X_0, Y_0, \dots, X_l, Y_l\big) - \mathbb{E}_{\theta_0} g^\theta_n(X_{l+m}) \,\Big| \\
&\quad = \Big|\, \mathbb{E}_{\theta_0}\big(g^\theta_n(X_{l+m}) \,\big|\, X_l\big) - \mathbb{E}_{\theta_0} g^\theta_n(X_{l+m}) \,\Big| \\
&\quad = \Big|\, \int g^\theta_n(x)\big(\delta_{X_l}(Q^{\theta_0})^m\big)(dx) - \int g^\theta_n(x)\big(\nu^{\theta_0}(Q^{\theta_0})^m\big)(dx) \,\Big|. \qquad (3.10)
\end{aligned}$$

Annahme 3.8 (iii) impliziert aber $\|g_n\|_{\sup} \leq \kappa^{-1}$ und wegen Annahme 3.8 (ii) erfüllt Q^{θ_0} die Doeblin-Bedingung mit Konstante ε, womit

$$(3.10) \leq \frac{1}{\kappa} \big\| \delta_{X_l}(Q^{\theta_0})^m - \nu^{\theta_0}(Q^{\theta_0})^m \big\|_{TV} \leq \frac{2}{\kappa}(1-\varepsilon)^m$$

nach der Submultiplikativität des Dobrushin-Koeffizienten. Zusammen mit Lemma 3.10 folgt daraus schließlich

$$\Big|\, \mathbb{E}_{\theta_0}\big(D^\theta_{l+m+n} \,\big|\, X_0, Y_0, \dots, X_l, Y_l\big) - \mathbb{E}_{\theta_0} D^\theta_{l+m+n} \,\Big| \leq \frac{2}{\kappa}(1-\varepsilon)^m + \frac{4}{\kappa^2}(1-\varepsilon^2)^{n-1}.$$

Setzen wir nun speziell $m = n-2$ und $m = n-1$ ergibt sich für alle $k \geq 2$ die obere Schranke

$$\Big|\, \mathbb{E}_{\theta_0}\big(D^\theta_{l+k} \,\big|\, X_0, Y_0, \dots, X_l, Y_l\big) - \mathbb{E}_{\theta_0} D^\theta_{l+k} \,\Big| \leq \underbrace{(1-\varepsilon^2)^{-1}\Big(\frac{2}{\kappa} + \frac{4}{\kappa^2\varepsilon^2}\Big)}_{=:C} \underbrace{\sqrt{(1-\varepsilon^2)}}_{=:\rho}{}^k.$$

Damit erfüllen $(Z_k)_{k\geq 0}$ und $(\mathcal{F}_k)_{k\geq 0}$ mit $Z_k = D^\theta_k - \mathbb{E}_{\theta_0} D^\theta_k$ und $\mathcal{F}_k = \sigma(X_0, Y_0, \dots, X_k, Y_k)$ die Voraussetzung von Proposition 3.7.

(3) Nach Lemma 3.11 ist die Folge $\ell_n : \Theta \to \mathbb{R}$ gleichgradig Lipschitz-stetig; es bezeichne K eine Schranke an die Lipschitzhalbnorm. Wegen

$$|\ell(\theta) - \ell(\theta')| \leq |\ell(\theta) - \ell_N(\theta)| + |\ell(\theta') - \ell_N(\theta')| + L\|\theta - \theta'\|_2 \longrightarrow K\|\theta - \theta'\|_2 \quad \mathbb{P}^{\theta_0}\text{-f.s.}$$

nach Schritt (2) ist auch die (deterministische) Funktion $\ell : \Theta \to \mathbb{R}$ Lipschitz-stetig zur Konstante K. Da Θ nach Annahme 3.8 (i) kompakt ist, existiert für jedes $\delta > 0$ eine endliche Überdeckung durch offene Kugeln vom Radius δ mit Mittelpunkten in Θ. D.h. für jedes $\delta > 0$ existiert eine endliche Teilmenge $\Theta_\delta \subset \Theta$, mit $\inf_{\theta' \in \Theta_\delta} \|\theta - \theta'\|_2 < \delta$ für jedes $\theta \in \Theta$. Nach der Dreiecksungleichung und Lemma 3.11 ist

$$\sup_{\theta\in\Theta} \big|\, \ell_N(\theta) - \ell(\theta) \,\big| \leq 2K\delta + \max_{\theta\in\Theta_\delta} \big|\, \ell_N(\theta) - \ell(\theta) \,\big| \leq 2K\delta + \sum_{\theta\in\Theta_\delta} \big|\, \ell_N(\theta) - \ell(\theta) \,\big|.$$

Aber da $\ell_N(\theta) \to \ell(\theta)$ $\mathbb{P}_{\theta_0}$-f.s. für jedes $\theta \in \theta$ und Θ_δ endlich ist, folgt daraus

$$\limsup_{N\to\infty} \sup_{\theta\in\Theta} \big|\, \ell_N(\theta) - \ell(\theta) \,\big| \le 2K\delta \quad \mathbb{P}^{\theta_0}\text{-f.s.}$$

Da $\delta > 0$ beliebig war, impliziert dies $\sup_{\theta\in\Theta} \big|\, \ell_N(\theta) - \ell(\theta) \,\big| \to 0$ $\mathbb{P}^{\theta_0}$-f.s. □

Nach aller Vorarbeit können wir nun das Hauptresultat dieses Abschnitts formulieren.

