# Entropie und Große Abweichungen

Notizen zum gleichnamigen Modul im SoSe 2018

Gerhard Keller
Department Mathematik
Universität Erlangen-Nürnberg

Version vom 9. Juli 2018

## Inhaltsverzeichnis

# Literatur


[1] https://de.wikipedia.org/wiki/Entropie

[2] https://en.wikipedia.org/wiki/Second_law_of_thermodynamics

[3] https://de.wikipedia.org/wiki/Ludwig_Boltzmann

[4] https://de.wikipedia.org/wiki/Entropie_(Informationstheorie)

[5] https://en.wikipedia.org/wiki/Kullback-Leibler_divergence

[6] https://en.wikipedia.org/wiki/Principle_of_maximum_entropy

[7] J. Aczél, B. Forte, and C.T. Ng. Why Shannon and Hartley entropies are "natural". *Adv. Appl. Probab.*, 6:131–146, 1974.

[8] N. Barton and H. de Vladar. Statistical mechanics and the evolution of polygenic quantitative traits. *Genetics*, 181:997–1011, 2009.

[9] Raphaël Cerf. *On Cramér's Theory in Infinite Dimensions.* Number 23 in Panoramas et Synthèses. Société Mathématique de France, 2007.

[10] Imre Csiszár. $i$-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

[11] Imre Csiszár. Axiomatic Characterizations of Information Measures. *Entropy*, 10(3):261–273, September 2008.

[12] Didier Dacunha-Castelle and Marie Duflo. *Probability and Statistics, Volume I.* Springer, 1986.

[13] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Springer, 1998.

[14] D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms.* Cambridge Univ. Press, 2009. Bibliothek: 18MI/mat 5.1-24.

[15] M.I. Freidlin and A.D. Wentzell. On small random perturbations of dynamical systems. *Russian Math. Surveys*, 25:1–55, 1970.

[16] Amos Golan, George Judge, and Douglas Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data.* John Wiley, 1996.

[17] Olav Kallenberg. *Foundations of Modern Probability.* Springer, second edition, 2002.

[18] Achim Klenke. *Wahrscheinlichkeitstheorie.* Springer, 2005. Als e-book: http://dx.doi.org/10.1007/978-3-540-77571-3.

[19] Wolfgang König. Große Abweichungen, Techniken und Anwendungen. Vorlesungsskript, Universität Leipzig, 2006.

[20] Ingo Müller. *A History of Thermodynamics.* Springer, 2007.

[21] Steven Orey. Large deviations for the empirical field of curie-weiss models. *Stochastics*, 25:3–14, 1988.

[22] Sung Y. Park and Anil K. Bera. Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics*, 150(2):219–230, June 2009.

[23] S.R.S. Varadhan. Asymptotic probabilities and differential equations. *Comm. Pure Appl. Math.*, 19:261–286, 1966.

[24] Anita Winter. Die Theorie der großen Abweichungen und Anwendungen. Vorlesungsskript, TU München (basierend auf dem Skript von W. König, 2009/10.

[25] Hermann Witting. *Mathematische Statistik I.* Teubner, 1985.

[26] X. Wu. Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*, 115:347–354, 2003.

**Aufgaben*:**
2.3 (Leonie Wicht) √
2.4 (Moritz Hanika) √
2.7 (Friedrich Wagner) √
3.3 (Jonas Neumann) √
3.4
3.5
3.6
4.1
5.1 (Moritz Hanika)
5.2 (Stephan Gärttner)
5.3 (Leonie Wicht) √
5.4
6.2
6.3
6.5
7.1
7.3
10.1

**Entropie (Karikatur)**

- Prozesse mit Umwandlung zwischen Energieformen in isolierten Systemen
- *Rudolf Clausius* (1865): In ,,isolierten Systemen"

$$dS \geqslant \frac{\delta Q}{T} \qquad \text{Entropieänderung}$$

mit Gleicheit in ,,idealen reversiblen" Systemen.

$Q, T, V, \ldots$ makroskopische Größen

$\uparrow$

- *Ludwig Boltzmann* (1877): Statistik mikroskopischer Größen,
  Entropie als Maß der ,,Zufälligkeit" einer Verteilung

**Große Abweichungen**
Spürbare Abweichungen vom statistischen Mittel sind bei sehr großen Systemen extrem unwahrscheinlich. Diese Wahrscheinlichkeit ist um so kleiner, je größer die Abweichung vom Mittel ist, und diese Abweichung kann durch Entropiedifferenzen gemessen werden.

**Vorbemerkung** Die im Skript enthaltenen **Aufgaben** sind dazu da, dass Sie sich mit Grundbegriffen, Rechentechniken und Argumentationsweisen etwas vertrauter machen. Die **Aufgaben*** sind zum Vorrechnen in den Übungen gedacht.

# 1 Einige Begriffe der Stochastik

Hier ein paar kurze Erinnerungen an die Maßtheorie und evtl. die Stochastik. Details dazu findet man in jedem einschlägigen Lehrbuch, z.B. in [18].

- Ist $(\Omega, \mathcal{F}, P)$ ein W'raum und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable (ZV), so ist die *Verteilung* $P_X$ von $X$ das durch $P_X(A) = P(X^{-1}(A))$ definierte Wahrscheinlichkeitsmaß auf dem Raum $(\mathbb{R}, \mathcal{B})$, wo $\mathcal{B}$ die Borelsche $\sigma$-Algebra auf $\mathbb{R}$ ist. Genauso definiert man $P_X$ für $\mathbb{R}^d$-wertige ZVn und allgemeiner auch für Zufallsgrößen $X : \Omega \to M$, die Werte in einem beliebigen messbaren Raum $(M, \mathcal{M})$ annehmen.
- Sind $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ ZVn, so ist ihre *gemeinsame Verteilung* $P_{X_1,\ldots,X_n}$ die Verteilung der $\mathbb{R}^n$-wertigen ZV $(X_1, \ldots, X_n)$. Die Verteilung der einzelnen $X_i$ erhält man daraus durch $P_{X_i}(A) = P_{X_1,\ldots,X_n}(\mathbb{R}^{i-1} \times A \times \mathbb{R}^{n-i})$.
- Die ZVn $X_1, \ldots, X_n$ sind unabhängig gdw. $P_{X_1,\ldots,X_n} = P_{X_1} \times \ldots \times P_{X_n}$.
- Die Kovarianzmatrix der $\mathbb{R}^n$-wertigen ZV $X = (X_1, \ldots, X_n)^t$ ist

  $$\mathrm{Var}(X) = E[X \cdot X^t] - E[X] \cdot E[X]^t,$$

  also $\mathrm{Var}(X)_{ij} = \mathrm{Cov}(X_i, X_j)$. Sie ist positiv semi-definit.
- Seien $(M, \mathcal{M}, Q)$ ein $\sigma$-endlicher Maßraum, $P$ ein W'maß auf $(M, \mathcal{M})$.
  - $P \ll Q$ („$P$ ist absolut stetig zu $Q$“), falls $Q(A) = 0 \Rightarrow P(A) = 0$ für alle $A \in \mathcal{M}$.
  - $P \ll Q$ gdw. $P = fQ$ für eine W'dichte $f$ auf $(M, \mathcal{M}, Q)$, d.h. $P(A) = \int_A f dQ$ für alle $A \in \mathcal{M}$. Bezeichnung: $f = \frac{dP}{dQ}$ (Satz von Radon-Nikodym).
- *Jensensche Ungleichung*: Ist $J \subseteq \mathbb{R}$ ein Interval, $X$ eine integrierbare ZV mit Werten in $J$ und $h : J \to \mathbb{R}$ strikt konvex, so ist

  $$\int h(X) dP \geqslant h\left(\int X dP\right)$$

  mit Gleichheit genau dann, wenn $X$ $P$-f.s. konstant gleich $\int X dP$ ist.[12, 3.2.16(c)]

# 2 Entropie

Der Begriff *Entropie* wurde 1865 von *Rudolf Clausius* im Rahmen thermodynamischer Untersuchungen an idealen reversiblen Wärmemaschinen eingeführt. Der momentane Zustand einer solchen Maschine wird durch wenige makroskopische Größen wie Volumen $V$, Temperatur $T$, im System vorhandene Wärmemenge $Q$ u.ä. beschrieben, und die Thermodynamik beschreibt Zusammenhänge zwischen solchen Größen, z.B. $\frac{\delta Q}{T} \geqslant 0$ bei Prozessen, die in isolierten Systemen ablaufen. Wegen ihrer Bedeutung bezeichnete Clausius die Größe $\frac{Q}{T}$ als Entropie, später wurde das präzisiert, indem man $dS = \frac{\delta Q}{T}$ als Entropieänderung definierte. Natürlich ist das eigentlich viel komplizierter, und da ich selbst kein Experte für Thermodynamik bin, verweise ich auf weitere Quellen [1], [2], [20]. Wichtig ist, dass es sich um eine Theorie für *makroskopische* Größen handelt, die Systeme mit extrem vielen mikroskopischen Freiheitsgraden beschreiben, z.B. Gase mit $10^{23}$ Molekülen pro Liter. *Ludwig Boltzmann* [3] und andere haben aufgezeigt, wie wichtige Aspekte der Thermodynamik mit Hilfe statistischer Überlegungen aus mikroskopischen Eigenschaften des Systems hergeleitet werden können. Dabei muss man sich von der Betrachtung individueller Mikrozustände etwas lösen und sich überlegen, durch wieviele verschiedene Mikrozustände ein Makrozustand definiert werden kann. Systeme mit extrem vielen Freiheitsgraden werden sich dann fast immer in einem Makrozustand befinden, der durch eine überwältigende Zahl von Mikrozuständen realisiert werden kann. Das einfachste mathematische Beispiel dieser Art lernen wir am Ende dieses Abschcnitts kennen.

Will man diesen Überlegungen eine solide mathematische Basis geben, so gelangt man zu unserem Thema *Entropie und Große Abweichungen*. Zunächst einmal führen wir dabei Entropie als eine Größe ein, die den Abstand zweier Wahrscheinlichkeitsverteilungen voneiander misst.

## 2.1 Entropie von Wahrscheinlichkeitsvektoren

**Ziel:** Ordne jeder ZV $X$, die nur endlich viele Werte annimmt, eine reelle Zahl zu, die den Informationsgehalt von $X$ in folgendem Sinn misst: Vor der Realisierung misst Sie die Unsicherheit, die durch die Realisierung in einen Informationsgewinn übergeht. Sie misst also den erwarteten Informationsgewinn. Beachte, dass die Verteilung $P_X$ ein W'vektor ist.

**Definition 2.1** *Sei $p = (p_1, \ldots, p_n)$ ein W'vektor.*

$$H(p) := -\sum_i p_i \log p_i = -\sum_i \varphi(p_i) \,\text{mit}\, \varphi(t) = t \log t \,\text{und}\, \varphi(0) = 0$$

*ist die **Entropie** von $p$. Ist $X$ eine ZV, die $n$ Werte annimmt, o.B.d.A. die Werte $1, \ldots, n$, so ist $H(X) := H(P_X)$.*

**Bemerkung 2.2** $\varphi : [0, \infty) \to \mathbb{R}$ ist stetig und $\min \varphi = -\frac{1}{e}$.

**Aufgabe 2.1** Überzeugen Sie sich davon, insbesondere von der Stetigkeit an der Stelle 0. Zeigen Sie auch, dass $\varphi$ strikt konvex ist.

**Satz 2.3** ***(Fundamentale Eigenschaften von $H$)***

*i. $H$ ist für jedes feste $n$ eine symmetrische Funktion.*

*ii. $H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n)$*

*iii. $(p_1, p_2) \mapsto H(p_1, p_2)$ ist stetig.*

*iv. $H(X, Y) \leqslant H(X) + H(Y)$ mit Gleichheit, falls $X$ und $Y$ unabhängige ZVn.*

*Beweis:* i) und ii) sind offensichtlich, iii) folgt aus Bemerkung 2.2, iv) später. □

Die Eigenschaften i)-iv) einer auf W'vektoren definierten Funktion werden oft als wesentlich für die Messung von Unsicherheit in einer ZV oder Verteilung angesehen. Es wurde gezeigt [7]:

**Satz 2.4** *$H$ wird durch i)–iv) bis auf einen konstanten Faktor (d.h. bis auf die Basis des Logarithmus) eindeutig bestimmt.*

Einen Überblick in die weit verzweigte Literatur zum Charakterisierungsproblem für $H$ findet man in [11].

Eigenschaft iv) aus Satz 2.3 und einiges mehr werden im folgenden Satz zusammengefasst. Dazu benötigen wir den Begriff der **bedingten Entropie**:

$$H(Y|X) := \sum_i P[X = x_i] \cdot H(P_{Y|\{X=x_i\}})$$

**Satz 2.5** ***(Weitere Eigenschaften von $H$)***

*i.* $H \geqslant 0$

*ii.* $H(p) \leqslant \log n$ *mit Gleichheit gdw.* $p = (\frac{1}{n}, \ldots, \frac{1}{n})$.

*iii.* $H(X,Y) = H(X) + H(Y|X)$

*iv.* $H(Y|X) \leqslant H(Y)$ *mit Gleichheit gdw. $X$ und $Y$ unabhängig.*

*v.* $H(X,Y) \leqslant H(X) + H(Y)$ *mit Gleichheit gdw. $X$ und $Y$ unabhängig.*

*Beweis:*

i. Offensichtlich, da $0 \leqslant p_i \leqslant 1$.

ii. $\varphi$ ist strikt konvex ($\varphi'' > 0$), also folgt aus der Jensenschen Ungleichung

$$H(p) = -n \sum_i \frac{1}{n} \varphi(p_i) \leqslant -n\varphi\left(\sum_i \frac{p_i}{n}\right) = -n\varphi(\frac{1}{n}) = n\frac{1}{n}\log(n) = \log(n)$$

mit Gleichheit genau dann wenn $p_1 = p_2 = \ldots = p_n$, also wenn $p_i = \frac{1}{n}$ für alle $i$.

iii. O.B.d.A. sei $P_X = (p_1, \ldots, p_m)$ und $P_Y = (q_1, \ldots, q_n)$. Sei außerdem $P_{X,Y} = (r_{11}, \ldots, r_{mn})$, d.h. $P(X = i, Y = j) = r_{ij}$. Dann ist

$$\begin{aligned} H(Y|X) &= -\sum_i p_i \sum_j \frac{r_{ij}}{p_i} \log \frac{r_{ij}}{p_i} \\ &= -\sum_{i,j} r_{ij} \log r_{ij} + \sum_{i,j} r_{ij} \log p_i \\ &= H(X,Y) + \sum_i p_i \log p_i \\ &= H(X,Y) - H(X) \end{aligned}$$

iv. Mit der Notation von iii. folgt aus der Jensenschen Ungleichung

$$\begin{aligned} H(Y|X) &= -\sum_{i,j} r_{ij} \log \frac{r_{ij}}{p_i} \\ &= -\sum_{i,j} r_{ij} \log q_j - \sum_{i,j} r_{ij} \log \frac{r_{ij}}{p_i q_j} \\ &= -\sum_j q_j \log q_j - \sum_{i,j} p_i q_j \varphi\left(\frac{r_{ij}}{p_i q_j}\right) \\ &\leqslant H(Y) - \varphi\left(\sum_{i,j} p_i q_j \frac{r_{ij}}{p_i q_j}\right) \\ &= H(Y) - \varphi(1) \\ &= H(Y) \end{aligned}$$

mit Gleichheit genau dann, wenn alle $\frac{r_{i\,j}}{p_i q_j}$ den gleichen Wert $w$ annehmen, d.h. wenn $r_{ij} = wp_iq_j$ gilt. Summation über $i$ und $j$ ergibt dann $1 = w$. Daher tritt Gleichheit genau dann auf, wenn $X$ und $Y$ unabhängig sind.

v. Folgt aus iii und iv.

□

Eine informative, eher elementar gehaltene Web-Seite zur wahrscheinlichkeitstheoretischen Entropie ist [4].

**Bemerkung 2.6** Die Differenz $I(X;Y) := H(Y) - H(Y|X)$ aus Punkt iv wird als **Transinformation** von $X$ und $Y$ bezeichnet.

**Aufgabe 2.2** Zeigen sie:

a) $I(X;Y) = 0 \Leftrightarrow X$ und $Y$ unabhängig.

b) $I(X;Y) = H(X) + H(Y) - H(X,Y)$, insbesondere $I(X;Y) = I(Y;X)$.

**Aufgabe* 2.3** Zeigen sie: $I(X;Y) = H(Y) \Leftrightarrow$ es gibt eine Abb. $f$ so dass $Y = f(X)$.

Bezeichne $\mathrm{GV}_n$ die Gleichverteilung auf $\{1,\ldots,n\}$. Dann besagt Aussage ii, dass $H(p) \leqslant H(\mathrm{GV}_n)$ mit Gleichheit gdw. $p = \mathrm{GV}_n$. Deshalb scheint folgendes Maß für die Abweichung von $p$ von der Gleichverteilung interessant:

$$\begin{aligned} D(p\|\,\mathrm{GV}_n) &:= H(\mathrm{GV}_n) - H(p) = \sum_i p_i(\log n + \log p_i) \\ &= \sum_i p_i \log \frac{p_i}{1/n} = E_p\left[\log \frac{p_i}{1/n}\right] = \sum_i \frac{1}{n}\frac{p_i}{1/n} \log \frac{p_i}{1/n} \qquad (1) \\ &= \sum_i \frac{1}{n}\varphi\left(\frac{p_i}{1/n}\right) = E_{\mathrm{GV}_n}\left[\varphi\left(\frac{p_i}{1/n}\right)\right]. \end{aligned}$$

Beachte, dass $\frac{p_i}{1/n}$ die Dichte des W'maßes $p$ zum W'maß $\mathrm{GV}_n$ auf $\{1,\ldots,n\}$ ist. Daher ist

$$D(p\|\,\mathrm{GV}_n) = \int \log \frac{dp}{d\,\mathrm{GV}_n} dp = \int \varphi\left(\frac{dp}{d\,\mathrm{GV}_n}\right) d\,\mathrm{GV}_n\,.$$

Aus Satz 2.5 folgt sofort:

- $D(p\|\,\mathrm{GV}_n) \geqslant 0$ mit Gleichheit gdw. $p = \mathrm{GV}_n$.

Analoge Aussagen zu weiteren Punkten in Satz 2.5 werden - in einem allgemeineren Rahmen - im nächsten Abschnitt hergeleitet.

**Beispiel 2.7 (Entropie und Bionomialkoeffizienten)**
Stirling-Formel: $n! = n^n e^{-n+O(\log n)}$. Also

$$\begin{aligned} \binom{n}{k} &= \frac{n!}{k!(n-k)!} = \frac{n^n}{k^k(n-k)^{n-k}} e^{O(\log n)} \\ &= \left(\left(\frac{k}{n}\right)^{k/n}\left(1-\frac{k}{n}\right)^{1-k/n}\right)^{-n} e^{O(\log n)} \\ &= \exp(\log(\ldots\ldots) + O(\log n)) \\ &= \exp\left(-n\left[\varphi\left(\frac{k}{n}\right) + \varphi\left(1-\frac{k}{n}\right)\right] + O(\log n)\right) \\ &= \exp\left(n \cdot H\left(\frac{k}{n}\right) + O(\log n)\right), \end{aligned}$$

wobei $H\left(\frac{k}{n}\right)$ abkürzend für $H\left(\frac{k}{n}, 1-\frac{k}{n}\right)$ steht.

**Beispiel 2.8 (Große Abweichungen für die Binomialverteilung)**
Seien $\xi_1, \ldots, \xi_n$ u.i.v. ZVn mit $P\{\xi_i = 0\} = \frac{1}{2} = P\{\xi_i = 1\}$ und sei $X_n := \xi_1 + \ldots + \xi_n$. $X_n$ ist also binomialverteilt mit Parametern $n$ und $\frac{1}{2}$. Für $I = (a, b) \subset [0, 1]$ ist

$$P\left\{\frac{1}{n}X_n \in I\right\} \quad = \quad \sum_{k=0}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n 1_{\{k/n \in I\}}$$
$$\begin{cases} \leqslant & 2^{-n} \cdot n \cdot \max\left\{\binom{n}{k} : a < \frac{k}{n} < b\right\} \\ \geqslant & 2^{-n} \cdot \max\left\{\binom{n}{k} : a < \frac{k}{n} < b\right\} \end{cases},$$

also

$$\begin{aligned} \frac{1}{n}\log P\left\{\frac{1}{n}X \in I\right\} &\to -\log 2 + \lim_{n\to\infty} \frac{1}{n} \log \max\left\{\exp\left(n \cdot H\left(\frac{k}{n}\right) + O(\log n)\right) : a < \frac{k}{n} < b\right\} \\ &= -\log 2 + \lim_{n\to\infty} \left(\max\left\{H\left(\frac{k}{n}\right) : a < \frac{k}{n} < b\right\} + O\left(\frac{\log n}{n}\right)\right) \\ &= -H(\mathrm{GV}_2) + \sup\{H(x) : x \in I\} \\ &= -\left(H(\mathrm{GV}_2) + \inf\{-H(x) : x \in I\}\right) \\ &= -\inf\{H(\mathrm{GV}_2) - H(x) : x \in I\} \\ &= -\inf_{x\in I} D((x, 1-x) \| \mathrm{GV}_2) \end{aligned}$$

Für das abgeschlossene Intervall $I = [a, b]$ erhält man genau das gleiche Ergebnis.

**Aufgabe* 2.4** Verallgemeinern Sie obiges Beispiel für binomialverteiltes $X$ mit Parametern $n$ und $p \in (0, 1)$ (statt $p = \frac{1}{2}$). Begründen Sie insbesondere den Übergang von der viert- zur drittletzten Zeile genau!

## 2.2 Relative Entropie, Kullback-Leibler Divergenz

Die Ausdrücke für $D(p\| \mathrm{GV}_n)$ in (1) lassen sich weitgehend verallgemeinern:

**Definition 2.9** *Sei $(M, \mathcal{M})$ ein messbarer Raum, $P, Q$ W'maße auf $(M, \mathcal{M})$. Definiere*

$$D(P\|Q) := \begin{cases} \int \log \frac{dP}{dQ} dP = \int \varphi\left(\frac{dP}{dQ}\right) dQ & \text{falls } P \ll Q \\ +\infty & \text{sonst.} \end{cases}$$

*$D(P\|Q)$ heißt **relative Entropie, Kullback-Leibler Divergenz, Informationsdivergenz**, oder ... . Da $\varphi \geqslant -\frac{1}{e}$ ist, ist diese Größe immer wohldefiniert.*

*Haben $P$ und $Q$ Dichten $f$ bzw. $g$ bzgl. eines $\sigma$-endlichen Referenzmaßes $\mu$ auf $(M, \mathcal{M})$ und ist $\{f > 0\} \subseteq \{g > 0\}$, so ist $P = f\mu = \frac{f}{g} g\mu = \frac{f}{g} Q$, also $P \ll Q$ mit $\frac{dP}{dQ} = \frac{f}{g}$, so dass*

$$D(P\|Q) = \int f \log \frac{f}{g} d\mu = \int g\varphi\left(\frac{f}{g}\right) d\mu. \tag{2}$$

*In leichter Erweiterung dieser Definitionen schreiben wir für $P = f\mu$ auch*

$$D(P\|\mu) = \int f \log f d\mu = \int \varphi(f) d\mu = \int \log f dP,$$

*falls dieses Integral wohldefiniert ist (auch wenn $\mu$ kein Wahrscheinlichkeitsmaß ist).*

Dabei können Sie z.B. an $M = \mathbb{R}^d$ und $\mu$ =Lebesgue-Maß denken.

**Aufgabe 2.5** Überzeugen Sie sich im Detail von der Gültigkeit von Gleichung (2).

Einen Überblick und Literaturangaben zur Kullback-Leibler-Divergenz findet man auf [5]. Siehe auch [12].

**Satz 2.10** *$D(P\|Q) \geqslant 0$ mit Gleichheit gdw. $P = Q$.*

*Beweis:* Sei o.B.d.A. $D(P\|Q) < \infty$. Wegen der Jensenschen Ungleichung ist dann $D(P\|Q) = \int \varphi\left(\frac{dP}{dQ}\right) dQ \geqslant \varphi\left(\int \frac{dP}{dQ} dQ\right) = \varphi(1) = 0$ mit Gleichheit gdw. $\frac{dP}{dQ}$ $Q$-f.s. konstant, also $Q$-f.s. 1 ist. □

**Satz 2.11** *Seien $X, Y$ ZVn, sei $\mu$ das Lebesgue-Maß auf $\mathbb{R}$, und sei $P_X = f\mu$, $P_Y = g\mu$, $P_{X,Y} = h\mu^2$ und $h_u(v) = \frac{h(u,v)}{f(u)}$. ($h_u$ ist also die bedingte W'dichte von $Y$ gegeben $X = u$ und $f(u) = \int h(u,v)dv$, $g(v) = \int h(u,v)du$.) Dann ist*

$$D(P_{X,Y}\|\mu^2) = D(P_X\|\mu) + D(P_{Y|X}\|\mu)$$

*falls diese drei Größen $> -\infty$ (und damit auch wohldefiniert) sind, wobei $D(P_{Y|X}\|\mu) = \int_{\mathbb{R}^2} f(u)\varphi(h_u(v))d\mu^2(u,v)$.*

*Es gilt eine Verallgemeinerung für Zufallsgrößen mit Werten in allgemeinen Maßräumen.*

*Beweis:* Da $D(P_{X,Y}\|\mu^2) > -\infty$ ist, kann man den Satz von Fubini anwenden: Da $\int h_u(v)dv = \frac{1}{f(u)} \int h(u,v)dv = 1$, gilt

$$\begin{aligned}
D(P_{X,Y}\|\mu^2) &= \int h \log h d\mu^2 = \int\int h(u,v) \log h(u,v) dv du \\
&= \int_{\{f(u)>0\}} f(u) \left( \int h_u(v)(\log f(u) + \log h_u(v)) dv \right) du \\
&= \int_{\{f(u)>0\}} f(u) \left( \log f(u) + \int h_u(v) \log h_u(v) dv \right) du \\
&= \int f(u) \log f(u) du + \int\int f(u)\varphi(h_u(v)) dv du \\
&= D(P_X\|\mu) + D(P_{Y|X}\|\mu)
\end{aligned}$$

□

**Aufgabe 2.6** Überzeugen Sie sich, dass alle Integrale und Umformungen im letzten Beweis wohldefiniert sind.

**Satz 2.12** *Unter den Annahmen des letzten Satzes gilt:*

$$D(P_{X,Y}\|\mu^2) \geqslant D(P_X\|\mu) + D(P_Y\|\mu).$$

*Ist $D(P_{X,Y}\|\mu^2) < \infty$, so tritt Gleichheit auf gdw. $X$ und $Y$ unabhängig sind.*

*Beweis:* Wegen der Jensenschen Ungleichung ist für jedes $v$

$$\int_{\mathbb{R}} \varphi(h_u(v)) f(u) du \geqslant \varphi\left( \int_{\mathbb{R}} h_u(v) f(u) du \right) = \varphi\left( \int_{\mathbb{R}} h(u,v) du \right) = \varphi(g(v))$$

mit Gleichheit genau dann wenn $h_u(v) = \int_{\mathbb{R}} h_{\tilde{u}}(v) f(\tilde{u}) d\tilde{u} = \int_{\mathbb{R}} h(\tilde{u}, v) d\tilde{u} = g(v)$ für $\mu$-f.a. $u$. Also ist

$$\begin{aligned}
D(P_{Y|X}\|\mu) &= \int_{\mathbb{R}} f(u) \left( \int_{\mathbb{R}} \varphi(h_u(v)) dv \right) du \\
&= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} f(u)\varphi(h_u(v)) du \right) dv \\
&\geqslant \int_{\mathbb{R}} \varphi(g(v)) dv = D(P_Y\|\mu)
\end{aligned}$$

mit Gleichheit genau dann, wenn $h_u(v) = g(v)$ für $\mu \times \mu$-f.a. $(u, v)$, d.h. wenn $h(u, v) = f(u)g(v)$. Die Behauptung folgt jetzt aus Satz 2.11. □

**Aufgabe* 2.7** Sei $(M, \mathcal{M}, \mu)$ ein W'raum, $P, Q$ weitere W'maße auf $(M, \mathcal{M})$. Zeigen Sie für $0 < \alpha < 1$:

$$D(\alpha P + (1-\alpha)Q\|\mu) \leqslant \alpha D(P\|\mu) + (1-\alpha)D(Q\|\mu).$$

# 3 Gibbs-Verteilungen

Wie wählt man geeignete Verteilungen, wenn man zufällige Effekte modellieren will? Eine Richtschnur ist das *Prinzip der maximalen Entropie* [6], das wir hier zu einem Prinzip der *minimalen relativen Entropie* zu einem Referenzmaß $\mu$ erweitern. Ist das Referenzmaß die Gleichverteilung auf einer endlichen Menge, so sind beide Prinzipien äquivalent. In Situationen mit kontinuierlichem Beobachtungsraum kann es sich aber auch um das (normierte) Lebesgue-Maß auf einem Würfel, einer Kugeloberfläche o.ä. handeln, und auch $\sigma$-endliche $\mu$ (wie das Lebesguemaß auf $\mathbb{R}$) als Referenzmaß können sinnvoll sein.

Daher nehmen wir jetzt an, dass $(M, \mathcal{M}, \mu)$ ein $\sigma$-endlicher Maßraum ist, z.B. $M = [0,1]$ oder wieder $M = \mathbb{R}$ und $\mu$ =Lebesgue-Maß. $\mu$ ist für uns eine „Gleichverteilung", ist es ein W'maß, so ist eine ZV $X$ mit $P_X = \mu$ maximal zufällig. Wollen wir aber eine Situation modellieren, in der z.B. der Erwartungswert $E$ einer $[0,\infty)$-wertigen ZVn bekannt ist – aber sonst nichts, so suchen wir ein Wahrscheinlichkeitsmaß $P \ll \mu$ auf $[0,\infty)$, das $D(P\|\mu)$ unter der Nebenbedingung $\int x dP(x) = E$ minimiert.

Allgemeiner sollen die Erwartungswerte mehrerer *Observablen* (d.h. $\mu$-f.s. endlicher, messbarer Funktionen) $T_1, \dots T_d : M \to \mathbb{R}$ vorgegeben sein. Sei daher für $\gamma = (\gamma_1, \dots, \gamma_d)$

$$\begin{aligned}
\mathcal{D} &:= \{f : f \text{ W'dichte bzgl. } \mu\} \\
\mathcal{D}_\gamma &:= \left\{ f \in \mathcal{D} : T \in L^1_{f\mu}, \int T f d\mu = \gamma \right\} \\
&= \left\{ f \in \mathcal{D} : T_i \in L^1_{f\mu}, \int T_i f d\mu = \gamma_i \forall i = 1, \dots, d \right\}
\end{aligned}$$

Zur Abkürzung schreiben wir $D(f\|\mu) := D(f\mu\|\mu)$ für $f \in \mathcal{D}$. Gesucht ist dann:

$$f \in \mathcal{D}_\gamma \text{ mit } D(f\|\mu) = \min\{D(g\|\mu) : g \in \mathcal{D}_\gamma\}.$$

Unter geeigneten Integrabilitätsannahmen an die $T_i$ wird sich herausstellen, dass das gesuchte $f$ von der Form

$$f_\vartheta := \exp(-\psi(\vartheta) + \langle \vartheta, T\rangle) := \exp\left(-\psi(\vartheta) + \sum_{i=1}^d \vartheta_i T_i\right) \quad \text{für ein } \vartheta \in \mathbb{R}^d \text{ ist,}$$

wobei $\psi : \mathbb{R}^d \to (-\infty, \infty]$, $\psi(\vartheta) := \log \int \exp\langle \vartheta, T\rangle d\mu$ so gewählt ist, dass $\int f_\vartheta d\mu = 1$. Bezeichne

$$\begin{aligned}
\Theta &:= \left\{ \vartheta \in \mathbb{R}^d : \int \exp\langle \vartheta, T\rangle \, d\mu < \infty \right\} \\
&= \left\{ \vartheta \in \mathbb{R}^d : f_\vartheta \text{ existiert als Wahrscheinlichkeitsdichte zu } \mu \right\}.
\end{aligned}$$

**Bemerkung 3.1** $\Theta$ ist eine konvexe Menge.

**Aufgabe 3.1** Beweisen Sie diese Bemerkung.

**Definition 3.2** *Für $\vartheta \in \Theta$ ist $P_\vartheta := f_\vartheta \mu$ eine **Gibbs-Verteilung**.*

**Beispiel 3.3**

a) Die Bernoulli-Maße mit Parameter $p \in (0,1)$ auf $M = \{0,1\}^n$ schreibt man folgendermaßen als Gibbs-Verteilungen: Sei $T(\omega) = \sum_{k=1}^n \omega_k$. Dann ist $p^{T(\omega)}(1-p)^{n-T(\omega)}$ die Dichte des Bernoulli-Maßes mit Parameter $p$ zum Zählmaß $\mu$ auf $M$, denn $\sum_{\omega \in M} p^{T(\omega)}(1-p)^{n-T(\omega)} = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = 1$. Diese Dichte schreiben wir um:

$$\begin{aligned}
p^{T(\omega)}(1-p)^{n-T(\omega)} &= (1-p)^n \left(\frac{p}{1-p}\right)^{T(\omega)} = e^{-(-n\log(1-p)) + \log\frac{p}{1-p} \cdot T(\omega)} \\
&= e^{-\psi(\vartheta) + \vartheta \cdot T(\omega)} = f_\vartheta(\omega)
\end{aligned}$$

wobei $\vartheta = \log \frac{p}{1-p}$, also $p = \frac{e^\vartheta}{1+e^\vartheta}$, und $\psi(\vartheta) = -n \cdot \log(1-p) = n \cdot \log(1+e^\vartheta)$. Es ist $\Theta = \mathbb{R}$.

b) Die Normalverteilung mit Erwartungswert $m$ und Varianz $\sigma^2$ schreibt man folgendermaßen als Gibbs-Verteilung: Sei $\mu$ das Lebesgue-Maß auf $\mathbb{R}$. Dann ist

$$\begin{aligned}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) &= \exp\left(-\big(\log\sqrt{2\pi\sigma^2} + \frac{m^2}{2\sigma^2}\big) + \frac{m}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right) \\ &= \exp\left(-\psi(\vartheta) + \vartheta_1 T_1(x) + \vartheta_2 T_2(x)\right) = f_\vartheta(x),\end{aligned}$$

wobei $T_1(x) = x$, $T_2(x) = x^2$, $\vartheta_1 = \frac{m}{\sigma^2}$, $\vartheta_2 = -\frac{1}{2\sigma^2}$ und $\psi(\vartheta) = \log\sqrt{2\pi\sigma^2} + \frac{m^2}{2\sigma^2} = \log\sqrt{-\pi/\vartheta_2} - \frac{\vartheta_1^2}{4\vartheta_2}$. Es ist $\Theta = \mathbb{R} \times (-\infty, 0)$.

**Aufgabe 3.2** Schreiben Sie die Exponentialverteilungen, die Poissonverteilungen und die Gamma-Verteilungen als Gibbs-Verteilungen zu geeigneten Referenzmaßen $\mu$.

**Satz 3.4** ***(Momente von Gibbs-Verteilungen)*** *Sei $\mu$ ein $\sigma$-endliches Maß auf $(M, \mathcal{M})$ und seien $T_1, \dots, T_n$ Observablen.*

*Für $\vartheta \in \Theta$ mögen $E_\vartheta$ und $Cov_\vartheta$ den Erwartungswert bzw. die Kovarianz bzgl. des Wahrscheinlichkeitsmaßes $P_\vartheta = f_\vartheta \mu$ bezeichnen. Dann ist $T \in L^2_{P_\vartheta}$, und es gilt*

1. *Für $j = 1, \dots, d$ und $\vartheta \in \mathring{\Theta}$ ist*

$$\frac{\partial}{\partial\vartheta_j}\psi(\vartheta) = \int_M T_j f_\vartheta \, d\mu = E_\vartheta[T_j], \text{ also } D\psi(\vartheta) = E_\vartheta[T] =: \Gamma(\vartheta).$$

2. *Für $i, j = 1, \dots, d$ und $\vartheta \in \mathring{\Theta}$ ist*

$$\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\psi(\vartheta) = \mathrm{Cov}_\vartheta(T_i, T_j), \text{ also } D^2\psi(\vartheta) = \mathrm{Var}_\vartheta(T) = D\Gamma(\vartheta).$$

$\mathrm{Var}_\vartheta(T) = (\mathrm{Cov}_\vartheta(T_i, T_j))_{ij}$ *ist eine positiv semidefinite Matrix. Sie ist positiv definit genau dann, wenn*

> *die Familie $\{1, T_1, \dots, T_d\}$ im Raum der $\mu$-Äquivalenzklassen messbarer Funktionen linear unabhängig ist.* (3)

*(Aus dieser Bedingung sieht man, dass die Matrix entweder für alle $\vartheta \in \mathring{\Theta}$ oder für keines positiv definit ist.)*

*Die Funktion $\vartheta \mapsto \gamma(\vartheta)$ ist sogar unendlich of differenzierbar in $\mathring{\Theta}$.*

*Beweis:*

1. Da

$$\psi(\vartheta) = \log \int_M \exp\langle\vartheta, T\rangle d\mu,$$

folgt aus dem Satz von der Differenzierbarkeit parameterabhängiger Integrale

$$\begin{aligned}\frac{\partial}{\partial\vartheta_j}\psi(\vartheta) &= e^{-\psi(\vartheta)} \cdot \frac{\partial}{\partial\vartheta_j} \int \exp\left(\sum_{i=1}^d \vartheta_i T_i\right) d\mu = e^{-\psi(\vartheta)} \cdot \int T_j \exp\langle\vartheta, T\rangle d\mu \qquad (4) \\ &= \int T_j f_\vartheta \, d\mu = E_\vartheta[T_j]\end{aligned}$$

sobald wir die Integrabilitätsbedingung dieses Satzes nachgewiesen haben: Sei dazu $\tilde{\vartheta} \in \mathring{\Theta}$. Es gibt $\eta > 0$ derart, dass $\tilde{\vartheta} + te_j \in \Theta$ für alle $t \in \mathbb{R}$ mit $|t| \leqslant \eta$. Sei nun $|t| < \frac{\eta}{2}$. Dann ist

$$T_j \, \exp\left(\sum_i (\tilde{\vartheta} + te_j)_i T_i\right) = \frac{1}{t - (\pm\eta)} \cdot \exp\langle\tilde{\vartheta} \pm \eta e_j, T\rangle \cdot (t - (\pm\eta)) T_j e^{(t-(\pm\eta))T_j}.$$

Betrachte nun zunächst den Fall $T_j(\omega) \geq 0$. Dann wenden wir diese Zerlegung für $+\eta$ an und erhalten

$$\begin{aligned}\left|\frac{\partial}{\partial\vartheta_j}e^{\sum_i \vartheta_i T_i(\omega)}\right|_{|\vartheta=\tilde{\vartheta}+te_j} &= \left|T_j(\omega)e^{\sum_i \vartheta_i T_i(\omega)}\right|_{|\vartheta=\tilde{\vartheta}+te_j} \\ &\leqslant \frac{2}{\eta}\cdot e^{\langle\tilde{\vartheta}+\eta e_j,T\rangle}\cdot\left|(t-\eta)T_j(\omega)\, e^{(t-\eta)T_j(\omega)}\right| \\ &\leqslant \frac{2}{\eta e}\cdot e^{\langle\tilde{\vartheta}+\eta e_j,T\rangle}\end{aligned}$$

denn $\max_{x\leq 0}|xe^x| = \max_{x\geq 0} xe^{-x} = e^{-1}$. Ist $T_j(\omega) \leq 0$, so wenden wir dieselbe Überlegung auf $-\eta$ an und erhalten schließlich für alle $|t| < \frac{\eta}{2}$

$$\left|\frac{\partial}{\partial\vartheta_j}e^{\sum_i \vartheta_i T_i}\right|_{|\vartheta=\tilde{\vartheta}+te_j} \leq \frac{2}{\eta e}\cdot\max\{e^{\langle\tilde{\vartheta}+\eta e_j,T\rangle}, e^{\langle\tilde{\vartheta}-\eta e_j,T\rangle}\} \in L^1_\mu.$$

2. Unter Berücksichtigung der Formel für $\frac{\partial}{\partial\vartheta_j}\psi(\vartheta)$ folgt ähnlich wie im vorherigen Teil (aber mit etwas mehr Aufwand bei der Überprüfung der Integrabilitätsvoraussetzung):

$$\begin{aligned}\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\psi(\vartheta) &= \frac{\partial}{\partial\vartheta_i}E_\vartheta[T_j] = \frac{\partial}{\partial\vartheta_i}\int e^{-\psi(\vartheta)}T_j e^{\langle\vartheta,T\rangle}d\mu \\ &= \frac{\partial}{\partial\vartheta_i}(e^{-\psi(\vartheta)})\cdot\int T_j e^{\langle\vartheta,T\rangle}d\mu + e^{-\psi(\vartheta)}\cdot\frac{\partial}{\partial\vartheta_i}\int T_j e^{\sum_{k=1}^d\vartheta_k T_k}d\mu \\ &= -E_\vartheta[T_i]E_\vartheta[T_j] + E_\vartheta[T_iT_j] = \mathrm{Cov}_\vartheta(T_i,T_j).\end{aligned}$$

Diese Matrix der 2. Ableitungen ist als Kovarianzmatrix positiv semidefinit, denn für $\lambda\in\mathbb{R}^d$ ist

$$\lambda^t\,\mathrm{Var}_\vartheta(T)\lambda = \sum_{i=i}^d\sum_{j=1}^d\lambda_i\lambda_j\,\mathrm{Cov}_\vartheta(T_i,T_j) = \mathrm{Var}\left(\sum_{i=1}^d\lambda_i T_i\right) = \mathrm{Var}_\vartheta(\langle\lambda,T\rangle)\geqslant 0$$

mit Gleichheit genau dann, wenn $\langle\lambda,T\rangle$ $f_\vartheta\mu$-f.s. konstant ist. Da $f_\vartheta > 0$ ist, ist positive Definitheit, d.h. strikte Positivität für alle $\lambda\in\mathbb{R}^d\setminus\{0\}$, äquivalent dazu, dass die Familie $\{1,T_1,\ldots,T_n\}$ im Raum der $\mu$-Äquivalenzklassen messbarer Funktionen linear unabhängig ist.