Satz 3.12. *Angenommen, es gilt Annahme 3.8 und die Funktion $\theta \mapsto \ell(\theta)$ besitze ein eindeutiges Maximum in $\theta = \theta_0$ $\mathbb{P}_{\theta_0}$-f.s. Dann ist der MLE konsistent.*

Beweis. Nach Voraussetzung ist $\theta_0 = \operatorname{argmax}_{\theta\in\Theta} \ell(\theta)$ und nach Definition des MLEs gilt $\hat\theta_N \in \operatorname{argmax}_{\theta\in\theta} \ell_N(\theta)$. Es folgt nach Nullergänzung mit $\pm\ell_N(\theta)$, Subadditivität des Supremums sowie Proposition 3.9

$$\begin{aligned}
0 &\le \ell(\theta_0) - \ell(\hat\theta_N) \\
&= \sup_{\theta\in\Theta} \ell(\theta) - \ell(\hat\theta_N) \\
&\le \sup_{\theta\in\Theta} \big(\ell(\theta) - \ell_N(\theta)\big) + \sup_{\theta\in\Theta} \ell_N(\theta) - \ell(\hat\theta_N) \\
&= \sup_{\theta\in\Theta} \big(\ell(\theta) - \ell_N(\theta)\big) + \ell_N(\hat\theta_N) - \ell(\hat\theta_N) \\
&\le 2 \sup_{\theta\in\Theta} \big|\, \ell(\theta) - \ell_N(\theta) \,\big| \longrightarrow 0 \quad \mathbb{P}_{\theta_0}\text{-f.s.} \quad (N\to\infty).
\end{aligned}$$

Seien $M = \{\omega \in \Omega |\, \ell(\hat\theta_N(\omega)) \to \ell(\theta_0)\}$ und $\omega \in M$. Da $\Theta \subset \mathbb{R}^d$ nach Annahme 3.8 (i) kompakt ist, existiert zu jeder Teilfolge $(N_n)_{n\in\mathbb{N}}$ eine Teilteilfolge $(N_n(m))_{m\in\mathbb{N}}$, entlang der $\hat\theta_{N_n(m)}(\omega)$ konvergent ist gegen einen Grenzwert $\theta(\omega) \in \Theta$. Stetigkeit von ℓ impliziert

$$\ell(\hat\theta_{N_n(m)}(\omega)) \longrightarrow \ell(\theta(\omega)) = \ell(\theta_0) \quad \text{für } m \to \infty$$

(als gleichmäßiger Grenzwert stetiger Funktionen ist ℓ stetig – s.o. auch (3)). Aber daraus folgt $\theta(\omega) = \theta_0$, da θ_0 einzige Maximalstelle ist, womit $\hat\theta_N(\omega) \to \theta_0$ für alle $\omega \in M$. □

Bemerkung 3.13. *Der von uns geführte Konsistenzbeweis hängt essentiell an der Identifizierbarkeit des Modells in dem Sinne, dass der Grenzwert $\ell(\theta) = \lim_{N\to\infty} \mathbb{E}_{\theta_0}\big(\ell_N(\theta)\big)$ ein eindeutiges Maximum in θ_0 hat. Man kann aber zeigen, dass unter Annahme 3.8 gilt:*

(i) $\ell(\theta) \le \ell(\theta_0)$ *für alle* $\theta \in \Theta$ *und*

(ii) $\ell(\theta) = \ell(\theta_0) \Leftrightarrow \mathbb{P}_\theta^{(Y_k)_{k\ge0}} = \mathbb{P}_{\theta_0}^{(Y_k)_{k\ge0}}$.

Sind also für je zwei Parameter θ und θ' die Verteilungsgesetze der unendlich langen gesamten Beobachtungsfolge unterschiedlich, ist die Eindeutigkeit der Maximalstelle bereits gewährleistet.

3.3.2 Ausblick: Asymptotische Normalität

Um asymptotische Normalität des MLEs zu beweisen, kann man der klassischen Strategie (ii) – (iii) zu Beginn von Abschnitt 3.3 folgen und die entsprechende Konvergenz der Ableitungen der Likelihood-Funktion beweisen. Die dafür erforderlichen Argumente sind im Wesentlichen dieselben, die wir auch für den Konsistenzbeweis herangezogen haben. Anstelle des starken Gesetzes der großen Zahlen aus Proposition 3.7 benötigen wir hier allerdings einen geeigneten CLT für abhängige Zufallsvariablen.

Literatur

Olivier Cappé, Eric Moulines and Tobias Rydén, (2009). Inference in Hidden Markov Models, Springer.

Ramon van Handel, (2008). Hidden Markov models, Lecture Notes, Princeton.