Ähnlich zeigt man die Existenz der höheren Ableitungen von $\psi$. □

**Korollar 3.5** *Sei $\Gamma : \mathring{\Theta}\to\mathbb{R}^d, \vartheta\mapsto E_\vartheta[T]$. ($\Gamma(\Theta)$ ist die Menge aller Erwartungswertvektoren von $T$, die unter den W'maßen $P_\vartheta = f_\vartheta\mu$ angenommen werden können.)*

*i.* $D\Gamma(\vartheta) = \mathrm{Var}_\vartheta(T)$

*ii. Unter der linearen Unabhängigkeitsbedingung (3) ist $\Gamma : \mathring{\Theta}\to\Gamma(\mathring{\Theta})$ ein Diffeomorphismus.*

*iii. Sei $d = 1$, $\Theta = \mathring{\Theta} = \mathbb{R}$ und $\mu$ ein W'maß, und es sei wieder die lineare Unabhängigkeitsbedingung (3) erfüllt. Dann ist $\Gamma(\mathring{\Theta})$ das offene Intervall $(\alpha,\beta)$ mit $\alpha := \mu\text{-ess inf}\,T$ und $\beta := \mu\text{-ess sup}\,T$.*

*Beweis:* **(von *i* und *ii*)** $D\Gamma(\vartheta) = D^2\psi(\vartheta) = \mathrm{Var}_\vartheta(T)$. Unter Annahme (ii) ist $D\Gamma(\vartheta)$ positiv definit. Damit ist $D\Gamma(\vartheta)$ invertierbar, so dass $\Gamma$ ein lokaler Diffeomorphismus und $\Gamma(\mathring{\Theta})$ offen ist. Für $\vartheta,\tilde{\vartheta}\in\mathring{\Theta}$ gilt außerdem: Sei $v = \tilde{\vartheta}-\vartheta\neq 0$. Dann ist

$$\langle v,\Gamma(\tilde{\vartheta})-\Gamma(\vartheta)\rangle = \langle v,\Gamma(\vartheta+v)\rangle - \langle v,\Gamma(\vartheta)\rangle = \int_0^1\langle v, D\Gamma(\vartheta+tv)v\rangle dt > 0,$$

also $\Gamma(\tilde{\vartheta})-\Gamma(\vartheta)\neq 0$. Also ist $\Gamma$ injektiv und damit bijektiv von $\mathring{\Theta}$ auf $\Gamma(\mathring{\Theta})$. □

**Aufgabe* 3.3** Beweisen Sie Teil *iii* des Korollars. Hier sind einige Teilschritte:

1. $\Gamma(\mathring{\Theta})$ ist ein offenes Intervall.
2. $\alpha \leqslant \Gamma(\vartheta) \leqslant \beta$ für alle $\vartheta \in \mathring{\Theta}$.
3. Zu jedem $b < \beta = \mu\text{-ess sup}\, T$ gibt es ein $\vartheta \in \mathbb{R}$, für das $\int_\Theta f_\vartheta \cdot T\, d\mu = \int_\Theta e^{\vartheta T - \gamma(\vartheta)} \cdot T\, d\mu \geqslant b$. Dazu kann man folgendermaßen vorgehen: Sei $b < b' < \beta$. Betrachte $A_0 = \{T \leqslant b\}$, $A_1 = \{b < T \leqslant b'\}$ und $A_2 = \{T > b'\}$, und beachte, dass $P_\vartheta(A_2) > 0$ (warum?).
   - Zeigen Sie: $\limsup_{\vartheta\to\infty} \frac{P_\vartheta(A_0)}{P_\vartheta(A_2)} = 0$.
   - Folgern Sie: $\lim_{\vartheta\to\infty} P_\vartheta(A_0) = 0$.
   - Zeigen Sie $\liminf_{\vartheta\to\infty} \int f_\vartheta \cdot T\, d\mu \geqslant b$. Zerlegen Sie dazu das Integral in die Anteile über $A_0$ und $A_1 \cup A_2$. Vielleicht wollen Sie für das Integral über $A_0$ benutzen, dass $T \cdot f_\vartheta = T\, e^T \cdot f_{\vartheta-1} \cdot e^{\psi(\vartheta-1)-\psi(\vartheta)}$.
4. Eine entsprechende Aussage gilt für $a > \alpha$.
5. Folgern Sie aus 2., 3. und 4., dass $\Gamma(\mathring{\Theta}) = (\alpha, \beta)$.

Um zu zeigen, dass $\liminf_{\vartheta\to\infty} \int_{A_0} f_\vartheta \cdot T\, d\mu \geqslant 0$, kann man folgendermaßen vorgehen: Zunächst ist

$$\int_{A_0} f_\vartheta \cdot T\, d\mu = \int_{\{T\leqslant b\}} T e^T \cdot f_{\vartheta-1}\, d\mu \cdot e^{\psi(\vartheta-1)-\psi(\vartheta)} \geqslant -\frac{1}{e} \cdot \frac{P_{\vartheta-1}(A_0)}{P_{\vartheta-1}(A_2)} \cdot P_{\vartheta-1}(A_2)\, e^{\psi(\vartheta-1)-\psi(\vartheta)}.$$

Da

$$P_{\vartheta-1}(A_2)\, e^{\psi(\vartheta-1)-\psi(\vartheta)} = \int_{\{T>b'\}} e^{-T} f_\vartheta\, d\mu \leqslant e^{-b'},$$

folgt aus $\limsup_{\vartheta\to\infty} \frac{P_\vartheta(A_0)}{P_\vartheta(A_2)} = 0$, dass in der Tat $\liminf_{\vartheta\to\infty} \int_{A_0} f_\vartheta \cdot T\, d\mu \geqslant 0$. Also ist

$$\begin{aligned}\liminf_{\vartheta\to\infty} \int f_\vartheta \cdot T\, d\mu &\geqslant \liminf_{\vartheta\to\infty} \int_{A_0} f_\vartheta \cdot T\, d\mu + \liminf_{\vartheta\to\infty} \int_{A_1\cup A_2} f_\vartheta \cdot T\, d\mu \\ &\geqslant b \cdot \liminf_{\vartheta\to\infty} P_\vartheta(A_1 \cup A_2) = b \cdot (1 - \limsup_{\vartheta\to\infty} P_\vartheta(A_0)) = b.\end{aligned}$$

**Satz 3.6** ***(Minimierung der relativen Entropie unter Nebenbedingungen)***
*Sei $\mu$ ein $\sigma$-endliches Maß auf $(M, \mathcal{M})$. Ist $\gamma = \Gamma(\vartheta)$ für ein $\vartheta \in \mathring{\Theta}$ und ein $\gamma \in \mathbb{R}^d$, so ist $f_\vartheta \in \mathcal{D}_\gamma$,*

$$D(f_\vartheta \| \mu) = \min\{D(g\|\mu) : g \in \mathcal{D}_\gamma\}$$

*und $D(f_\vartheta\|\mu) = D(g\|\mu)$ genau dann, wenn $g = f_\vartheta$ $\mu$-f.s. Explizit ist*

$$D(f_\vartheta\|\mu) = \langle \vartheta, \Gamma(\vartheta)\rangle - \psi(\vartheta).$$

*Beweis: Vorbemerkung:* Ein naiver Variationsansatz liefert für einen Minimierer $g \in \mathcal{D}_\gamma$ und jedes $\delta g$ mit $g + \delta g \in \mathcal{D}_\gamma$ unter Beachtung von $\log(g+\delta g) = \log g + \log\left(1 + \frac{\delta g}{g}\right)$:

$$\begin{aligned} 0 \leqslant D(g+\delta g\|\mu) - D(g\|\mu) &= \int (g+\delta g)\log(g+\delta g) d\mu - \int g \log g d\mu \\ &= \int g \log\left(1+\frac{\delta g}{g}\right) d\mu + \int \delta g \left(\log g + \log\left(1+\frac{\delta g}{g}\right)\right) d\mu \\ &= \int g \frac{\delta g}{g} d\mu + \int \delta g \log g d\mu + O((\delta g)^2) \\ &= 0 + \int \delta g \log g d\mu + O((\delta g)^2)\ , \end{aligned}$$

wobei die Bedeutung des Fehlerterms $O((\delta g)^2)$ nicht präzisiert wird. Für $\delta g \to 0$ wird die letzte Zeile vom Integralterm dominiert, und da man $\delta g$ in der ganzen Rechnung durch $-\delta g$ ersetzen kann, muss $\int \delta g \log g d\mu = 0$ sein für alle $\delta g$ mit $\int \delta g d\mu = 0$ und $\int \delta g T_i d\mu = 0$ $(i = 1, \ldots, d)$. In der Notation der Linearen Algebra kann man das so formulieren:

$$\log g \in (\text{span}\{1, T_1, \ldots, T_d\}^\perp)^\perp = \text{span}\{1, T_1, \ldots, T_d\},$$

so dass $\log g$ eine Linearkombination von $1, T_1, \ldots, T_d$ und daher ein $\log f_\vartheta$ sein muss. *Problematisch* an diesem ,,Beweis" ist, dass man i.A. nur weiß, dass $\int |\delta g| d\mu$ klein ist, nicht aber $\delta g \in L^2_\mu$. Insbesondere ist der Ausdruck $O((\delta g)^2)$ sicher nicht wohldefiniert.

*Nun zum formalen Beweis* (angelehnt an [10, Theorem 3.1]):

Da $\gamma = \Gamma(\vartheta) = E_\vartheta[T]$, ist auch $f_\vartheta \in \mathcal{D}_\gamma$. Unter Berücksichtigung von $\log f_\vartheta = -\psi(\vartheta) + \langle \vartheta, T\rangle$ folgt

$$D(f_\vartheta \| \mu) = \int f_\vartheta \log f_\vartheta d\mu = -\psi(\vartheta) + \int \langle \vartheta, T\rangle f_\vartheta d\mu = \langle \vartheta, \Gamma(\vartheta)\rangle - \psi(\vartheta) < \infty.$$

Sei nun $g \in \mathcal{D}_\gamma$ mit $D(g\|\mu) < \infty$. Insbesondere ist $T \in L^1_{g\mu}$, also auch $\log f_\vartheta \in L^1_{g\mu}$. Es folgt:

$$\begin{aligned}
D(g\|\mu) - D(f_\vartheta\|\mu) &= \int g \log g \, d\mu - \int f_\vartheta \log f_\vartheta \, d\mu = \int \frac{g}{f_\vartheta} \log \frac{g}{f_\vartheta} \cdot f_\vartheta \, d\mu + \int (g - f_\vartheta) \log f_\vartheta \, d\mu \\
&= \int \varphi\left(\frac{g}{f_\vartheta}\right) \cdot f_\vartheta \, d\mu + \int (g - f_\vartheta) \cdot (-\psi(\vartheta) + \langle \vartheta, T\rangle) \, d\mu \\
&\geqslant \varphi\left(\int \frac{g}{f_\vartheta} f_\vartheta d\mu\right) + \langle \vartheta, \gamma\rangle - \langle \vartheta, \gamma\rangle \\
&= \varphi(1) \\
&= 0
\end{aligned}$$

wobei in der dritten Zeile die Jensensche Ungleichung für das Wahrscheinlichkeitsmaß $f_\vartheta \mu$ benutzt wurde. Insgesamt tritt Gleichheit auf genau dann, wenn $\frac{g}{f_\vartheta}$ $f_\vartheta\mu$-f.s konstant ist. Da $f_\vartheta > 0$ ist, ist das äquivalent dazu, dass $\frac{g}{f_\vartheta}$ $\mu$-f.s constant ist, und da sowohl $f_\vartheta$ als auch $g$ Wahrscheinlichkeitsdichten zu $\mu$ sind, folgt $g = f_\vartheta$ $\mu$-f.s. □

**Beispiel 3.7** Sei $M = \{0,1\}^n$, $\mu$ das Zählmaß auf $M$ und $T(\omega) = \sum_{k=1}^n \omega_k$, siehe Beispiel 3.3a. Sei $\gamma = pn$ für ein $p \in (0,1)$. Dann minimiert die Dichte der Bernoulliverteilung zum Parameter $p$, d.h. die Dichte $b_{n,p}(\omega) = p^{T(\omega)}(1-p)^{n-T(\omega)}$, das Funktional $D(g\|\mu)$ unter der Nebenbedingung $\int Tg \, d\mu = \gamma$.

Das folgt so: Nach Satz 3.6 ist die gesuchte Dichte von der Form $f_\vartheta = e^{-\psi(\vartheta)+\vartheta T}$ für ein $\vartheta$ aus $\Theta = \mathring{\Theta} = \mathbb{R}$ . Wie in Beispiel 3.3a zeigt man, dass $\psi(\vartheta) = n\log(1+e^\vartheta)$ , so dass

$$\Gamma(\vartheta) = D\psi(\vartheta) = n\frac{e^\vartheta}{1+e^\vartheta} \quad \text{und} \quad \Gamma(\mathring{\Theta}) = \Gamma(\mathbb{R}) = (0,n).$$

Daher ist

$$D\psi(\vartheta) = \gamma = pn \text{ gdw. } \frac{e^\vartheta}{1+e^\vartheta} = p \text{ gdw. } \frac{1}{1+e^\vartheta} = 1-p,$$

so dass wir als minimierende Dichte erhalten:

$$f_\vartheta(\omega) = e^{-\psi(\vartheta)+\vartheta T(\omega)} = \left(\frac{1}{1+e^\vartheta}\right)^n e^{\vartheta T(\omega)} = p^{T(\omega)}(1-p)^{n-T(\omega)} = b_{n,p}(\omega),$$

was wieder genau mit der Form aus Beispiel 3.3a übereinstimmt.

**Aufgabe* 3.4** Zeigen Sie, dass die Exponentialverteilung zum Parameter $\lambda$ die relative Entropie zum Lebesgue-Maß $\mu$ auf $[0,\infty)$ unter der Nebenbedingung ,,Erwartungswert= $\frac{1}{\lambda}$ " minimiert.

**Aufgabe* 3.5** Zeigen Sie, dass die Normalverteilung mit Erwartungswert $m$ und Varianz $\sigma^2$ die relative Entropie zum Lebesgue-Maß $\mu$ auf $\mathbb{R}$ unter den Nebenbedingungen ,,Erwartungswert= $m$" und ,,2. Moment= $m^2 + \sigma^2$ " minimiert. Bestimmen Sie explizit die Abbildung $\Gamma$ und ihre Inverse für dieses Beispiel.

**Aufgabe* 3.6** Sei $(M, \mathcal{M}, \mu)$ ein $\sigma$-endlicher Maßraum und $T_1, \dots T_d : M \to \mathbb{R}$ Observablen, also messbare, $\mu$-f.s. endliche Funktionen. Bezeichne

$$V := \operatorname{span}(\{1, T_1, \dots, T_d\}) \quad \text{im Raum der } \mu\text{-Äquivalenzklassen messbarer Funktionen.}$$

Seien $\vartheta, \vartheta' \in \Theta$. Charakterisieren Sie mit Hilfe von $V$, wann $f_\vartheta = f'_\vartheta$ $\mu$-f.s. ist.

Anwendungen der ,,Maximum-Entropie-Methode“ in der Ökonometrie findet man in der Monographie [16], Beispiele in den Arbeiten [22] und [26]. Eine Anwendung in der Populationsgenetik, die sich in ihrer Herangehensweise an der statistischen Mechanik orientiert, ist z.B. [8].

# 4 Anwendung in der Statistik: exponentielle Familien

## 4.1 Minimierung der Kullback-Leibler Divergenz und Maximum-Likelihood

Da es sich in diesem Abschnitt um Fragestellungen der Statistik handelt, benutze ich hier die Bezeichnung *Kullback-Leibler Divergenz* statt relativer Entropie. In der parametrischen Schätz- und Testtheorie spielen *exponentielle Familien* eine große Rolle. Viele wichtige Familien von Verteilungen (z.B. die Normalverteilungen) gehören dazu.

**Definition 4.1** *(Vergl. [25, Abschn. 1.7]) Sei $(M, \mathcal{M}, \nu)$ ein $\sigma$-endlicher Maßraum. Eine Familie $(Q_\lambda : \lambda \in \Lambda)$ von W'maßen auf $(M, \mathcal{M})$ heißt* exponentielle Familie *(auch* Exponentialfamilie*), falls es ein $d \in \mathbb{N}$, $A : \Lambda \to \mathbb{R}$, $\zeta : \Lambda \to \mathbb{R}^d$ und messbare $h : M \to [0, \infty)$ und $T : M \to \mathbb{R}^d$ gibt, so dass*

$$\frac{dQ_\lambda}{d\nu}(x) = A(\lambda) \cdot h(x) \cdot \exp\langle \zeta(\lambda), T(x)\rangle.$$

Mit $\nu$ ist auch $\mu := h\nu$ $\sigma$-endlich (warum?) und $\frac{dQ_\lambda}{d\mu}(x) = A(\lambda) \cdot \exp\langle \zeta(\lambda), T(x)\rangle$. Setzt man nun $f_\vartheta := \exp(-\psi(\vartheta) + \langle \vartheta, T\rangle)$ für $\vartheta \in \Theta := \zeta(\Lambda) \subseteq \mathbb{R}^d$, so ist $\frac{dQ_\lambda}{d\mu} = f_{\zeta(\lambda)}$ bis auf Umparametrisierung die Dichte einer Gibbs-Verteilung. $\vartheta := \zeta(\lambda)$ wird als der *natürliche Parameter* der Familie bezeichnet. Also:

> Exponentialfamilien in natürlicher Parametrisierung sind bei geeigneter Wahl des Referenzmaßes dasselbe wie Familien von Gibbs-Verteilungen.

Das entspricht der in [12, Abschn. 7.1.6] gewählten Definition einer Exponentialfamilie, von der wir auch hier im Weiteren ausgehen.

**Die Idee des Schätzens (ganz allgemein und etwas vage)** Beim Schätzen geht es in der Statistik immer darum, aus einer vorgegebenen Familie von Verteilungen (dem *statistischen Modell*) eine Verteilung auszuwählen, die einen beobachteten Datensatz möglichst gut beschreibt. Handelt es sich um eine parametrisierte Familie (z.B. die Familie aller eindimensionalen Normalverteilungen mit Varianz 1), so läuft das darauf hinaus, den - evtl. mehrdimensionalen - Parameter dieser Verteilung zu bestimmen (z.B. den Erwartungswert einer Normalverteilung). Das wesentliche konzeptionelle Problem besteht darin, die Idee „einen beobachteten Datensatz möglichst gut beschreiben" mathematisch zu präzisieren. Im folgenden wird ein allgemeiner Ansatz dazu präsentiert, aus dem sich sogar die Wahl des statistischen Modells ergibt.

**Schätzung einer unbekannten Verteilung durch Minimierung der Kullback-Leibler Divergenz** Sei nun $(M, \mathcal{M}, \mu)$ ein $\sigma$-endlicher Maßraum, und seien $X_1, \ldots, X_N$ u.i.v. $M$-wertige Beobachtungen mit unbekannter Verteilung $P_X \ll \mu$.

Sind die $X_i$ nicht direkt beobachtbar, sondern kennt man nur die beobachteten Mittelwerte $\frac{1}{N}\sum_{i=1}^N T_k(X_i)$ von Observablen $T_1, \ldots, T_d : M \to \mathbb{R}$, $T = (T_1, \ldots, T_d) : M \to \mathbb{R}^d$, so kann man folgende Überlegung anstellen, um auf Basis dieser beobachteten Werte die unbekannte Verteilung $P_X$ zu schätzen: Die gesuchte Verteilung soll die empirischen (gleich: beobachteten) Mitelwerte der $T_i$ reproduzieren, darüberhinaus aber möglichst großen Zufall repräsentieren. Formal: Die geschätzte Verteilung hat diejenige Dichte $f$ bzgl. $\mu$, die die Kullback-Leibler Divergenz $D(f\|\mu)$ unter der Nebenbedingung $\int_M T f d\mu = \hat{\gamma} := \frac{1}{N}\sum_{i=1}^N T(X_i)$ minimiert. Damit ist $f$ von der Form

$$f_\vartheta = \exp(-\psi(\vartheta) + \langle \vartheta, T\rangle)$$

mit dem aus den Beobachtungen bestimmten Parameter $\vartheta = \hat{\vartheta} := \Gamma^{-1}(\hat{\gamma})$. Dabei muss man voraussetzen, dass $\hat{\gamma} \in \Gamma(\mathring{\Theta})$, dass also das beobachtete $\hat{\gamma}$ überhaupt als Erwartungswertvektor unter einer Verteilung der gewählten Familie auftreten kann.

**Satz 4.2** *Denselben Schätzwert $\hat{\vartheta} = \Gamma^{-1}(\hat{\gamma})$ erhält man durch eine* Maximum-Likelihood-Schätzung *des Parameters $\vartheta$ in der exponentiellen Familie $(f_\vartheta \mu : \vartheta \in \Theta)$.*

*Beweis:* Für eine Maximum-Likelihood-Schätzung von $\vartheta$ ist, bei gegebenen $X_1, \dots, X_N$, die Dichte der Verteilung von $(X_1, \dots, X_N)$, also die Produktdichte $\prod_{i=1}^N f_\vartheta(X_i)$ – äquivalent dazu deren Logarithmus – durch Wahl von $\vartheta$ zu maximieren:

$$\begin{aligned}\log \prod_{i=1}^N f_\vartheta(X_i) &= -N\psi(\vartheta) + \sum_{i=1}^N \langle \vartheta, T(X_i)\rangle = N\left(-\psi(\vartheta) + \left\langle \vartheta, \frac{1}{N}\sum_{i=1}^N T(X_i)\right\rangle\right) \\ &= N\left(-\psi(\vartheta) + \langle \vartheta, \hat{\gamma}\rangle\right),\end{aligned}$$

also nach Satz 3.4

$$D_\vartheta \left(\log \prod_{i=1}^N f_\vartheta(X_i)\right) = N(-\Gamma(\vartheta) + \hat{\gamma}).$$

Dieser Ausdruck wir null für $\vartheta = \Gamma^{-1}(\hat{\gamma})$. Die 2. Ableitung $-ND_\vartheta\Gamma(\vartheta) = -N \operatorname{Var}_\vartheta(T)$ ist überall negativ definit, so dass tatsächlich ein Maximum vorliegt. □

## 4.2 Kullback-Leibler Divergenz und Fisher-Information

Sei $(f_\vartheta : \vartheta \in \Theta \subseteq \mathbb{R}^d)$ eine Familie von Wahrscheinlichkeitsdichten auf dem $\sigma$-endlichen Maßraum $(M, \mathcal{M}, \mu)$ (nicht notwendig eine Exponentialfamilie). Wir nehmen an, dass die Zufallsvariablen $D_\vartheta(\log f_\vartheta) : \Omega \to \mathbb{R}^d$ bzgl. $P_\vartheta = f_\vartheta \mu$ quadratingerierbar sind und definieren die *Fisher-Informationsmatrix* als die Kovarianzmatrix dieser ZV unter $P_\vartheta$,

$$I(\vartheta) := \operatorname{Var}_\vartheta(D_\vartheta(\log f_\vartheta)).$$

$I(\vartheta)$ ist also positiv semidefinit.

Ist speziell $f_\vartheta = \exp(-\psi(\vartheta) + \langle \vartheta, T\rangle)$ eine *Exponentialfamilie*, so ist für $\vartheta \in \mathring{\Theta}$:

$$D_\vartheta(\log f_\vartheta) = (D_\vartheta(-\psi(\vartheta) + \langle \vartheta, T\rangle) = T - E_\vartheta[T]$$

und daher

$$I(\vartheta) = \operatorname{Var}_\vartheta(D_\vartheta(\log f_\vartheta))) = \operatorname{Var}_\vartheta(T). \tag{5}$$

**Satz 4.3** ***(Cramer-Rao-Ungleichung)***
*Sei $g : \Theta \to \mathbb{R}^k$ eine Funktion der natürlichen Parameter und sei die $\mathbb{R}^k$-wertige ZV $Y$ ein* erwartungstreuer Schätzer *von $g(\vartheta)$, d.h. es gelte $E_\vartheta[Y] = g(\vartheta)$ für alle $\vartheta \in \Theta$. Ist $\operatorname{Var}_\vartheta(Y_i) < \infty$ für alle $i = 1, \dots, k$ und $\vartheta \in \Theta$ und ist die Fisher-Informationsmatrix $I(\vartheta)$ invertierbar (d.h. positiv definit), so gilt unter milden Regularitätsannahmen*

$$\operatorname{Var}_\vartheta(Y) \geqslant D_\vartheta g(\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t. \tag{6}$$

*Dabei wird $D_\vartheta g(\vartheta)$ als $k \times d$-Matrix aufgefasst. Die Ungleichung ist im üblichen Sinn von ,,Linke Seite minus Rechte Seite ist positiv semidefinit" zu verstehen. Für $k = 1$ ist es also eine gewöhnliche Ungleichung zwischen Zahlen. (Der Schätzer $Y$ heißt* effizient*, falls in dieser Ungleichung für alle $\vartheta$ Gleichheit gilt.)*

*Beweis:* Siehe z.B. [12, Theorem 7.2.16] oder [25, Satz 2.124].
*Skizze:* Sei zunächst $k = 1$. Unter geeigneten Annahmen kann man folgende parameterabhängigen Integrale differenzieren (in den mit (*) markierten Gleichheiten):

$$E_\vartheta[D_\vartheta(\log f_\vartheta)] = \int f_\vartheta D_\vartheta(\log f_\vartheta) d\mu = \int D_\vartheta f_\vartheta d\mu \overset{(*)}{=} D_\vartheta \int f_\vartheta d\mu = D_\vartheta 1 = 0,$$

so dass

$$D_\vartheta g(\vartheta) = D_\vartheta \int Y f_\vartheta d\mu \overset{(*)}{=} \int Y \cdot D_\vartheta f_\vartheta d\mu = \int Y \cdot D_\vartheta(\log f_\vartheta) f_\vartheta d\mu = E_\vartheta[(Y - E_\vartheta[Y]) \cdot D_\vartheta(\log f_\vartheta)].$$

Durch Multiplikation von rechts mit dem nicht-zufälligen $d$-Vektor $I(\vartheta)^{-1} D_\vartheta g(\vartheta)^t$ folgt:

$$\begin{aligned} D_\vartheta g(\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t &= E_\vartheta[\underbrace{(Y - E_\vartheta[Y])}_{\in \mathbb{R}^k = \mathbb{R}^1} \cdot \underbrace{(D_\vartheta(\log f_\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t)}_{\in \mathbb{R}^k = \mathbb{R}^1}] \\ &\leqslant \sqrt{\mathrm{Var}_\vartheta(Y)} \sqrt{E_\vartheta[(D_\vartheta(\log f_\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t)^2]}. \end{aligned}$$

Weiterhin ist wegen der Symmetrie von $I(\vartheta)^{-1}$ und weil $D_\vartheta(\log f_\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t$ ein Skalar ist: :

$$\begin{aligned} & E_\vartheta[(D_\vartheta(\log f_\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t)^2] \\ = \; & E_\vartheta[(D_\vartheta(\log f_\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t)^t \cdot (D_\vartheta(\log f_\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t)] \\ = \; & E_\vartheta[D_\vartheta g(\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta(\log f_\vartheta)^t \cdot D_\vartheta(\log f_\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t] \\ = \; & D_\vartheta g(\vartheta) \cdot I(\vartheta)^{-1} \cdot E_\vartheta[D_\vartheta(\log f_\vartheta)^t \cdot D_\vartheta(\log f_\vartheta)] \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t \\ = \; & D_\vartheta g(\vartheta) \cdot I(\vartheta)^{-1} \cdot \mathrm{Var}_\vartheta[D_\vartheta(\log f_\vartheta)] \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t \\ = \; & D_\vartheta g(\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t, \end{aligned}$$

wobei auch $E_\vartheta[D_\vartheta(\log f_\vartheta)] = 0$ benutzt wurde. Daher:

$$\mathrm{Var}_\vartheta(Y) \geqslant D_\vartheta g(\vartheta) I(\vartheta)^{-1} D_\vartheta g(\vartheta)^t.$$

Im Fall $k > 1$ sei $u \in \mathbb{R}^k$ beliebig. Dann erfüllen $\langle u, g\rangle$ und $\langle u, Y\rangle$ die Voraussetzungen des Satzes für $k = 1$, so dass

$$\begin{aligned} u^t \cdot \mathrm{Var}_\vartheta(Y) \cdot u &= \mathrm{Var}_\vartheta(\langle u, Y\rangle) \\ &\geqslant D_\vartheta\langle u, g(\vartheta)\rangle \cdot I(\vartheta)^{-1} \cdot D_\vartheta\langle u, g(\vartheta)\rangle^t \\ &= u^t \cdot D_\vartheta g(\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t \cdot u. \end{aligned}$$

□

**Bemerkung 4.4** Ist $f_\vartheta = \exp(-\psi(\vartheta) + \langle \vartheta, T\rangle)$ eine *Exponentialfamilie* und will man $g(\vartheta) = E_\vartheta[T_i]$ schätzen, so ist $T_i$ selbst ein *effizienter* erwartungstreuer Schätzer, denn für $\vartheta \in \mathring{\Theta}$ ist $I(\vartheta) = \mathrm{Var}_\vartheta(T)$ nach (5) und

$$D_\vartheta g(\vartheta) = e_i^t \cdot D_\vartheta E_\vartheta[T] = e_i^t \cdot D_\vartheta \Gamma(\vartheta) = e_i^t \cdot \mathrm{Var}_\vartheta(T) = e_i^t \cdot I(\vartheta),$$

so dass

$$D_\vartheta g(\vartheta) \cdot I(\vartheta)^{-1} \cdot D_\vartheta g(\vartheta)^t = e_i^t \cdot I(\vartheta) \cdot I(\vartheta)^{-1} \cdot I(\vartheta) \cdot e_i = (I(\vartheta))_{i,i} = \mathrm{Var}_\vartheta(T_i).$$

**Bemerkung 4.5** Unter geeigneten Regularitäts- und Integrabilitätsannahmen besteht folgender Zusammenhang zwischen $I(\vartheta)$ und der Kullback-Leibler Divergenz: Seien $\vartheta, \vartheta' \in \mathring{\Theta}$. Dann kann man zeigen, dass

$$D(P_{\vartheta'} \| P_\vartheta) = \frac{1}{2}(\vartheta' - \vartheta)^t I(\vartheta)(\vartheta' - \vartheta) + o(\|\vartheta' - \vartheta\|^2) \quad \text{im Limes } \vartheta' \to \vartheta. \tag{7}$$

**Beispiel 4.6** Die Normalverteilungen $\mathcal{N}(m, \sigma^2)$ bilden eine Exponentialfamilie mit $T_1(x) = x$, $T_2(x) = x^2$, $\vartheta_1 = \frac{m}{\sigma^2}$, $\vartheta_2 = -\frac{1}{2\sigma^2}$. $T_1$ ist ein erwartungstreuer effizienter Schätzer von $g_1(\vartheta) = E_\vartheta[T_1] = m = -\frac{\vartheta_1}{2\vartheta_2}$, $T_2$ von $g_2(\vartheta) = E_\vartheta[T_2] = \sigma^2 + m^2 = -\frac{1}{2\vartheta_2} + \frac{\vartheta_1^2}{4\vartheta_2^2}$. (Erwartungstreues Schätzen von $\sigma^2$ wird hier nicht erfasst.)

**Aufgabe* 4.1** a) Weisen Sie Gleichung (7) zunächst für exponentielle Familien mit natürlicher Parametrisierung nach. (Das ist eine recht direkte Rechnung, die die vorher hergeleiteten Formeln benutzt.)

b) Weisen Sie die Gleichung dann für allgemeine Familien $P_\vartheta = f_\vartheta \mu$ unter geeigneten Annahmen an die Parameterabhängigkeit der $f_\vartheta$ nach. (Dabei werden Sie wahrscheinlich [12, Proposition 7.2.16] benutzen wollen.)

**Zusammengefasst** (für den Fall $g(\vartheta) = \vartheta$):

> Bei „kleiner“ Fisher-Informationsmatrix $I(\vartheta)$ hat jeder erwartungstreue Schätzer $Y$ von $\vartheta$ eine große Varianz (Gleichung (6): $\mathrm{Var}_\vartheta(Y) \geqslant I(\vartheta)^{-1}$) , weil sich Verteilungen zu benachbarten Parametern nur wenig unterscheiden (Gleichung (7): $D(P_{\vartheta'} \| P_\vartheta) = \frac{1}{2} I(\vartheta)(\vartheta' - \vartheta)^2 + o((\vartheta' - \vartheta)^2)$).
>
> Will man nicht $\vartheta$, sondern $\alpha\vartheta$ für ein festes $\alpha \in \mathbb{R} \setminus \{0\}$ schätzen, so ist $\alpha Y$ ein erwartungstreuer Schätzer, und Gleichung (7) liefert, wenig überraschend, dass $\mathrm{Var}_\vartheta(\alpha Y) \geqslant \alpha I(\vartheta)^{-1} \alpha = \alpha^2 I(\vartheta)^{-1}$.

# 5 Entropie und Konvexität

## 5.1 Halbstetigkeit

**Definition 5.1** *Eine auf einem metrischen (oder topologischen) Raum definierte Funktion* $f : E \to (-\infty, \infty]$ *heißt* ***unterhalbstetig*** *oder* ***halbstetig von unten****, engl.* lower semicontinuous*, falls* $f^{-1}((-\infty, a])$ *für jedes* $a \in \mathbb{R}$ *abgeschlossen ist.*

**Bemerkung 5.2** Indikatorfunktionen offener Mengen sind unterhalbstetig, denn für offene $G \subseteq E$ und $f = 1_G$ gilt:

$$f^{-1}((-\infty, a]) = \begin{cases} \emptyset & \text{falls } a < 0 \\ E \setminus G & \text{falls } 0 \leqslant a < 1 \\ E & \text{falls } a \geqslant 1. \end{cases}$$

**Aufgabe* 5.1** Zeigen Sie, dass auf einem metrischen Raum $E$ jede der folgenden Bedingungen zur Halbstetigkeit von unten äquivalent ist:

i. Sind $x, x_n \in E$, $\lim_{n\to\infty} x_n = x$, so ist $\limsup_{n\to\infty} f(x_n) \geqslant f(x)$.

ii. Sind $x, x_n \in E$, $\lim_{n\to\infty} x_n = x$, so ist $\liminf_{n\to\infty} f(x_n) \geqslant f(x)$.

iii. Für alle $x \in E$ ist $\lim_{\varepsilon\downarrow 0} \inf f(U_\varepsilon(x)) = f(x)$.

**Aufgabe* 5.2** a) Sind $f_n : E \to \mathbb{R}$ stetig, $f_1 \leqslant f_2 \leqslant \ldots$ und $f(x) = \sup_n f_n(x)$ für $x \in E$, so ist $f$ halbstetig von unten. (Tatsächlich lässt sich auch umgekehrt jede von unten halbstetige Funktion als punktweiser Limes einer wachsenden Folge stetiger Funktionen darstellen. Zeigen Sie auch das, unter der Zusatzannahme, dass $E$ kompakt ist.)

b) Ist $(f_\lambda | \lambda \in \Lambda)$ eine beliebige Familie stetiger Funktionen von $E$ nach $\mathbb{R}$, so wird durch $f(x) := \sup_{\lambda\in\Lambda} f_\lambda(x)$ eine unterhalbstetige Funktion $f : E \to (-\infty, \infty]$ definiert.

**Aufgabe* 5.3** Sei $X$ ein reeller Vektorraum, $F_\lambda : X \to \mathbb{R}$ $(\lambda \in \Lambda)$ lineare Funktionale und $x_\lambda \in X$ $(\lambda \in \Lambda)$. Definiere $F : X \to (-\infty, \infty]$ durch $F(x) := \sup_{\lambda\in\Lambda} (F_\lambda(x) + x_\lambda)$. Zeigen Sie, dass $F$ ein konvexes Funktional ist, d.h. dass $F(\alpha x + (1-\alpha)y) \leqslant \alpha F(x) + (1-\alpha)F(y)$ für alle $x, y \in X$ und alle $\alpha \in [0, 1]$.

## 5.2 Variationsprinzip für die relative Entropie

Sei $(M, \mathcal{M}, \mu)$ ein W'raum, insbesondere also $D(\nu\|\mu) \geqslant 0$ für alle W'maße $\nu$ auf $(M, \mathcal{M})$. Es sei $\mathrm{mb}(M)$ der Raum aller $\mathcal{M}$-messbaren Funktionen von $M$ nach $\mathbb{R}$, $\mathrm{mb}_b(M) = \{u \in \mathrm{mb}(M) : u \text{ beschränkt}\}$. Für $u \in \mathrm{mb}(M)$ definieren wir

$$\Psi(u) := \log \int e^u d\mu \in (-\infty, \infty]. \tag{8}$$

Beachte, dass $\Psi(\langle \vartheta, T\rangle) = \psi(\vartheta)$ in der Notation von Kapitel 3.

**Lemma 5.1** *a) Sei* $u \in \mathrm{mb}(M)$ *und sei* $\nu$ *ein W'maß auf* $(M, \mathcal{M})$ *mit* $\int u^+ d\nu < \infty$*. Dann ist*

$$\int u\, d\nu - D(\nu\|\mu) \leqslant \Psi(u) \tag{9}$$

*mit Gleichheit genau dann, wenn* $\Psi(u) < \infty$ *und* $\nu = e^{u-\Psi(u)}\mu$*. (Ist* $\Psi(u) < \infty$ *und* $\nu = e^{u-\Psi(u)}\mu$*, so ist* $\int u^+ d\nu < \infty$ *äquivalent zu* $D(\nu\|\mu) < \infty$*.)*

*b) Für jedes* $u \in \mathrm{mb}(M)$ *ist*

$$\begin{aligned} \Psi(u) &= \sup\left\{ \int u\, d\nu - D(\nu\|\mu) : \nu \text{ W'maß auf } M, \int u^+ d\nu < \infty \right\}. \\ &= \sup\left\{ \int u\, d\nu - D(\nu\|\mu) : \nu \text{ W'maß auf } M, \int |u|\, d\nu < \infty, D(\nu\|\mu) < \infty \right\} \end{aligned} \tag{10}$$

*c)* $\Psi : (\mathrm{mb}_b(M), \|.\|_\infty) \to (-\infty, \infty)$ *ist konvex und unterhalb stetig.*

*d) Für jedes W'Maß* $\nu$ *auf* $(M, \mathcal{M})$ *ist*

$$D(\nu\|\mu) = \sup\left\{\int u\, d\nu - \Psi(u) : u \in \mathrm{mb}_b(M)\right\}. \tag{11}$$

*Ist* $\nu \ll \mu$*, so gilt Gleichheit für* $u = \log \frac{d\nu}{d\mu}$ *(auch wenn das nicht beschränkt ist).*

*Beweis:* a) Wir beginnen mit zwei Extremfällen:
- Sei $D(\nu\|\mu) = \infty$. Dann liegt strikte Ungleichheit in (9) vor. Wäre in dieser Situation $\Psi(u) < \infty$ und $\nu = e^{u-\Psi(u)}\mu$, so wäre $D(\nu\|\mu) = \int u - \Psi(u) d\nu \leqslant \int u^+ d\nu - \Psi(u) < \infty$.
- Sei $\Psi(u) = \infty$. Dann kann Gleichheit in (9) nicht auftreten, da die linke Seite $< \infty$ ist.
- Sei also $D(\nu\|\mu) < \infty$, insbesondere $\nu = f\mu$ für eine W'dichte $f$, und $\Psi(u) < \infty$. Dann ist nach Jensen (angewandt auf das Wahrscheinlichkeitsmaß $e^{u-\Psi(u)}\mu$ )

$$\begin{aligned}
\Psi(u) - \int u\, d\nu + D(\nu\|\mu) &= \int (\Psi(u) - u + \log f) \cdot f\, d\mu = \int \log \frac{f}{e^{u-\Psi(u)}} \cdot f\, d\mu \\
&= \int \frac{f}{e^{u-\Psi(u)}} \cdot \log \frac{f}{e^{u-\Psi(u)}} e^{u-\Psi(u)} d\mu = \int \varphi\left(\frac{f}{e^{u-\Psi(u)}}\right) e^{u-\Psi(u)} d\mu \\
&\geqslant \varphi\left(\int \frac{f}{e^{u-\Psi(u)}} \cdot e^{u-\Psi(u)} d\mu\right) = \varphi\left(\int f\, d\mu\right) = 0
\end{aligned}$$

mit Gleichheit genau dann, wenn $\frac{f}{e^{u-\Psi(u)}}$ konstant $e^{u-\Psi(u)}\mu$-f.s. ist. Da $e^{u-\Psi(u)}$ eine strikt positive Wahrscheinlichkeitsdichte ist, ist das äquivalent zu $f = e^{u-\Psi(u)}$ $\mu$-f.s.

Schließlich: Für $\nu = e^{u-\Psi(u)}\mu$ ist $0 \leqslant D(\nu\|\mu) = \int (u - \Psi(u)) d\nu = \int u\, d\nu - \Psi(u)$, also $D(\nu\|\mu) < \infty$ genau dann wenn $\int u^+ d\nu < \infty$.
b) Das erste „$\geqslant$“ in (10) folgt aus Teil a), das zweite ist trivial. Zu zeigen bleibt:

Es gibt W'maße $\nu_r$ auf $M$ mit
$\int |u|\, d\nu_r < \infty$, $D(\nu_r\|\mu) < \infty$ und $\Psi(u) \leqslant \sup_r(\int u\, d\nu_r - D(\nu_r\|\mu))$.

Für $r > 1$ sei

$$u_r(x) = \begin{cases} u(x) & \text{falls } u(x) \leqslant r \\ -\log u(x) & \text{falls } u(x) > r \end{cases}.$$

Dann ist $u_r \leqslant u$, $\lim_{r\to\infty} u_r(x) = u(x)$ für alle $x \in M$ und

$$\Psi(u_r) = \log \int e^{u_r} d\mu = \log\left(\int_{\{u\leqslant r\}} e^u d\mu + \int_{\{u>r\}} \frac{1}{u} d\mu\right) < \infty,$$

da $\int_{\{u>r\}} \frac{1}{u} d\mu \leqslant \frac{1}{r}$, und

$$\lim_{r\to\infty} \Psi(u_r) = \log \int e^u d\mu = \Psi(u)$$

nach dem Satz von der monotonen Konvergenz. Betrachte die W'maße $\nu_r = e^{u_r - \Psi(u_r)}\mu$. Es ist ist$\int |u|\, d\nu_r < \infty$, denn

$$\int u^+ d\nu_r = e^{-\Psi(u_r)}\left(\int_{\{u\leqslant r\}} u^+ e^u d\mu + \int_{\{u>r\}} u\frac{1}{u} d\mu\right) \leqslant e^{-\Psi(u_r)}(re^r + 1) < \infty$$

und

$$\int u^- d\nu_r = \int_{\{u<0\}} (-u) e^{u_r - \Psi(u_r)} d\mu = \int_{\{u<0\}} (-u) e^{-(-u)} d\mu \cdot e^{-\Psi(u_r)} \leqslant \frac{1}{e} e^{-\Psi(u_r)} < \infty.$$

Da $u_r$ nach oben beschränkt ist, ist $D(\nu_r\|\mu) < \infty$, und es folgt aus Teil a), dass $\Psi(u_r) = \int u_r\, d\nu_r - D(\nu_r\|\mu)$ und daher

$$\Psi(u) \leqslant \sup_r \Psi(u_r) = \sup_r \int u_r\, d\nu_r - D(\nu_r\|\mu) \leqslant \sup_r \int u\, d\nu_r - D(\nu_r\|\mu) \leqslant \Psi(u).$$

c) Sei $V$ die Menge aller W'maße $\nu$ auf $(M, \mathcal{M})$ mit $D(\nu\|\mu) < \infty$. Betrachte die stetigen affinen Funktionale $F_\nu : \mathrm{mb}_b(M) \to \mathbb{R}, u \mapsto \int u\, d\nu - D(\nu\|\mu)$. Wegen b) ist $\Psi(u) = \sup_{\nu\in V} F_\nu(u)$ für alle $u \in \mathrm{mb}_b(M)$, so dass $\Psi$ als punktweises Supremum affiner Funktionale konvex und als punktweises Supremum stetiger Funktionale unterhalbstetig ist.
d) „$\geqslant$“ folgt sofort aus (9). Für die Umkehrung betrachten wir zwei Fälle:
- Ist $\nu = f\mu$, so setze $u_{r,s} = \log(f \wedge r) \vee (-s)$ für $r, s > 1$. Dann ist $-s \leqslant u_{r,s} \leqslant \log r$, insbesondere $\int u_{r,s}^+\, d\nu \leqslant \log r < \infty$, und für festes $r$ geht im Limes $s \to \infty$

$$\begin{aligned}\int u_{r,s}\, d\nu - \Psi(u_{r,s}) &= \int \Big(\log(f \wedge r) \vee (-s)\Big) \cdot f\, d\mu - \log \int (f \wedge r) \vee e^{-s} d\mu \\ &\to \int \log(f \wedge r) \cdot f\, d\mu - \log \int f \wedge r\, d\mu\end{aligned}$$

(*Begründung?*). Für $r \to \infty$ geht dieser Ausdruck gegen

$$\int \log f \cdot f\, d\mu - \log \int f\, d\mu = D(\nu\|\mu).$$

- Ist $\nu \not\ll \mu$, so gibt es ein $A \in \mathcal{M}$ mit $\mu(A) = 0$ und $\nu(A) > 0$. Für $u_r = r \cdot 1_A$ geht dann

$$\int u_r\, d\nu - \Psi(u_r) = r \cdot \nu(A) - \log \int e^{r\cdot 1_A}\, d\mu = r \cdot \nu(A) \to \infty = D(\nu\|\mu) \text{ im Limes } r \to \infty.$$

□

Sei weiterhin $(M, \mathcal{M}, \mu)$ ein W'raum, $T : M \to \mathbb{R}^d$ messbar, $\vartheta \in \mathbb{R}^d$.

**Korollar 5.3**

$$\begin{aligned}\psi(\vartheta) &= \log \int e^{\langle \vartheta, T\rangle} d\mu \\ &= \sup\left\{ \int \langle \vartheta, T\rangle\, d\nu - D(\nu\|\mu) : \nu \text{ W'maß auf } (M, \mathcal{M}), \int |\langle \vartheta, T\rangle|\, d\nu < \infty \right\}. \quad (12)\end{aligned}$$

*Weiterhin gilt:*

*a) Für $\vartheta \in \Theta$ wird das Supremum nur durch das W'maß $\nu = f_\vartheta \mu$ realisiert.*

*b) Für $\vartheta \in \Theta$ ist*

$$\psi(\vartheta) = \langle \vartheta, \Gamma(\vartheta)\rangle - D(f_\vartheta\|\mu) = \sup\{\langle \vartheta, \Gamma(\vartheta')\rangle - D(f_{\vartheta'}\|\mu) : \vartheta' \in \Theta\}, \quad (13)$$

*und das Supremum wird nur für $\vartheta = \vartheta'$ angenommen. (Zur Erinnerung: $\Gamma(\vartheta) = E_\vartheta[T]$.)*

*c) Für $\vartheta, \vartheta' \in \Theta$ ist*

$$\langle \vartheta, \Gamma(\vartheta')\rangle \leqslant \psi(\vartheta) + D(f_{\vartheta'}\|\mu)$$

*mit Gleichheit gdw. $\vartheta = \vartheta'$.*

*d) Für $\vartheta \in \Theta$ ist*

$$D(f_\vartheta\|\mu) = \sup\{\langle \vartheta', \Gamma(\vartheta)\rangle - \psi(\vartheta') : \vartheta' \in \Theta\}. \quad (14)$$

*Beweis:* Da $\psi(\vartheta) = \Psi(\langle\vartheta, T\rangle)$ ist (12) ein Spezialfall von (10) mit $u = \langle\vartheta, T\rangle$. Insbesondere ist $\Psi(u) = \psi(\vartheta) < \infty$ für $\vartheta \in \Theta$. Mit Lemma 5.1a folgt daraus Aussage a) dieses Korollars.

Für $\nu = f_{\vartheta'}\mu$ und $u = \langle\vartheta, T\rangle$ folgt aus (9)

$$\langle\vartheta, \Gamma(\vartheta')\rangle - D(f_{\vartheta'}\|\mu) = \langle\vartheta, E'_\vartheta[T]\rangle - D(f_{\vartheta'}\|\mu) = \int \langle\vartheta, T\rangle f_{\vartheta'}\, d\mu - D(f_{\vartheta'}\|\mu) \leqslant \Psi(u) = \psi(\vartheta)$$

mit Gleichheit genau dann, wenn $\vartheta' = \vartheta$, da $\int |u|\, d\nu \leqslant \sum_i |\vartheta_i| \cdot \int |T_i| f_{\vartheta'}\, d\mu < \infty$. Das ist b).
c) folgt aus b).
d) folgt aus c) mit Vertauschung von $\vartheta$ und $\vartheta'$. □

**Satz 5.4** *Sei $M$ ein kompakter metrischer Raum mit Borel-$\sigma$-Algebra $\mathcal{M}$ und $\mu$ ein W'maß auf $\mathcal{M}$. Sei $\Psi : C(M) \to (-\infty, \infty], \Psi(u) = \log \int e^u d\mu$. Dann ist*

$$D(\nu\|\mu) = \sup\left\{\int u\, d\nu - \Psi(u) : u \in C(M)\right\}$$

*für jedes W'maß $\nu$ auf $\mathcal{M}$.*

*Beweis:* Das ist im Wesentlichen (11). Man beachte, dass bei gegebenen W'maßen $\mu$ und $\nu$ und $u \in \mathrm{mb}_b(M)$ sowohl $\int u\, d\nu$ als auch $\Psi(u)$ (gleichzeitig!) durch stetige beschränkte $\tilde{u}$ approximiert werden können [17, Lemma 1.35]:

Sei $u \in \mathrm{mb}_b(M)$, $\varepsilon > 0$, und setze $S := \sup e^u$. Wähle $\tilde{u} \in C(M)$ so, dass

$$\int |u - \tilde{u}|\, d(\nu + \mu) < \frac{\varepsilon^2}{8S} \int e^u d\mu \leqslant \frac{\varepsilon^2}{8} \qquad \text{und} \qquad e^{\tilde{u}} \leqslant 2S.$$

Dann ist $\left|\int u\, d\nu - \int \tilde{u}\, d\nu\right| < \frac{\varepsilon^2}{8} < \varepsilon^2$ und

$$\begin{aligned}
\int e^{\tilde{u}} d\mu &= \int_{\{\tilde{u}-u\leqslant\varepsilon/2\}} e^{\tilde{u}} d\mu + \int_{\{\tilde{u}-u>\varepsilon/2\}} e^{\tilde{u}} d\mu \\
&\leqslant e^{\varepsilon/2} \int e^u d\mu + 2S \cdot \mu\left\{|\tilde{u} - u| > \frac{\varepsilon}{2}\right\} \\
&\leqslant e^{\varepsilon/2} \int e^u d\mu + 2S \cdot \frac{2}{\varepsilon} \cdot \frac{\varepsilon^2}{8S} \cdot \int e^u d\mu \\
&\leqslant \int e^u d\mu \cdot \left(e^{\varepsilon/2} + \frac{\varepsilon}{2}\right),
\end{aligned}$$

woraus durch Logrithmieren folgt

$$\Psi(\tilde{u}) \leqslant \Psi(u) + \log\left(e^{\varepsilon/2}\left(1 + \frac{\varepsilon}{2}\right)\right) \leqslant \Psi(u) + \varepsilon.$$

Also:

$$\int \tilde{u}\, d\nu - \Psi(\tilde{u}) \geqslant \int u\, d\nu - \varepsilon^2 - \Psi(u) - \varepsilon.$$

□

Die Dualitätsbeziehungen in Korollar 5.3 und Satz 5.4 sind Spezialfälle der folgenden Situation:

**Definition 5.5** *Sei $X$ ein topologischer Vektorraum (z.B. ein normierter Vektorraum), $X^*$ sein Dualraum (der Raum aller stetigen linearen Abbildungen von $X \to \mathbb{R}$). Sei $\Lambda : X \to (-\infty, \infty]$ eine konvexe Funktion. Die durch*

$$\Lambda^* : X^* \to (-\infty, \infty],\ \Lambda^*(z) = \sup\{\langle x, z\rangle - \Lambda(x) : x \in X\}$$

*definierte Abbildung heißt* ***Legendre-Fenchel Transformierte*** *von $\Lambda$.*

**Bemerkung 5.6** a) Nach Definition ist $\Lambda^*(z) + \Lambda(x) \geqslant \langle x, z\rangle$ für alle $x \in X$ und $z \in X^*$.

b) Als Supremum affiner Funktionen ist $\Lambda^*$ konvex und unterhalbstetig. (Beachte, dass $X^*$ mit der dualen Topologie selbst wieder ein topologischer Vektorraum ist.)

**Bemerkung 5.7** Im Fall $X = \mathbb{R}^d$, also auch $X^* = \mathbb{R}^d$, und falls $\int e^{\langle \vartheta, T\rangle} d\mu < \infty$ für alle $\vartheta \in X$, so dass $\Theta = X$ ist, kann man Gleichung (14) so formulieren:

$$D(f_\vartheta \| \mu) = \sup\{\langle \vartheta', \Gamma(\vartheta)\rangle - \psi(\vartheta') : \vartheta' \in X\} = \psi^*(\Gamma(\vartheta)) \quad \text{für } \vartheta \in \Theta = X.$$

Interpretiert man nun $\Gamma : \Theta \to \mathbb{R}^d$ als Abbildung von $X$ nach $X^*$ und gilt die lineare Unabhängigkeitsannahme (3), so folgt daraus wegen (13) für alle $z \in \Gamma(\mathring{\Theta}) = \Gamma(X)$:

$$\psi^*(z) = D(f_{\Gamma^{-1}z} \| \mu) = \langle \Gamma^{-1}(z), z\rangle - \psi(\Gamma^{-1}(z)) \ ,$$

und das ist differenzierbar in $z \in \Gamma(\mathring{\Theta})$. Für $z = \Gamma(\vartheta) \in \Gamma(\mathring{\Theta})$ gilt insbesondere $\psi^*(z) + \psi(\vartheta) = \langle \vartheta, z\rangle$ (und nicht nur „$\geqslant$“).

Satz 5.4 kann nun folgendermaßen formuliert werden:

**Satz 5.8** *Sei $M$ ein kompakter metrischer Raum mit Borel-$\sigma$-Algebra $\mathcal{M}$. Sei $X = C(M)$, also $X^*$ der Raum aller endlichen signierten Borel-Maße auf $M$. Sei $\Psi : C(M) \to (-\infty, \infty], \Psi(u) = \log \int e^u d\mu$. Dann ist*

$$D(\nu \| \mu) = \Psi^*(\nu)$$

*für jedes W'maß $\nu \in X^*$.*

**Aufgabe* 5.4** Wir bezeichnen mit $\mathcal{N}_\sigma$ die zentrierte Normalverteilung auf dem $\mathbb{R}^d$ mit Kovarianzmatrix $\sigma^2 E$. Für ein W'maß $Q$ auf dem $\mathbb{R}^d$ ist die *Faltung* $Q * \mathcal{N}_\sigma$ das durch

$$\int u \, d(Q * \mathcal{N}_\sigma) = \int \int u(x + y) \, d\mathcal{N}_\sigma(y) dQ(x)$$

definierte W'maß auf $\mathbb{R}^d$. Ist $u : \mathbb{R}^d \to \mathbb{R}$ messbar und beschränkt, so ist $u * \mathcal{N}_\sigma$ die messbare (sogar beliebig oft differenzierbare) beschränkte Funktion

$$(u * \mathcal{N}_\sigma)(x) = \int u(x + y) \, d\mathcal{N}_\sigma(y).$$

Insbesondere gilt:

$$\int u \, d(Q * \mathcal{N}_\sigma) = \int (u * \mathcal{N}_\sigma) \, dQ.$$

Zeigen Sie: Sind $Q, R$ W'maße auf $\mathbb{R}^d$, so ist

a) $D(R * \mathcal{N}_\sigma \| Q * \mathcal{N}_\sigma) \leqslant D(R\|Q)$ für jedes $\sigma > 0$ und

b) $\lim_{\sigma \to 0} D(R * \mathcal{N}_\sigma \| Q * \mathcal{N}_\sigma) = D(R\|Q)$. *Hinweis:* Hier hilft evtl. ein Korollar zum Satz von Lusin (Korollar VIII.1.19 im Buch *Maß- und Integrationstheorie* von J. Elstrodt).

# 6 Große Abweichungen

Eine sehr schöne Ausarbeitung zum Thema „Große Abweichungen“ ist das Vorlesungsskript [19] von Wolfgang König, dem ich auch einige Anregungen verdanke.

## 6.1 Vorbereitungen

**Definition 6.1** *Sei $\mu$ die Verteilung einer $\mathbb{R}^d$-wertigen ZV $X$. Dann heißt*

$$\psi_X : \mathbb{R}^d \to (-\infty, \infty], \psi_X(\vartheta) = \log \int e^{\langle \vartheta, x\rangle} d\mu(x) = \log E[e^{\langle \vartheta, X\rangle}]$$

*die* logarithmische Laplace-Transformierte *oder die* logarithmische momentenerzeugende Funktion *von $X$. Sei $\Theta = \{\vartheta \in \mathbb{R}^d : \psi_X(\vartheta) < \infty\}$.*

**Aufgabe 6.1** Seien $X^\sigma$, $\sigma > 0$, $\mathbb{R}^d$-wertige normalverteilte zentrierte Zufallsvariablen mit Kovarianzmatrix $\sigma^2 E$, $E = \mathrm{diag}(1, \ldots, 1)$. Zeigen Sie: $\psi_{X^\sigma}(\vartheta) = \frac{\sigma^2}{2}\langle \vartheta, \vartheta\rangle$ und $\psi^*_{X^\sigma}(z) = \frac{1}{2\sigma^2}\langle z, z\rangle$.

Seien $X_1, X_2, \ldots$ ZVn mit Verteilung $\mu$. Notation: $S_n = X_1 + \ldots + X_n$, $\bar{S}_n = n^{-1} S_n$.

**Lemma 6.1** *Sind die $X_i$ u.i.v., so ist $\psi_{\bar{S}_n}(n\vartheta) = \psi_{S_n}(\vartheta) = n\psi_X(\vartheta)$.*

*Beweis:* $\log E[e^{\langle n\vartheta, \bar{S}_n\rangle}] = \log E[e^{\langle \vartheta, S_n\rangle}] = n \log E[e^{\langle \vartheta, X\rangle}]$. □

**Lemma 6.2** *Seien $N \in \mathbb{N}$, und seien $a^i_\varepsilon$, $\varepsilon > 0$, $i = 1, \ldots, N$, nichtnegative Zahlen. Dann gilt*

$$\limsup_{\varepsilon\to 0} \left(\varepsilon \cdot \log \sum_{i=1}^N a^i_\varepsilon\right) = \max_{i=1,\ldots,N} \limsup_{\varepsilon\to 0} \varepsilon \log(a^i_\varepsilon).$$

*Beweis:* [18, Lemma 23.9]. „$\geqslant$“ ist klar. „$\leqslant$“ : Es gibt $\varepsilon_k \searrow 0$ and $i_k \in \{1, \ldots, N\}$ so dass

$$\limsup_{\varepsilon\to 0} \varepsilon \log \sum_{i=1}^N a^i_\varepsilon \;\leqslant\; \limsup_{\varepsilon\to 0} \varepsilon \log(N \cdot \max_{i=1,\ldots,N} a^i_\varepsilon) = \lim_{k\to\infty} \varepsilon_k \log(\max_{i=1,\ldots,N} a^i_{\varepsilon_k}) = \lim_{k\to\infty} \varepsilon_k \log a^{i_k}_{\varepsilon_k}.$$

Es gibt ein $i_0 \in \{1, \ldots, N\}$, so dass $i_{k_j} = i_0$ für unendlich viele Indizes $k_1 < k_2 < k_3 < \ldots$. Daher ist

$$\limsup_{\varepsilon\to 0} \varepsilon \log \sum_{i=1}^N a^i_\varepsilon \leqslant \lim_{j\to\infty} \varepsilon_{k_j} \log a^{i_0}_{\varepsilon_{k_j}} \leqslant \limsup_{\varepsilon\to 0} \varepsilon \log a^{i_0}_\varepsilon \leqslant \max_{i=1,\ldots,N} \limsup_{\varepsilon\to 0} \varepsilon \log a^i_\varepsilon$$

□

## 6.2 Die Grundidee

Seien $X_1, X_2, \ldots$ u.i.v. $\mathbb{R}^d$-wertige ZVn mit Verteilung $\mu$ und bezeichne $\Gamma = D\psi_X$. Sei $z = \Gamma(\vartheta) \in \Gamma(\mathring{\Theta}) \subseteq \mathbb{R}^d$, sei $\delta > 0$ und sei $U_\delta(z)$ die offene $\delta$-Umgebung von $z$. Dann ist mit Lemma 6.1 und Bemerkung 5.7

$$\begin{aligned}
& P\{\bar{S}_n \in U_\delta(z)\} \\
= & \int_{\{\bar{S}_n \in U_\delta(z)\}} e^{\psi_{\bar{S}_n}(n\vartheta) - \langle n\vartheta, z\rangle} \cdot e^{-\langle n\vartheta, \bar{S}_n - z\rangle - \psi_{\bar{S}_n}(n\vartheta) + \langle n\vartheta, \bar{S}_n\rangle} dP \\
= & \exp\left(\psi_{\bar{S}_n}(n\vartheta) - \langle n\vartheta, z\rangle\right) \cdot \int_{\{\bar{S}_n \in U_\delta(z)\}} \exp\left(-\psi_{\bar{S}_n}(n\vartheta) + \langle n\vartheta, \bar{S}_n\rangle \overbrace{\pm(n\delta\|\vartheta\|)}^{\text{zufällig}}\right) dP \\
= & \exp\left(\psi_{\bar{S}_n}(n\vartheta) - \langle n\vartheta, z\rangle \overbrace{\pm(n\delta\|\vartheta\|)}^{\text{nicht zufällig}}\right) \cdot \int_{\{\bar{S}_n \in U_\delta(z)\}} e^{-\psi_{\bar{S}_n}(n\vartheta) + \langle n\vartheta, \bar{S}_n\rangle} dP, \qquad (15)
\end{aligned}$$

wobei $\pm(x)$ hier immer einen Wert aus dem Intervall $[-|x|,|x|]$ bezeichnet. Indem man die Integration bzgl. $P$ durch Integration mit der Verteilung $\mu^{\otimes n}$ des $(\mathbb{R}^d)^n$-wertigen Zufallsvektors $(X_1,\dots,X_n)$ ersetzt und außerdem Lemma 6.1 beachtet, folgt daraus

$$\begin{aligned} & P\{\bar S_n \in U_\delta(z)\} \\ = \;& \exp\left(n(\psi_X(\vartheta) - \langle\vartheta,z\rangle \pm(\delta\|\vartheta\|))\right) \cdot \int_{\{n^{-1}(x_1+\dots+x_n)\in U_\delta(z)\}} \underbrace{e^{-n\psi_X(\vartheta)+\langle\vartheta,x_1+\dots+x_n\rangle} d\mu^{\otimes n}(x_1,\dots x_n)}_{=d(f_\vartheta\mu)^{\otimes n}(x_1,\dots,x_n)} \\ = \;& e^{-n(\psi_X^*(z)\pm(\delta\|\vartheta\|))} P_\vartheta\{\bar S_n \in U_\delta(z)\}. \end{aligned} \tag{16}$$

Da $z = \Gamma(\vartheta) = \int_{\mathbb{R}} x f_\vartheta(x) d\mu(x)$, konvergiert $P_\vartheta\{\bar S_n \in U_\delta(z)\}$ nach dem schwachen Gesetz der großen Zahl gegen 1. Daher ist nach Bemerkung 5.7

$$\liminf_{n\to\infty} \text{ und } \limsup_{n\to\infty} \frac{1}{n}\log P\{\bar S_n \in U_\delta(z)\} \;=\; -\psi_X^*(z) \pm (\delta\|\vartheta\|) = -D(f_\vartheta\|\mu) \pm (\delta\|\vartheta\|). \tag{17}$$

**Bemerkung 6.2** Sowohl die Unabhängigkeitsannahme als auch die Voraussetzung, dass $z = \Gamma(\vartheta) \in \Gamma(\mathring{\Theta})$ sein soll, sind nicht immer erfüllt. Ohne diese Annahmen gilt aber immer noch (15), so dass (jetzt für *jedes* $\vartheta \in \mathbb{R}^d$)

$$\begin{aligned} P\{\bar S_n \in U_\delta(z)\} &= e^{\psi_{\bar S_n}(n\vartheta)-\langle n\vartheta,z\rangle\pm(n\delta\|\vartheta\|)} \int_{\{\bar S_n\in U_\delta(z)\}} \underbrace{e^{-\psi_{\bar S_n}(n\vartheta)+\langle n\vartheta,\bar S_n\rangle} dP}_{\text{W'maß}} \\ &\leqslant e^{\psi_{S_n}(\vartheta)-\langle n\vartheta,z\rangle+n\delta\|\vartheta\|}. \end{aligned} \tag{18}$$

**Für den Rest dieser Bemerkung setzen wir voraus**, dass

$$\psi(\vartheta) := \lim_{n\to\infty} \frac{1}{n}\psi_{S_n}(\vartheta) \in [-\infty,+\infty] \quad \text{für alle } \vartheta\in\mathbb{R}^d \text{ existiert.} \tag{19}$$

Wir erinnern daran, dass $\psi^*(z) = \sup_{\vartheta\in\mathbb{R}^d}\langle\vartheta,z\rangle - \psi(\vartheta)$, und definieren für $r>0$ (nahe bei 0)

$$\psi_r^* := \min\{\psi^* - r, r^{-1}\}.$$

Dann gibt zu jedem $r>0$ und $z\in\mathbb{R}^d$ ein $\vartheta_{z,r}\in\mathbb{R}^d$ mit $\langle\vartheta_{z,r},z\rangle - \psi(\vartheta_{z,r}) > \psi_r^*(z)$, und es folgt, dass für jedes $\delta>0$

$$\begin{aligned} \limsup_{n\to\infty}\frac{1}{n}\log P\{\bar S_n\in U_\delta(z)\} &\leqslant \psi(\vartheta_{z,r}) - \langle\vartheta_{z,r},z\rangle + \delta\|\vartheta_{z,r}\| \\ &< -\psi_r^*(z) + \delta\|\vartheta_{z,r}\| \end{aligned} \tag{20}$$

Damit ist die Bemerkung beendet.

Es folgt ein erster, noch etwas eingeschränkter Satz über große Abweichungen:

**Satz 6.3** *Seien $X_1, X_2, \dots$ $\mathbb{R}^d$-wertige ZVn, $\psi$ wie in (19) definiert.*

*a) Sind die $X_i$ u.i.v, also $\psi = \psi_X$ nach Lemma 6.1, und ist $G\subseteq\mathbb{R}^d$ offen, so ist*

$$\liminf_{n\to\infty}\frac{1}{n}\log P\{\bar S_n\in G\} \geqslant -\inf_{z\in G\cap\Gamma(\mathring{\Theta})}\psi^*(z).$$

*Dabei ist $\inf_{z\in\emptyset}\psi^*(z) = \infty$ und $\psi^*(z) = D(f_{\Gamma^{-1}(z)}\|\mu)$ für $z\in\Gamma(\mathring{\Theta})$.*

*b) Ist $K\subseteq\mathbb{R}^d$ kompakt, so ist*

$$\limsup_{n\to\infty}\frac{1}{n}\log P\{\bar S_n\in K\} \leqslant -\inf_{z\in K}\psi^*(z).$$

*Beweis:* a) Sei $z = \Gamma(\vartheta) \in G$, $\vartheta \in \mathring{\Theta}$. Dann ist $U_\delta(z) \subseteq G$ für jedes hinreichend kleine $\delta > 0$, so dass die Abschätzung aus (17) folgt.
b) Sei $r > 0$ wie in Bemerkung 6.2. Zu $z \in K$ sei $\delta_z := r \cdot \|\vartheta_{z,r}\|^{-1}$. Da $K$ kompakt ist, gibt es $z_1, \ldots, z_N \in K$, so dass $K \subseteq \bigcup_{i=1}^N U_{\delta_{z_i}}(z_i)$. Dann folgt aus (20) und Lemma 6.2, dass

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log P\{\bar{S}_n \in K\} &\leqslant \limsup_{n\to\infty} \frac{1}{n} \log \sum_{i=1}^N P\{\bar{S}_n \in U_{\delta_{z_i}}(z_i)\} \\
&\leqslant \max_{i=1,\ldots,N} \limsup_{n\to\infty} \frac{1}{n} \log P\{\bar{S}_n \in U_{\delta_{z_i}}(z_i)\} \\
&\leqslant \max_{i=1,\ldots,N} (-\psi_r^*(z_i)) + \delta_{z_i} \|\vartheta_{z_i,r}\| \\
&\leqslant -\inf_{z\in K} \psi_r^*(z) + r \\
&= -\inf_{z\in K} \min\{\psi^*(z) - r, r^{-1}\} + r \\
&= -\min\{\inf_{z\in K} \psi^*(z) - r, r^{-1}\} + r,
\end{aligned}$$

da Minimum und Infimum vertauschen. Im Limes $r \to 0$ folgt b). □

Ist $\Gamma(\Theta) = \mathbb{R}^d$, so ist dieser Satz gerade ein *schwaches Prinzip der großen Abweichungen* (*weak large deviations principle, weak LDP).* Für ein *starkes LDP* würde man fordern, dass Aussage b) für beliebige abgeschlossene $K \subseteq \mathbb{R}^d$ gilt. Solche ,,vollständigen" Sätze lernen wir in den nächsten Kapiteln kennen.

**Aufgabe* 6.2** Seien $X^\sigma$, $\sigma > 0$, $\mathbb{R}^d$-wertige normalverteilte zentrierte Zufallsvariablen mit Kovarianzmatrix $\sigma^2 E$, $E = \mathrm{diag}(1, \ldots, 1)$.

a) Zeigen Sie: $\liminf_{\sigma\to 0} \sigma^2 \cdot \log P\{X^\sigma \in G\} \geqslant -\inf_{z\in G} \frac{1}{2}\langle z, z\rangle$ für jede offene Teilmenge $G \subseteq \mathbb{R}^d$.

b) Zeigen Sie: $\limsup_{\sigma\to 0} \sigma^2 \cdot \log P\{X^\sigma \in K\} \leqslant -\inf_{z\in K} \frac{1}{2}\langle z, z\rangle$ für jede kompakte Teilmenge $K \subseteq \mathbb{R}^d$.

*Hinweis:* Man kann das nach dem Muster von Gleichung (16) und Satz 6.3 beweisen.

**Aufgabe* 6.3** Zeigen Sie in der Situation von Problem 6.2:

a) Für $r > 0$ ist

$$\lim_{\sigma\to 0} \sigma^2 \cdot \log P\{\|X^\sigma\|_2 \geqslant r\} = -\frac{r^2}{2}.$$

b) Zeigen Sie: $\limsup_{\sigma\to 0} \sigma^2 \cdot \log P\{X^\sigma \in A\} \leqslant -\inf_{z\in A} \frac{1}{2}\langle z, z\rangle$ gilt für jede abgeschlossene Teilmenge $A \subseteq \mathbb{R}^d$. *Hinweis:* Benutzen Sie auch das Ergebnis von Aufgabe 6.2.

## 6.3 Das LDP (Large Deviations Principle, Prinzip der großen Abweichungen)

Sei $M$ ein *polnischer Raum*, d.h. ein separabler topologischer Raum, dessen Topologie von einer vollständigen Metrik $d$ erzeugt wird. $\mathcal{M}$ sei die zugehörige Borelsche $\sigma$-Algebra. Jedes endliche Borel-Maß auf einem polnischen Raum ist regulär, d.h. jede Borelmenge kann bzgl. des Maßes von innen durch kompakte und von außen durch offene Mengen approximiert werden, siehe z.B. [18, Satz 13.6].

Motiviert durch Satz 6.3 treffen wir folgende Definitionen, die auf Varadhan [23] zurückgehen:

**Definition 6.4** *Eine unterhalbstetige Funktion $I : M \to [0, +\infty]$ heißt* Ratenfunktion. *Sie heißt* gut, *falls die Niveaumengen $\{x \in M : I(x) \leqslant s\}$ für alle $s < \infty$ kompakt (und nicht nur abgeschlossen) sind. Ist $A \subseteq M$, so bezeichnet* $\inf_A I := \inf\{I(x) : x \in A\}$.

**Definition 6.5** *Eine Folge von W'verteilungen $(\mu_n)_{n\geqslant 1}$ erfüllt ein* schwaches LDP *mit Ratenfunktion $I$, falls*

*a) für alle offenen* $G \subseteq M$ $\quad \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geqslant -\inf_G I$ *und*

*b) für alle kompakten* $K \subseteq M$ $\quad \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(K) \leqslant -\inf_K I$.

*Gilt b) sogar für alle abgeschlossenen* $K \subseteq M$*, so spricht man von einem* vollen LDP *oder einfach von einem* LDP.

*Man sagt auch, eine Folge* $(Z_n)_n$ *von ZVn erfüllt ein LDP, wenn die Folge ihrer Verteilungen ein LDP erfüllt.*

**Bemerkung 6.6** a) Verallgemeinerungen dieser Definition findet man z.B. in [18, Def. 23.7] oder [9, Def. 2.2].
b) Die Ratenfunktion in Abschnitt 6.2 ist $I = \psi^*$.

**Bemerkung 6.7** Äquivalent zum vollen LDP ist:

- Für alle messbaren $A \subseteq M$ ist

$$-\inf_{\mathring{A}} I \leqslant \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(A) \leqslant \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(A) \leqslant -\inf_{\bar{A}} I.$$

Ist die Ratenfunktion auf $\bar{A}$ stetig und ist $\overline{\mathring{A}} = \bar{A}$, so liegt Konvergenz vor.

**Aufgabe 6.4** Beweisen Sie diese Bemerkung.

Eine äquivalente Formulierung des LDP geht auf Freidlin und Wentzell [15] zurück:

**Satz 6.8** *Sei* $d$ *eine Metrik für den polnischen Raum* $M$*. Eine Folge von W'verteilungen* $(\mu_n)_{n\geqslant 1}$ *erfüllt ein* volles LDP *mit* guter *Ratenfunktion* $I$ *genau dann, wenn*

*A. für jedes* $x \in M$ *und jedes* $\varepsilon > 0$ $\quad \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(U_\varepsilon(x)) \geqslant -I(x)$ $\quad$ *und*

*B. für jedes* $s < \infty$ *und jedes* $\varepsilon > 0$ $\quad \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(M \setminus U_\varepsilon(\{I \leqslant s\})) \leqslant -s$.

*Genauer gilt: A)⇔a) und B)⇔b).*

*Beweis:* a) ⇒A): Für jedes $x \in M$ und $\varepsilon > 0$ gilt: $\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(U_\varepsilon(x)) \geqslant -\inf_{U_\varepsilon(x)} I \geqslant -I(x)$.
A) ⇒ a): Für jedes $x \in G$ gibt es ein $\varepsilon(x) > 0$ mit $U_{\varepsilon(x)}(x) \subseteq G$. Also: $\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geqslant \sup_{x\in G} \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(U_{\varepsilon(x)}(x)) \geqslant \sup_{x\in G}(-I(x)) = -\inf_G I$.
b) ⇒B): $F := M \setminus U_\varepsilon(\{I \leqslant s\})$ ist abgeschlossen und $\inf_F I \geqslant \inf_{\{I>s\}} I \geqslant s$.
B)⇒b): Sei $A \subseteq M$ abgeschlossen und sei $s_0 := \inf_A I \geqslant 0$. Ist $s_0 = 0$, so ist nichts zu zeigen. Sonst sei $s \in (0, s_0)$. Dann sind die kompakte Menge $\{I \leqslant s\}$ und die abgeschlossene Menge $A$ disjunkt, so dass $d(\{I \leqslant s\}, A) > 0$ und damit auch $d(U_\varepsilon(\{I \leqslant s\}), A) > 0$ für hinreichend kleine $\varepsilon > 0$, also $A \subseteq M \setminus U_\varepsilon(\{I \leqslant s\})$. Also: $\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(A) \leqslant -s$ und da $s < s_0 = \inf_A I$ beliebig war, folgt b). □

**Definition 6.9** *Eine Folge von W'verteilungen* $(\mu_n)_{n\geqslant 1}$ *heißt* exponentiell straff*, wenn es zu jedem* $s > 0$ *ein kompaktes* $K \subseteq M$ *gibt, so dass*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(M \setminus K) \leqslant -s.$$

**Satz 6.10** *Erfüllt eine exponentiell straffe Folge von W'verteilungen* $(\mu_n)_{n\geqslant 1}$ *ein schwaches LDP, so erfüllt sie auch ein volles LDP.*

*Beweis:* Sei $A \subseteq M$ abgeschlossen, $s > 0$ und $K$ kompakt wie in Definition 6.9. Dann ist

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(A) &\leqslant \limsup_{n\to\infty} \frac{1}{n} \log(\mu_n(A \cap K) + \mu_n(M \setminus K)) \\
&= \max\left\{ \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(A \cap K), \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(M \setminus K) \right\} \\
&\leqslant \max\{-\inf_{A\cap K} I, -s\} \leqslant \max\{-\inf_{A} I, -s\},
\end{aligned}$$

und da das für jedes $s > 0$ gilt, ist der Satz bewiesen.

□

**Satz 6.11** *Seien $X_1, X_2, \ldots$ u.i.v. $\mathbb{R}^d$-wertige ZVn, für die gilt $0 \in \mathring{\Theta}$. Dann ist die Folge der Verteilungen der $\bar{S}_n$ exponentiell straff.*

**Aufgabe* 6.5** Beweisen Sie Satz 6.11. Versuchen Sie es zunächst für $d = 1$.

# 7 Der Satz von Cramer und Anwendungen

## 7.1 Der Satz von Cramer in $\mathbb{R}$

(Fast) ohne einschränkende Annahmen an das Intervall $\Gamma(\Theta) \subseteq \mathbb{R}$ haben wir den folgenden Satz, der die Situation von Bemerkung 6.7 (d.h. $\bar{A} = \bar{\mathring{A}}$) mit $A = [x, \infty)$ illustriert.

**Satz 7.1** ***(Cramer)*** *Seien $X_1, X_2, \ldots$ u.i.v. $\mathbb{R}$-wertige ZVn mit Verteilung $\mu$ und $\gamma := EX_i \in (-\infty, \infty)$. Sei $0 \in \mathring{\Theta}$. Bezeichne $\bar{S}_n := \frac{1}{n}(X_1 + \ldots + X_n)$.*

*a) Für jedes $x \geqslant \gamma$ mit $x \in \Gamma(\mathring{\Theta})$ gilt*

$$\lim_{n\to\infty} \frac{1}{n} \log P\left\{\bar{S}_n \geqslant x\right\} = -\psi^*(x) = -D(f_{\vartheta_x} \| \mu) \ , \ \textit{wobei } \vartheta_x = \Gamma^{-1}(x) \ . \tag{21}$$

*b) Ist $[0, \infty) \subseteq \Theta$, so gilt das auch für $x \geqslant \gamma$, die nicht in $\Gamma(\mathring{\Theta})$ liegen.*

*c) Ist $x \leqslant \gamma$, so gelten a) und b) entsprechend für $P\left\{\bar{S}_n \leqslant x\right\}$.*

**Bemerkung 7.2** Den Bezug zu einem allgemeinen LDP sieht man folgendermaßen: Sei $x \geq \gamma$ wie im Satz. Die Menge $A = [x, \infty)$ ist abgeschlossen, und es ist $\inf_A \psi^* = \psi^*(x)$., denn nach Definition ist $\psi^*(x) = \sup_{\vartheta\in\mathbb{R}}(\vartheta x - \psi(\vartheta)) \geq 0\, x - \psi(0) = 0$ und für $x = \gamma$ ist $\psi^*(\gamma) = D(f_{\Gamma^{-1}(\gamma)} \| \mu) = D(f_0 \| \mu) = 0$, so dass aus der Konvexität von $\psi^*$ folgt, dass $\psi^*$ monoton wachsend auf $[\gamma, \infty)$ ist.

*Beweis:* Für jedes $x \geqslant \gamma$ und jedes $\vartheta \geqslant 0$ ist

$$\begin{aligned} P\{\bar{S}_n \geqslant x\} &\leqslant P\{\vartheta S_n \geqslant n\vartheta x\} = P\{e^{\vartheta S_n} \geqslant e^{n\vartheta x}\} \\ &\leqslant e^{-n\vartheta x} \int e^{\vartheta S_n} dP = e^{-n\vartheta x + \psi_{S_n}(\vartheta)} = e^{n(\psi(\vartheta) - \vartheta x)}. \end{aligned}$$

(Beachte, dass das auch für $\psi(\vartheta) = \infty$ richtig ist.) Es folgt, dass

$$\frac{1}{n} \log P\{\bar{S}_n \geqslant x\} \leqslant \inf_{\vartheta \geqslant 0} (\psi(\vartheta) - x\vartheta) = -\sup_{\vartheta \geqslant 0} (x\vartheta - \psi(\vartheta)) \ .$$

Da $\psi(0) = 0$ und da $\psi$ konvex ist, ist $\psi(\vartheta) \geqslant \vartheta \cdot D\psi(0) = \vartheta \cdot \int x \, d\mu(x) = \vartheta\gamma$ für alle $\vartheta \in \mathbb{R}$, so dass wegen $x \geqslant \gamma$ für $\vartheta < 0$ gilt: $x\vartheta - \psi(\vartheta) \leqslant \gamma\vartheta - \vartheta\gamma = 0$. Also kann man das Supremum auf alle $\vartheta \in \mathbb{R}$ ausdehnen, ohne es zu verändern. Man erhält so $\frac{1}{n} \log P\{\bar{S}_n \geqslant x\} \leqslant -\psi^*(x)$.
a) Für die untere Abschätzung sei nun $x \in \Gamma(\mathring{\Theta})$. Sei $\delta > 0$ so, dass auch $z := x + \delta \in \Gamma(\mathring{\Theta})$. Für jedes $r \in (0, \delta)$ ist dann $U_r(z) \subseteq [x, \infty)$. Aus (17) folgt

$$\liminf_{n\to\infty} \frac{1}{n} \log P\{\bar{S}_n \geqslant x\} \quge \liminf_{n\to\infty} \frac{1}{n} \log P\{\bar{S}_n \in U_r(z)\} \geqslant -\psi^*(z) \pm (r|\vartheta_z|).$$

Im Limes $r \to 0$ verschwindet der Fehlerterm. Im Limes $\delta \to 0$ geht dann $z \to x$, und da $z \mapsto \Gamma^{-1}(z)$ und $z \mapsto \psi^*(z) = \langle \Gamma^{-1}(z), z\rangle - \psi(\Gamma^{-1}(z))$ für $z \in \Gamma(\mathring{\Theta})$ stetig (sogar differenzierbar) sind, gilt $\lim_{z\to x} \psi^*(z) = \psi^*(x)$. Daraus folgt die gesuchte Abschätung.
b) Ist $[0, \infty) \subseteq \Theta$, so lässt sich die untere Abschätzung auch für $x \notin \Gamma(\mathring{\Theta})$ zeigen. Zunächst beachtet man, dass das Intervall $\Gamma(\mathring{\Theta}) = (\alpha, \beta)$ die rechte Grenze $\beta = \mu\text{-ess sup } X$ hat (siehe Korollar 3.5iii). Da $x \geqslant \gamma = E[X_i] = \Gamma(0) \in \Gamma(\Theta)$, ist also $x \geqslant \beta$ und daher $\beta < \infty$.

- Ist $x > \beta$, so ist $P\{\bar{S}_n \geqslant x\} = 0$ für alle $n$, und es ist $\psi^*(x) = \sup_{\vartheta\in\mathbb{R}}(\vartheta x - \psi(\vartheta)) = \infty$, da $\psi'(\vartheta) = \Gamma(\vartheta) \leqslant \beta < x$ für alle $\vartheta$. Also sind beide Seiten von (21) gleich $-\infty$.

– Ist $x = \beta$, so ist $P\{\bar{S}_n \geqslant x\} = P\{X_1 = \ldots = X_n = \beta\} = \mu\{\beta\}^n$, so dass die linke Seite von (21) gleich $\log \mu\{\beta\}$ ist. Zu zeigen bleibt $\log \mu\{\beta\} \geqslant -\psi^*(x)$: Für jedes $\delta > 0$ und $\vartheta > 0$ ist

$$\begin{aligned}\psi(\vartheta) &= \log\left(\int_{(-\infty,\beta-\delta]} e^{\vartheta t} d\mu(t) + \int_{(\beta-\delta,\beta]} e^{\vartheta t} d\mu(t)\right) \\ &\leqslant \log(e^{\vartheta(\beta-\delta)} + \mu((\beta-\delta,\beta])e^{\vartheta\beta}),\end{aligned}$$

woraus folgt

$$\begin{aligned}\psi^*(x) = \sup_{\vartheta\in\mathbb{R}} \vartheta x - \psi(\vartheta) &\geqslant \liminf_{\vartheta\to\infty} \vartheta x - \psi(\vartheta) \\ &\geqslant \liminf_{\vartheta\to\infty} [\vartheta\beta - \log(e^{\vartheta(\beta-\delta)} + \mu((\beta-\delta,\beta])e^{\vartheta\beta})] \\ &= -\limsup_{\vartheta\to\infty} \log(e^{-\vartheta\delta} + \mu((\beta-\delta,\beta])) \\ &= -\log\mu([\beta-\delta,\beta]).\end{aligned}$$

Im Limes $\delta \to 0$ folgt $\psi^*(x) \geqslant -\log\mu\{\beta\}$.

c) Bleibt der Fall $x \leqslant \gamma$ zu betrachten. Dann ist

$$P\left\{\frac{1}{n}(X_1 + \ldots + X_n) \leqslant x\right\} = P\left\{\frac{1}{n}(-X_1 - \ldots - X_n) \geqslant -x\right\}$$

und $-x > -\gamma = E[-X]$. Da $\psi_{-X}(\vartheta) = \psi_X(-\vartheta)$, ist

$$\psi^*_{-X}(-x) = \sup_{\vartheta\in\mathbb{R}}(-\vartheta x - \psi_{-X}(\vartheta)) = \sup_{\vartheta\in\mathbb{R}}(-\vartheta x - \psi_X(-\vartheta)) = \psi^*_X(x),$$

und die Abschätzung folgt aus dem schon bewiesenen Fall. □

**Aufgabe* 7.1** Seien $X_1, X_2, \ldots$ u.i.v. $\mathbb{R}$-wertige ZVn mit Verteilung $\mu$. Es gelte $\Theta = \mathbb{R}$ und $\Gamma(\Theta) = \mathbb{R}$. Benutzen Sie die Sätze 6.3, 6.10 und 6.11, um ein LDP für die Folge $(\bar{S}_n)_{n>0}$ zu zeigen. Leiten Sie daraus den Satz von Cramer (7.1) her.

## 7.2 Anwendung: Neyman-Pearson Tests

Seien $X_1, \ldots, X_n$ u.i.v. reellwertige ZVn. (Oft nehmen sie nur die Werte 1 und 0 für „Erfolg“ bzw. „Misserfolg“ an.) Es sei bekannt, dass die $X_i$ entweder die Verteilung $\mu_0$ oder die Verteilung $\mu_1$ haben. Auf Basis der $n$ Beobachtungen ist nun die Hypothese, dass es $\mu_0$ ist, gegen die Alternative, dass es $\mu_1$ ist, zu testen.

Wir nehmen an, dass $\mu_0 \approx \mu_1$ und betrachten den log-Likelihood Quotienten $h(x) := \log \frac{d\mu_1}{d\mu_0}(x)$. Sei $Y_i := h(X_i)$. Ein Neymann-Pearson Test für das obige Testproblem ist von der Form

$$T_n(X_1, \ldots, X_n) := 1_{\{Y_1+\ldots+Y_n \geq nc\}}$$

für ein $c \in \mathbb{R}$. Ist $T_n = 1$, so wird die Hypothese abgelehnt. Seien

$$\alpha_n := P_0(T_n = 1) \quad \text{und} \quad \beta_n := P_1(T_n = 0)$$

die Wahrscheinlichkeiten, dass die Hypothese abgelehnt wird, obwohl $\mu_0$ vorliegt (*Fehler 1. Art*), bzw. dass die Hypothese akzeptiert wird, obwohl $\mu_1$ vorliegt (*Fehler 2. Art*). Aus der Testtheorie weiß man, dass Neyman-Pearson Tests im folgenden Sinne optimal sind: Kein anderer Test kann bei gleichem $\alpha_n$ ein kleineres $\beta_n$ haben und umgekehrt.

Sei nun $\psi_0(\vartheta) := \log E_{P_0}[e^{\vartheta Y}] = \log \int e^{\vartheta h(x)} d\mu_0(x)$, insbesondere $\psi_0(0) = \psi_0(1) = 0$, so dass $[0, 1] \subseteq \Theta$.

**Satz 7.3** *Sei sogar* $[0,1] \subseteq \mathring{\Theta}$. *Bezeichne* $x_0 = -H(\mu_0\|\mu_1)$, $x_1 = H(\mu_1\|\mu_0)$, *und sei* $c \in (x_0, x_1)$. *Dann ist*

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n = -\psi_0^*(c) < 0 \quad \text{und} \quad \lim_{n\to\infty} \frac{1}{n} \log \beta_n = c - \psi_0^*(c) < 0.$$

*Beweis:* Es ist

$$x_0 = -\int \frac{d\mu_0}{d\mu_1} \log \frac{d\mu_0}{d\mu_1} d\mu_1 = \int \log \frac{d\mu_1}{d\mu_0} d\mu_0 = \int h d\mu_0 = E_{P_0}[Y] = \psi_0'(0) = \Gamma(0)$$

und

$$x_1 = \int \frac{d\mu_1}{d\mu_0} \log \frac{d\mu_1}{d\mu_0} d\mu_0 = \int \log \frac{d\mu_1}{d\mu_0} d\mu_1 = \int h d\mu_1 = E_{P_1}[Y] = \psi_0'(1) = \Gamma(1).$$

Insbesondere ist $c \in (x_0, x_1) \subseteq \Gamma(\mathring{\Theta})$. Da also $c > E_{P_0}[Y]$ ist, folgt aus dem Satz von Cramer:

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n = \lim_{n\to\infty} \frac{1}{n} \log P_0\{Y_1 + \ldots + Y_n \geq nc\} = -\psi_0^*(c).$$

Weiterhin ist $\psi_1(\vartheta) := \log E_{P_1}[e^{\vartheta Y}] = \log \int e^{\vartheta h(x)} d\mu_1(x) = \log \int e^{\vartheta h(x) + h(x)} d\mu_0(x) = \psi_0(1+\vartheta)$, also

$$\psi_1^*(x) = \sup_\vartheta(\vartheta x - \psi_0(1+\vartheta)) = \sup_\vartheta((1+\vartheta)x - \psi_0(1+\vartheta)) - x = \psi_0^*(x) - x.$$

Da $c < x_1 = E_1[Y]$, folgt aus der letzten Aussage des Satzes von Cramer

$$\lim_{n-\infty} \frac{1}{n} \log \beta_n \quad = \quad \lim_{n\to\infty} \frac{1}{n} \log P_1\{Y_1 + \ldots + Y_n < nc\} = -\psi_1^*(c) = c - \psi_0^*(c).$$

□

**Aufgabe 7.2** Zeigen Sie, dass tatsächlich beide Limiten echt negativ sind.

**Aufgabe* 7.3** Sei $\alpha \in (0,1)$. Für jedes $n \in \mathbb{N}$ sei $c_n \in \mathbb{R}$ so gewählt, dass

$$\alpha_n(c_n) := P_0(Y_1 + \cdots + Y_n \geqslant c_n \cdot n) = \alpha.$$

(Das geht immer, wenn die Verteilung $\mu_0$ keine Punktmassen hat.) Zeigen Sie, dass dann für die entsprechenden $\beta_n(c_n)$ gilt:

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n(c_n) = x_0 \ .$$

## 7.3 Anwendung: Die stationäre Verteilung von Warteschlangen

Sei $(U_n, V_n)_{n>0}$ eine Folge unabhängiger $\mathbb{R}^2$-wertiger Zufallsvariablen, $U_n, V_n \geqslant 0$. Interpretiert man $U_n$ als den Betrag, um den sich eine Warteschlange in der $n$-ten Zeitperiode durch Ankunft neuer „Kunden" aufbaut, und $V_n$ als den Betrag, um den sie in derselben Zeit durch Bearbeitung abgebaut werden kann, so verändert sich die Länge $L_n$ der Warteschlange um $X_n := U_n - V_n$, also $L_n = (L_{n-1} + X_n)^+$, denn die Länge kann natürlich nicht negativ werden. Insbesondere ist

$$L_n \geqslant L_{n-1} + X_n \geqslant \ldots \geqslant L_{m-1} + \sum_{i=m}^{n} X_i \geqslant \ldots \geqslant \sum_{i=1}^{n} X_i. \tag{22}$$

Wir betrachten den Fall, wo $E[X_i] < 0$, d.h. wo der erwartete Zuwachs kleiner als der erwartete Abbau ist. Wir nehmen zusätzlich an, dass $P[X_i > 0] > 0$, dass es also durchaus passieren kann, dass in einer Periode mehr Kunden ankommen als bedient werden können.

Ist $k-1 = k(n)-1$ der letzte Zeitpunkt vor und einschließlich $n$, zu dem $L_{k-1} = 0$, so ist $L_n = \sum_{i=k(n)}^{n} X_i$ und $\sum_{i=k(n)}^{m-1} X_i > 0$ für alle $m \in \{k(n)+1, \ldots, n\}$.

– Für $m > k(n)$ ist daher $\sum_{i=m}^{n} X_i = L_n - \sum_{i=k(n)}^{m-1} X_i < L_n$.

– Für $m < k(n)$ ist wegen (22):

$$\sum_{i=m}^{n} X_i = \sum_{i=m}^{k(n)-1} X_i + L_n \leqslant L_{k(n)-1} - L_{m-1} + L_n = -L_{m-1} + L_n \leqslant L_n.$$

Es folgt, dass

$$L_n = \max_{m=1,\ldots n} \sum_{i=m}^{n} X_i. \tag{23}$$

Sei nun $\psi(\vartheta) = \log E[e^{\vartheta X_i}]$. Der Einfachheit halber nehmen wir an, dass $[0,\infty) \subseteq \mathring{\Theta}$ ist. Wir wissen, dass $\psi$ konvex ist und dass $\psi(0) = 0$ und $\psi'(0) = E[X_i] < 0$. Außerdem gibt es ein $\vartheta_* > 0$, für das $\psi(\vartheta_*) = 0$ ist. Das folgt aus obiger Annahme $P[X_i > 0] > 0$, denn dann existiert ein $\delta > 0$ mit $P[X_i \geqslant \delta] > 0$, so dass für $\vartheta > 0$ gilt:

$$\psi(\vartheta) \geqslant \log \int_{\{X_i \geqslant \delta\}} e^{\vartheta X_i} dP \geqslant \log(e^{\vartheta\delta} P[X_i \geqslant \delta]) = \vartheta\delta + \log P[X_i \geqslant \delta] \to \infty \text{ mit } \vartheta \to \infty.$$

**Satz 7.4** *Sei* $[0,\infty) \subseteq \mathring{\Theta}$.

$$\lim_{z\to\infty} \frac{1}{z} \log \sup_{n>0} P[L_n > z] = \lim_{z\to\infty} \frac{1}{z} \log P[L_{\lceil z/\alpha \rceil} > z] = -\vartheta_* \quad (\text{wo } \alpha = \psi'(\vartheta_*) \text{ ist}).$$

Anmerkung: *Man kann zeigen, dass die $L_n$ inVerteilung gegen eine Zufallsvariable $L_\infty$ konvergieren. Man kannn sogar Zufallsvariablen $\tilde{L}_n \leqslant L_\infty$ finden, die genau so verteilt sind wie die $L_n$, und fast sicher gegen $L_\infty$ konvergieren. Damit lässt sich die Aussage diesess Satzes schreiben als*

$$\lim_{z\to\infty} \frac{1}{z} \log P[L_\infty > z] = -\vartheta_* \ .$$

*Beweis:* „$\leqslant$“: Sei $s \in (0, \vartheta_*)$, so dass $\psi(s) < 0$. Dann ist für jedes $z > 0$ und $n > 0$:

$$\begin{aligned}
P[L_n > z] &= P\left[\exists m \in \{1,\ldots,n\} : \sum_{i=m}^{n} X_i > z\right] \\
&\leqslant \sum_{m=1}^{n} P\left[\sum_{i=m}^{n} X_i > z\right] \\
&= \sum_{m=1}^{n} P\left[e^{s\sum_{i=m}^{n} X_i} > e^{sz}\right] \\
&\leqslant \sum_{m=1}^{n} e^{-sz} \prod_{i=m}^{n} E[e^{sX_i}] \\
&= e^{-sz} \sum_{m=1}^{n} e^{(n-m+1)\psi(s)} \\
&\leqslant e^{-sz} \frac{e^{\psi(s)}}{1 - e^{\psi(s)}},
\end{aligned}$$

da $\psi(s) < 0$. Es folgt:

$$\limsup_{z\to\infty} \frac{1}{z} \sup_{n>0} \log P[L_n > z] \leqslant -s,$$

und da $s \in (0, \vartheta_*)$ beliebig ist, gilt die gleiche Abschätzung auch mit $\vartheta_*$ statt $s$.
„$\geqslant$“: Sei $\alpha = \psi'(\vartheta_*) > 0 > \psi'(0) = E[X_i]$. Beachte, dass $L_n \geqslant \sum_{i=1}^n X_i$. Dann ist

$$\begin{aligned}
\liminf_{z\to\infty} \frac{1}{z} \sup_{n>0} \log P[L_n > z] &\geqslant \frac{1}{\alpha} \liminf_{z\to\infty} \frac{1}{\lceil z/\alpha \rceil} \log P[L_{\lceil z/\alpha \rceil} > \alpha \lceil z/\alpha \rceil] \\
&= \frac{1}{\alpha} \liminf_{n\to\infty} \frac{1}{n} \log P[L_n > \alpha n] \\
&\geqslant \frac{1}{\alpha} \liminf_{n\to\infty} \frac{1}{n} \log P\left[\sum_{i=1}^n X_i > \alpha n\right] \\
&= -\frac{1}{\alpha} \psi^*(\alpha)
\end{aligned}$$

wegen des Satzes von Cramer, und aus Bemerkung 5.7 folgt $\psi^*(\alpha) = \vartheta_* \cdot \alpha - \psi(\vartheta_*) = \vartheta_* \cdot \alpha$, da $\alpha = \Gamma(\vartheta_*)$ □

**Beispiel 7.5** Nimmt man an, dass die $U_n$ und $V_n$ voneinander unabhängig sind, so ist $\psi(\vartheta) = \psi_U(\vartheta) + \psi_V(-\vartheta)$. Ist die Zahl der ankommenden Kunden pro Zeitschritt Poisson($\lambda$)-verteilt und wird immer ein Kunde pro Zeiteinheit abgefertigt, so ist $\psi_U(\vartheta) = \lambda(e^\vartheta - 1)$ und $\psi_V(\vartheta) = \vartheta$ und daher $\psi(\vartheta) = \lambda(e^\vartheta - 1) - \vartheta$. Zur Bestimmung von $\vartheta_*$ durch $\psi(\vartheta_*) = 0$ ist also die Gleichung $\lambda = f(\vartheta_*) := \vartheta_*/(e^{\vartheta_*} - 1)$ zu lösen. Da $f'(\vartheta) = \frac{e^\vartheta - 1 - \vartheta e^\vartheta}{(e^\vartheta - 1)^2} = \frac{-1}{e^\vartheta - 1} \left(e^\vartheta + \frac{1}{e^\vartheta - 1}\right) < 0$, $\lim_{\vartheta \downarrow 0} f(\vartheta) = 1$ und $\lim_{\vartheta\to\infty} f(\vartheta) = 0$, gibt es für $0 < \lambda = E[U] < E[V] = 1$ genau eine Lösung $\vartheta_* > 0$.

# 8 Der Satz von Cramer im $\mathbb{R}^d$ und das Gärtner-Ellis Theorem

**Satz 8.1** ***(Gärtner-Ellis)*** *Seien $X_1, X_2, \ldots$ $\mathbb{R}^d$-wertige ZVn, für die*

$$\psi(\vartheta) := \lim_{n\to\infty} \frac{1}{n}\psi_{S_n}(\vartheta) \in \mathbb{R} \quad \textit{existiert für alle } \vartheta \in \mathbb{R}^d$$

*und für die $\psi$ differenzierbar ist.*

*Ist dann die Folge der Verteilungen $(P_{\bar{S}_n})_n$ exponentiell straff, so erfüllt sie ein volles LDP mit Ratenfunktion $\psi^*$.*

*Beweis:* **(Skizze)** Die obere Abschätzung für kompakte Mengen wurde schon in Satz 6.3 gezeigt. Aus der exponentiellen Straffheit folgt dann die obere Abschätzung für abgeschlossene Mengen (siehe Satz 6.10). Die untere Abschätzung erfolgt zunächst auch wie in (18): Sei $G \subseteq \mathbb{R}^d$ offen. Für jedes $z \in G$ und hinreichend kleine $\delta > 0$ ist

$$P\{\bar{S}_n \in G\} \geqslant P\{\bar{S}_n \in U_\delta(z)\} = e^{\psi_{S_n}(\vartheta) - \langle n\vartheta, z\rangle \pm (n\delta\|\vartheta\|)} \tilde{P}_\vartheta\{\bar{S}_n \in U_\delta(z)\}$$

mit $\tilde{P}_\vartheta = e^{-\psi_{S_n}(\vartheta) + \langle \vartheta, S_n\rangle} P$, also

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathrm{P}\{\bar{S}_n \in G\} \quge \psi(\vartheta) - \langle \vartheta, z\rangle \pm (\delta\|\vartheta\|) + \liminf_{n\to\infty} \frac{1}{n} \log \tilde{P}_\vartheta\{\bar{S}_n \in U_\delta(z)\}.$$

Doch dann kann man $\lim_{n\to\infty} \tilde{P}_\vartheta\{\bar{S}_n \in U_\delta(z)\} = 1$ wegen der fehlenden Unabhängigkeit nicht direkt aus einem schwachen Gesetz der großen Zahl unter $\tilde{P}_\vartheta$ folgern (selbst wenn $\vartheta = \Gamma(z)$!). Stattdessen zeigt man, dass

$$\lim_{n\to\infty} \tilde{P}_\vartheta\{\bar{S}_n \in M \setminus U_\delta(z)\} = 0 \tag{24}$$

indem man die obere Abschätzung auf die abgeschlossene Menge $M \setminus U_\delta(z)$ und den Prozess $\bar{S}_n$ unter $\tilde{P}_\vartheta$ anwendet. Details des Beweises (einer Verallgemeinerung dieses Satzes) findet man in [19, Satz 3.4.4]. □

**Satz 8.2** ***(Cramer im $\mathbb{R}^d$)*** *Seien $X_1, X_2, \ldots$ u.i.v. $\mathbb{R}^d$-wertige ZVn. Ist $\Theta = \mathbb{R}^d$, d.h. $|\psi_X(\vartheta)| < \infty$ für alle $\vartheta \in \mathbb{R}^d$, so erfüllt die Folge $(\bar{S}_n)_n$ ein volles LDP mit Ratenfunktion $\psi_X^*$. Es ist $\psi_X^*(z) = D(f_{\Gamma^{-1}(z)}\|\mu)$ für $z \in \Gamma(\mathring{\Theta})$.*

*Beweis:* Das ist im wesentlichen ein Korollar zum Satz von Gärtner-Ellis. Nur die exponentielle Straffheit ist noch zu zeigen. Die folgt aus Satz 6.11.

**Untere Abschätzung**: Im Fall unabhängiger $X_k$ kann man die etwas schwierigen Argumente der konvexen Analysis aus dem Beweis in [19, Satz 3.4.4] vermeiden. Sei wiederum $z \in G$ beliebig. Wir unterscheiden zwei Fälle:

A)

$$\exists c > 0\ \forall \vartheta \in \mathring{\Theta} :\ \langle \vartheta, z - \Gamma(\vartheta)\rangle \leqslant -c \ \text{ oder } \ \|z - \Gamma(\vartheta)\| \geqslant c. \tag{25}$$

Es reicht zu zeigen, dass

$$\psi_X^*(z) = \sup_{\vartheta\in\mathbb{R}} (\langle \vartheta, z\rangle - \psi_X(\vartheta)) = \infty. \tag{26}$$

Das wollen wir durch Angabe einer Kurve $t \mapsto \vartheta_t$ $(t \geqslant 0)$ erreichen, entlang derer $\langle \vartheta_t, z\rangle - \psi_X(\vartheta_t)$ unbeschränkt wächst. Wie sollte eine solche Kurve aussehen?

$$\begin{aligned} \langle z, \vartheta_t\rangle - \psi_X(\vartheta_t) &= \langle z, \vartheta_0\rangle + \int_0^t \langle z, \dot{\vartheta}_s\rangle ds - \psi_X(\vartheta_0) - \int_0^t D\psi_X(\vartheta_s)\cdot\dot{\vartheta}_s ds \\ &= \langle z, \vartheta_0\rangle - \psi_X(\vartheta_0) + \int_0^t \langle z - \Gamma(\vartheta_s), \dot{\vartheta}_s\rangle ds \end{aligned} \tag{27}$$

Um das Integral als Funktion von $t$ möglichst große zu machen, lösen wir das Anfangswertproblem

$$\dot{\vartheta}_t = z - \Gamma(\vartheta_t), \vartheta_0 = 0.$$

Da $\Gamma$ differenzierbar ist, ist es insbesondere lokal Lipschitz-stetig, und es gibt ein maximales Lösungsintervall $[0, T)$, $0 < T \leqslant \infty$, in dem die Lösung die folgende Eigenschaft hat:

$$\frac{d}{dt}\langle\dot{\vartheta}_t, \dot{\vartheta}_t\rangle = 2\langle\dot{\vartheta}_t, \ddot{\vartheta}_t\rangle = -2\langle\dot{\vartheta}_t, D\Gamma(\vartheta_t)\dot{\vartheta}_t\rangle \leqslant 0. \tag{28}$$

(Beachte, dass $D\Gamma$ positiv semidefinit ist.) Insbesondere ist $\sup_{0\leqslant t<T} \|\dot{\vartheta}_t\|^2 \leqslant \|\dot{\vartheta}_0\|^2 = \|z - \Gamma(\vartheta_0)\|^2 = \|z - \Gamma(0)\|^2 < \infty$. Wäre $T < \infty$, so wäre damit auch $\sup_{0\leqslant t<T} \|\vartheta_t\| < \infty$. Da $\Gamma$ auf ganz $\mathbb{R}^d$ definiert ist, würde das aber der Maximalität von $T$ widersprechen. Daher ist $T = \infty$, so dass die obige Rechnung wie folgt für alle $t > 0$ fortgeführt werden kann:

$$\psi_X^*(z) \geqslant \langle z, \vartheta_t\rangle - \psi_X(\vartheta_t) = -\psi_X(0) + \int_0^t \langle z - \Gamma(\vartheta_s), z - \Gamma(\vartheta_s)\rangle ds = \int_0^t \|z - \Gamma(\vartheta_s)\|^2 \, ds \ . \tag{29}$$

Sei $M := \int_0^\infty \|z - \Gamma(\vartheta_s)\|^2 \, ds$. Die Behauptung (26) folgt, falls $M = \infty$. Wir zeigen das nun durch Widerspruch: Angenommen, $M < \infty$. Sei $U := \{s \in [0, \infty) : \langle\vartheta_s, z - \Gamma(\vartheta_s)\rangle \leqslant -c\}$. Für $s \in U$ gilt dann:

$$\frac{d}{ds}\|\vartheta_s\|^2 = \frac{d}{ds}\langle\vartheta_s, \vartheta_s\rangle = 2\langle\vartheta_s, \dot{\vartheta}_s\rangle = 2\langle\vartheta_s, z - \Gamma(\vartheta_s)\rangle \leqslant -2c,$$

während für beliebige $s > 0$ nur $\frac{d}{ds}\langle\vartheta_s, \vartheta_s\rangle \leqslant 2\|\vartheta_s\| \cdot \|z - \Gamma(\vartheta_s)\|$ gilt. Außerdem ist $\|z - \Gamma(\vartheta_s)\| \geqslant c$ für $s \in U^c$ nach Annahme (25), also $M \geqslant \int_{U^c} \|z - \Gamma(\vartheta_s)\|^2 \, ds \geqslant c^2 \cdot \lambda(U^c)$, wo $\lambda$ das Lebesgue-Maß auf $\mathbb{R}$ bezeichnet. Es folgt $\lambda(U^c) \leqslant c^{-2} M < \infty$.

Daher ist für jedes $t > 0$:

$$\begin{aligned}
\|\vartheta_t\|^2 &= \|\vartheta_0\|^2 + \int_0^t \frac{d}{ds}\langle\vartheta_s, \vartheta_s\rangle \, ds = \int_0^t 1_U(s)\frac{d}{ds}\langle\vartheta_s, \vartheta_s\rangle \, ds + \int_0^t 1_{U^c}(s)\frac{d}{ds}\langle\vartheta_s, \vartheta_s\rangle \, ds \\
&\leqslant -2c \cdot \lambda([0, t] \cap U) + 2\int_0^t 1_{U^c}(s)\|\vartheta_s\| \cdot \|z - \Gamma(\vartheta_s)\| \, ds \\
&\leqslant -2c(t - \lambda(U^c)) + 2\left(\int_0^t 1_{U^c}(s)\|\vartheta_s\|^2 \, ds\right)^{1/2} \left(\int_0^t \|z - \Gamma(\vartheta_s)\|^2 \, ds\right)^{1/2} \ .
\end{aligned}$$

Setze $f(t) := \sup_{0\leqslant s\leqslant t} \|\vartheta_s\|$. Dann folgt für $t > 0$:

$$0 \leqslant \|\vartheta_t\|^2 \leqslant -2c(t - \lambda(U^c)) + 2\left(\lambda(U^c)f(t)^2\right)^{1/2} M^{1/2}$$

Wäre nun $\sup_t \|\vartheta_t\| < \infty$, so wäre auch $\sup_t f(t) < \infty$, d.h. der zweite Summand wäre beschränkt, während der erste mit $t \to \infty$ gegen $-\infty$ strebt – ein Widerspruch zur Positivität der Summe beider Terme. Also ist $\sup_t \|\vartheta_t\| = \infty$, und es gibt $t_1 < t_2 < \cdots \to \infty$ derart, dass $f(t_k) = \|\vartheta_{t_k}\| \to \infty$. Für große $k$ ist daher

$$f(t_k)^2 = \|\vartheta_{t_k}\|^2 \leqslant -2ct_k + 2c\lambda(U^c) + 2(M\lambda(U^c))^{1/2} \cdot f(t_k),$$

also

$$f(t_k) \cdot \left(f(t_k) - 2(M\lambda(U^c))^{1/2}\right) \leqslant -2ct_k + 2c\lambda(U^c) \ ,$$

der gesuchte Widerspruch, da die linke Seite für $k \to \infty$ gegen $+\infty$, die rechte Seite aber gegen $-\infty$ strebt.

B) Wenn Fall A) nicht zutrifft, so gilt:

$$\forall k \in \mathbb{N} \ \exists \vartheta_k \in \mathring{\Theta} : \ \langle\vartheta_k, z - \Gamma(\vartheta_k)\rangle \geqslant -k^{-1} \ \text{ und } \ \|z - \Gamma(\vartheta_k)\| \leqslant k^{-1} \ . \tag{30}$$

Setze $\delta_k := 2\|z - \Gamma(\vartheta_k)\|$ und $\rho_k := \delta_k/(2 + \|\vartheta_k\|)$. Für ausreichend große $k$ ist dann $U_{\rho_k}(\Gamma(\vartheta_k)) \subseteq U_{\delta_k}(z) \subseteq G$, und wir haben (mit den üblichen Überlegungen)

$$\begin{aligned} P\{\bar{S}_n \in G\} &\geqslant P\{\bar{S}_n \in U_{\rho_k}(\Gamma(\vartheta_k))\} \\ &= \int_{\{\bar{S}_n \in U_{\rho_k}(\Gamma(\vartheta_k))\}} e^{\langle n\vartheta_k, \bar{S}_n\rangle - n\psi_X(\vartheta_k) + \langle n\vartheta_k, \Gamma(\vartheta_k) - \bar{S}_n\rangle + \langle n\vartheta_k, z - \Gamma(\vartheta_k)\rangle - \langle n\vartheta_k, z\rangle + n\psi_X(\vartheta_k)} dP \\ &\geqslant e^{-n(\langle \vartheta_k, z\rangle - \psi_X(\vartheta_k))} \cdot e^{n(\langle \vartheta_k, z - \Gamma(\vartheta_k)\rangle - \|\vartheta_k\|\rho_k)} \cdot P_{\vartheta_k}\{\bar{S}_n \in U_{\rho_k}(\Gamma(\vartheta_k))\}. \end{aligned}$$

Nach dem schwachen Gesetz der großen Zahl ist $\lim_{n\to\infty} P_{\vartheta_k}\{\bar{S}_n \in U_{\rho_k}(\Gamma(\vartheta_k))\} = 1$ für jedes $k$. Also ist wegen (30) für jedes ausreichend große $k$

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \log P\{\bar{S}_n \in G\} &\geqslant -\left(\langle \vartheta_k, z\rangle - \psi_X(\vartheta_k)\right) + \langle \vartheta_k, z - \Gamma(\vartheta_k)\rangle - \delta_k \\ &\geqslant -\psi_X^*(z) - k^{-1} - 2k^{-1} \ , \end{aligned}$$

und die gesuchte untere Abschätzung folgt im Limes $k \to \infty$. □

# 9 Der Satz von Sanov

Sei $E$ ein *polnischer Raum*. $\mathcal{E}$ sei die zugehörige Borelsche $\sigma$-Algebra, und $\mathcal{P}$ sei der Raum der Borel-W'maße auf $E$. $\mathcal{P}$, versehen mit der Konvergenz der Topologie der schwachen Konvergenz ist ein topologischer Raum und trägt deshalb selbst eine Borel-$\sigma$-Algebra.

**Bemerkung 9.1** Ist $\nu \in \mathcal{P}$, so bilden die folgenden Mengen $U_{T,\delta}(\nu)$ eine *Umgebungsbasis* von $\nu$ (d.h. jede Umgebung von $\nu$ enthält eine Menge $U_{T,\delta}(\nu)$): Für $T_1, \ldots, T_d \in C_b(E)$, $d \in \mathbb{N}$ und $\delta > 0$ ist

$$U_{T,\delta}(\nu) := \left\{ \tilde{\nu} \in \mathcal{P} : \left| \int T_i d\tilde{\nu} - \int T_i d\nu \right| < \delta (i = 1, \ldots, d) \right\}.$$

Sind $\mu_n, \mu \in \mathcal{P}$, so ist $\lim_{n\to\infty} \mu_n = \mu$ im Sinne der schwachen Konvergenz, falls $\lim_{n\to\infty} \int u d\mu_n = \int u d\mu$ für alle stetigen beschränkten $u : E \to \mathbb{R}$. (In funktionalanalytischem Jargon würde man eher von schwach-$*$-Konvergenz sprechen. Details findet man z.B. in [18, Kapitel 13].)

**Bemerkung 9.2** Sei $\mu \in \mathcal{P}$. Dann ist $I : \mathcal{P} \to [0, \infty]$, $I(\nu) = D(\nu\|\mu)$, eine Ratenfunktion: In Satz 5.4 wurde für $\Psi(u) = \log \int e^u d\mu$ und für kompaktes $E$ gezeigt, dass

$$I(\nu) = D(\nu\|\mu) = \Psi^*(\nu) = \sup \left\{ \int u d\nu - \Psi(u) : u \in C_b(E) \right\},$$

so dass $I(\nu)$ als Supremum stetiger Funktionen unterhalbstetig ist. Das bleibt für allgemeine polnische Räume $E$ richtig. (Die Beweise in der Literatur und in Skripten sind leider oft nicht vollständig oder nicht korrekt.)

Seien nun $X_1, X_2, \ldots$ u.i.v. $E$-wertige ZVn mit Verteilung $\mu$. Natürlich ist $\mu \in \mathcal{P}$. Für $n > 0$ sei $\varepsilon_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ die empirische Verteilung von $X_1, \ldots, X_n$. Die $\varepsilon_n$ können als $\mathcal{P}$-wertige ZVn aufgefasst werden.

**Satz 9.3** ***(Sanov)*** *Die Folge von Verteilungen $(P_{\varepsilon_n})_{n\geqslant 1}$ auf $\mathcal{P}$ erfüllt ein volles LDP mit Ratenfunktion $I(\nu) = D(\nu\|\mu)$. Das heißt:*

*a) Für jede offene Menge $G \subseteq \mathcal{P}$ ist*

$$\liminf_{n\to\infty} \frac{1}{n} \log P\{\varepsilon_n \in G\} \geqslant - \inf_{\nu\in G} D(\nu\|\mu)$$

*b) Für jede abgeschlossene Menge $F \subseteq \mathcal{P}$ ist*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{\varepsilon_n \in F\} \leqslant - \inf_{\nu\in F} D(\nu\|\mu)$$

Dieser Satz wird in vielen unterschiedlichen Versionen formuliert. Die obige Version ist recht allgemein [17, Theorem 27.15] oder [19, Satz 2.4.1]. Eine noch allgemeinere (ohne topologische Annahme an den Raum $E$) findet man in [9, Kapitel 23], eine einfachere, bei der $E$ eine endliche Menge ist, in [18, Satz 23.13]. Hier geben wir einen *Beweis für den Fall, dass $E$ ein kompakter metrischer Raum* ist. Dann ist auch $\mathcal{P}$ mit der Topologie der schwachen Konvergenz kompakt. (Die im Fall eines polnischen $E$ notwendigen zusätzlichen Straffheitsüberlegungen findet man in [19, Satz 2.4.1].)

*Beweis:* **des Satzes von Sanov:** a) Sei $G \subseteq \mathcal{P}$ offen.

- Sei $\nu \in G$. Da $G$ offen ist, gibt es $T_1, \ldots, T_d \in C_b(E)$ und $\delta > 0$, so dass $U_{T,\delta}(\nu) \subseteq G$,wobei $T = (T_1, \ldots, T_d)$.
- Sei $\psi(\vartheta) = \log \int e^{\langle \vartheta, T\rangle} d\mu$ wie vorher. Wir betrachten die u.i.v. ZVn $Y_k = T(X_k)$ und $\bar{S}_{n,Y} = \frac{1}{n}(Y_1 + \ldots + Y_n) = \frac{1}{n} \sum_{k=1}^n T(X_k) = \int T \, d\varepsilon_n$. Beachte, dass $\psi_Y = \psi$.
- Es ist $\varepsilon_n \in U_{T,\delta}(\nu)$ genau dann, wenn $\bar{S}_{n,Y} \in B_\delta(E_\nu[T])$.

- Da $T$ beschränkt ist, ist in dieser Situation $\Theta = \mathring{\Theta} = \mathbb{R}^d$, und aus der unteren Abschätzung im Satz von Cramer für den $\mathbb{R}^d$ folgt

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log P\{\varepsilon_n \in G\} &\geqslant \liminf_{n\to\infty} \frac{1}{n} \log P\{\varepsilon_n \in U_{T,\delta}(\nu)\} \\
&= \liminf_{n\to\infty} \frac{1}{n} \log P\{\bar{S}_{n,Y} \in B_\delta(E_\nu[T])\} \\
&\geqslant -\inf\{\psi^*(z) : z \in B_\delta(E_\nu[T])\} \\
&\geqslant -\psi^*(E_\nu[T]) \\
&= -\sup\{\langle \vartheta, E_\nu[T]\rangle - \psi(\vartheta) : \vartheta \in \mathbb{R}^d\} \\
&= -\sup\left\{\int \langle \vartheta, T\rangle d\nu - \log \int e^{\langle \vartheta, T\rangle} d\mu : \vartheta \in \mathbb{R}^d\right\} \\
&\geqslant -\sup\left\{\int u d\nu - \log \int e^u d\mu : u \in \mathrm{mb}_b(E)\right\} \\
&= -D(\nu\|\mu)
\end{aligned}$$

  wegen Gleichung (11) aus Lemma 5.1.

b) Sei $F \subseteq \mathcal{P}$ abgeschlossen, also kompakt. Für die obere Abschätzung können wir o.B.d.A. annehmen, dass $s := \inf_{\nu\in F} D(\nu\|\mu) > 0$ und betrachten $\alpha \in (0, s)$. Für jedes $\nu \in F$ ist nach Satz 5.4

$$\alpha < D(\nu\|\mu) = \Psi^*(\nu) = \sup\left\{\int u d\nu - \log \int e^u d\mu : u \in C_b(E)\right\}.$$

Also gibt es ein $u_\nu \in C_b(E)$, so dass $\int u_\nu d\nu - \log \int e^{u_\nu} d\mu > \alpha$. Daher ist $\nu \in V_\nu$, wo

$$V_\nu := \left\{\tilde{\nu} : \ \tilde{\nu} \text{ W'maß auf } E, \int u_\nu d\tilde{\nu} - \log \int e^{u_\nu} d\mu > \alpha\right\} \quad \text{(offen)}.$$

Es ist

$$\begin{aligned}
P\{\varepsilon_n \in V_\nu\} &= P\left\{\int u_\nu d\varepsilon_n > \log \int e^{u_\nu} d\mu + \alpha\right\} \\
&= P\left\{\exp\left(n \int u_\nu d\varepsilon_n\right) > \exp\left(n\alpha + n \log \int e^{u_\nu} d\mu\right)\right\} \\
&\leqslant \exp\left(-n\alpha - n \log \int e^{u_\nu} d\mu\right) \cdot E\left[\exp\left(n \int u_\nu d\varepsilon_n\right)\right] \\
&= e^{-n\alpha} \cdot \left(\int e^{u_\nu} d\mu\right)^{-n} \cdot E[\exp(u_\nu(X_1) + \ldots + u_\nu(X_n))] \\
&= e^{-n\alpha},
\end{aligned}$$

und da $F$ durch endlich viele solche $V_\nu$ überdeckt werden kann, ist

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{\varepsilon_n \in F\} \leqslant -\alpha,$$

siehe Lemma 6.2. Da das für jedes $\alpha < s = \inf_{\nu\in F} D(\nu\|\mu)$ gilt, folgt die obere Abschätzung. □

**Proposition 9.4** *Die Ratenfunktion $I(\nu) = D(\nu\|\mu)$ im Satz von Sanov ist* gut.

*Beweis:* Sei $a > 0$. Zu zeigen ist, dass $\mathcal{K} := I^{-1}((-\infty, a]) \subseteq \mathcal{P}$ kompakt ist. Dazu reicht es, die Straffheit von $\mathcal{K}$ zu zeigen, denn daraus folgt die relative Kompaktheit (Satz von Prohorov), und da $\mathcal{K}$ wegen der Unterhalbstetigkeit von $I$ abgeschlossen ist, folgt daraus die Kompaktheit von $\mathcal{K}$.

Die Straffheit von $\mathcal{K}$ zeigt man so: Sei $\varepsilon > 0$. Setze $S := 2(a + e^{-1})\varepsilon^{-1}$. Da $\mu$ als Borel-Maß auf einem polnischen Raum regulär ist, gibt es eine kompakte Teilmenge $L$ von $E$ mit $\mu(E \setminus L) < \frac{\varepsilon}{2e^S}$. Daher ist für alle $\nu = f\mu \in \mathcal{K}$:

$$\begin{aligned}
\nu(E \setminus L) &= \int_{E\setminus L} f d\mu \quad = \quad \int_{(E\setminus L)\cap\{f\leqslant e^S\}} f d\mu + \int_{(E\setminus L)\cap\{f>e^S\}} \frac{\varphi\circ f}{\log f} d\mu \\
&\leqslant e^S \mu(E \setminus L) + S^{-1} \int_{\{f>e^S\}} \varphi \circ f d\mu \\
&\leqslant e^S \mu(E \setminus L) + S^{-1} \int_{\{f\geqslant 1\}} \varphi \circ f d\mu \\
&\leqslant \frac{\varepsilon}{2} + S^{-1} \left( I(\nu) - \int_{\{f<1\}} \varphi \circ f d\mu \right) \\
&\leqslant \frac{\varepsilon}{2} + S^{-1}(a + e^{-1}) \quad = \quad \varepsilon
\end{aligned}$$

Das zeigt die Straffheit von $\mathcal{K}$. □

Wir beschließen dieses Kapitel mit einem Sanov-artigen Satz, den wir als Korollar des Satzes von Gärtner und Ellis erhalten. Sein Beweis beruht auf dem Satz von Perron und Frobenius, den wir in der folgenden Bemerkung zusammen fassen:

**Bemerkung 9.5 (Satz von Perron und Frobenius)** Sei $A$ eine irreduzible, aperiodische $d \times d$-Matrix mit nichtnegativen Einträgen, d.h. es gebe ein $k \in \mathbb{N}$, für das $A^k$ nur strikt positive Koeffizienten hat. Der Satz von Perron und Frobenius besagt:

- $A$ hat einen einfachen führenden Eigenwert $\lambda > 0$, d.h. alle andere Eigenwerte von $A$ haben Betrag echt kleiner als $\lambda$. $u$ und $v$ bezeichnen den zugehörigen Links- bzw. Rechtseigenvektor, also $u^T A = \lambda u^T$, $Av = \lambda v$. $u$ und $v$ haben positive Koeffizienten und können durch $\langle u, \mathbf{1}\rangle = 1$ und $\langle u, v\rangle = 1$ eindeutig normiert werden. Dabei bezeichnet $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^d$.
- Es gibt ein $\epsilon > 0$, so dass für alle $f \in \mathbb{R}^d$ und $n \in \mathbb{N}$ gilt: $A^n f = \langle u, f\rangle \lambda^n v + O((\lambda - \epsilon)^n)\|f\|$, also insbesondere $A^n \mathbf{1} = \lambda^n v + O((\lambda - \epsilon)^n)$.

Darüberhinaus sind folgende Beobachtungen von Bedeutung:

- Setzt man $\pi_j := u_j \cdot v_j$, so ist $\pi$ ein W'vektor, denn $\langle \pi, \mathbf{1}\rangle = \langle u, v\rangle = 1$, und es gilt für die Matrix $\tilde{A}$, die definiert ist durch$\tilde{A}_{ij} = \lambda^{-1} v_i^{-1} A_{ij} v_j$:

  $$(\tilde{A}\mathbf{1})_i = \lambda^{-1} v_i^{-1} (Av)_i = v_i^{-1} v_i = 1 \quad \text{für alle } i,$$

  d.h. $\tilde{A}$ ist eine stochastische Matrix, und

  $$(\pi^T \tilde{A})_j = \sum_{i=1}^{d} u_i v_i \lambda^{-1} v_i^{-1} A_{ij} v_j = u_j v_j = (\pi^T)_j,$$

  d.h. $\pi$ ist der eindeutige stationäre W'vektor für $\tilde{A}$.

**Satz 9.6 (Sanov-Satz für Markovketten mit endlichem Zustandsraum)**
*Sei $Z_0, Z_1, Z_2, \ldots$ eine Markovkette mit Zustandsraum $E = \{1, \ldots, d\}$ und Übergangswahrscheinlichkeiten*

$$q_{ij} = P(Z_n = j | Z_{n-1} = i) \qquad (i, j = 1, \ldots, d).$$

*Die Matrix $Q$ sei irreduzibel und aperiodisch. Bezeichne $\varepsilon_n := \frac{1}{n} \sum_{k=1}^{n} \delta_{Z_k}$ die empirische Verteilung, also*

$$(\varepsilon_n)_i = \frac{1}{n} \#\{k \in \{1, \ldots n\} : Z_k = i\}.$$

*Dann erfüllt $(P_{\varepsilon_n})_{n>0}$ ein LDP mit Ratenfunktion $I(\mu) = \psi^*(\mu) = \sup_{\vartheta\in\mathbb{R}^d}(\langle\vartheta, \mu\rangle - \psi(\vartheta))$, $\psi(\vartheta) = \log \lambda_\vartheta$, wo $\lambda_\vartheta$ der eindeutig bestimmte führende Eigenwert der Matrix $Q_\vartheta$, $(Q_\vartheta)_{ij} = q_{ij} e^{\vartheta_j}$, ist.*

*Beweis:* Wir betrachten die $\mathbb{R}^d$-wertigen ZVn $X_k := \boldsymbol{e}_{Z_k}$, d.h. $X_k = \boldsymbol{e}_i$ gdw. $Z_k = i$. Dann ist

$$\bar{S}_n = \frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{e}_{Z_k} = \frac{1}{n}\sum_{k=1}^{n} \delta_{Z_k} = \varepsilon_n,$$

wobei die Einheitsvektoren $\boldsymbol{e}_i$ mit den Wahrscheinlichkeitsvektoren(=Wahrscheinlichkeitsmaßen) $\delta_i$ ($i \in \{1,\ldots,d\}$) identifiziert wurden. Daher reicht es ein LDP für $(P_{\bar{S}_n})_{n>0}$ zu zeigen. Das werden wir jetzt aus dem Gärtner-Ellis Theorem herleiten. Dazu muss

$$\psi(\vartheta) = \lim_{n\to\infty} \frac{1}{n}\psi_{S_n}(\vartheta) = \lim_{n\to\infty} \frac{1}{n}\log E[e^{\langle \vartheta, S_n\rangle}]$$

bestimmt werden.

Mit $Q$ sind auch die $Q_\vartheta$ irreduzibel und aperiodisch. Insbesondere haben sie einen einfachen führenden Eigenwert $\lambda_\vartheta > 0$ mit zugehörigem strikt positiven Eigenvektor $v_\vartheta$, der so normiert ist, dass $Q_\vartheta^n \mathbf{1} = \lambda_\vartheta^n v_\vartheta + O_\vartheta((\lambda_\vartheta - \epsilon_\vartheta)^n)$ für ein $\epsilon_\vartheta > 0$. (Das folgt aus dem Satz von Perron und Frobenius, siehe Bemerkung 9.5.)

Für jede Funktion $f : \{1,\ldots,d\} \to \mathbb{R}$ (also $f \in \mathbb{R}^d$) gilt nun:

$$\begin{aligned}
E[e^{\langle \vartheta, S_n\rangle} f_{Z_n}] &= E[e^{\langle \vartheta, S_{n-1}\rangle} \cdot e^{\langle \vartheta, X_n\rangle} f_{Z_n}] \\
&= E\left[e^{\langle \vartheta, S_{n-1}\rangle} \cdot E\left[e^{\vartheta_{Z_n}} f_{Z_n} | Z_0,\ldots,Z_{n-1}\right]\right] \\
&= E\left[e^{\langle \vartheta, S_{n-1}\rangle} \cdot E\left[e^{\vartheta_{Z_n}} f_{Z_n} | Z_{n-1}\right]\right] \\
&= E\left[e^{\langle \vartheta, S_{n-1}\rangle} \cdot \sum_{j=1}^{d} q_{Z_{n-1},j} e^{\vartheta_j} f_j\right] \\
&= E[e^{\langle \vartheta, S_{n-1}\rangle} \cdot (Q_\vartheta f)_{Z_{n-1}}]
\end{aligned}$$

Durch iterative Anwendung dieser Gleichung auf $f = \mathbf{1}$, $f = Q_\vartheta \mathbf{1}, \ldots, f = Q_\vartheta^{n-1}\mathbf{1}$ folgt:

$$E[e^{\langle \vartheta, S_n\rangle}] = E[(Q_\vartheta^n \mathbf{1})_{Z_0}] = \lambda_\vartheta^n E[(v_\vartheta)_{Z_0}] + O_\vartheta((\lambda_\vartheta - \epsilon_\vartheta)^n) = \lambda_\vartheta^n \cdot (E[(v_\vartheta)_{Z_0}] + o_\vartheta(1)),$$

so dass

$$\psi(\vartheta) = \lim_{n\to\infty} \frac{1}{n}\log E[e^{\langle \vartheta, S_n\rangle}] = \log \lambda_\vartheta.$$

Daher erfüllt $(P_{\bar{S}_n})_{n>0}$, und damit auch $(P_{\varepsilon_n})_{n>0}$, ein volles LDP mit guter Ratenfunktion $I(\mu) = \psi^*(\mu)$, wobei $\psi(\vartheta) = \log \lambda_\vartheta$. (Beachte, dass $E = \{1,\ldots,d\}$ kompakt ist, und interpretiere $\mu$ als Wahrscheinlichkeitsvektor aus dem $\mathbb{R}^d$.) □

**Bemerkung 9.7** Durch störungstheoretische Betrachtungen kann man zeigen, dass

$$\frac{d}{d\vartheta}\log \lambda_\vartheta = E_\vartheta[X_0]$$

und

$$\frac{d^2}{d\vartheta^2}\log \lambda_\vartheta = V_\vartheta[X_0] + 2\sum_{k=1}^{\infty} \mathrm{Cov}_\vartheta[X_0, X_k],$$

wo $E_\vartheta, V_\vartheta, \mathrm{Cov}_\vartheta$ bzgl. des stationären Markov-Maßes zu $\pi_\vartheta$ und $\widetilde{Q_\vartheta}$ aus Bemerkung 9.5 gebildet werden. Beachte dabei, dass $E_\vartheta[X_0] = E_\vartheta[\boldsymbol{e}_{Z_0}] = \pi_\vartheta$.

# 10 Das Kontraktionsprinzip

**Satz 10.1** *Seien $E$ und $F$ polnische Räume, und sei $T : E \to F$ stetig. Erfüllt die Folge von Wahrscheinlichkeitsverteilungen $(\mu_n)_{n\geqslant 1}$ auf $E$ ein volles LDP mit guter Ratenfunktion $I$, so erfüllt die Folge $(\mu_n \circ T^{-1})_{n\geqslant 1}$ von Wahrscheinlichkeitsverteilungen auf $F$ ein volles LDP mit guter Ratenfunktion $J(y) := \inf\{I(x) : T(x) = y\}$. (Hier ist ggf. $\inf \emptyset = \infty$.)*
***Zur Verdeutlichung:** Ist $\mu_n = P_{Z_n}$ , so ist $\mu_n \circ T^{-1} = P_{T(Z_n)}$.*

*Beweis:* - $J$ ist eine gute Ratenfunktion: Sei $a \in \mathbb{R}$. Es gilt:

$$\begin{aligned} y \in J^{-1}((-\infty, a]) \quad &\Rightarrow \quad \exists x_n \in T^{-1}\{y\} : \limsup_{n\to\infty} I(x_n) \leqslant a \\ &\Rightarrow \quad \exists x'_n \in T^{-1}\{y\} : \limsup_{n\to\infty} I(x'_n) \leqslant a \text{ und } (x'_n)_{n\geqslant 1} \text{ konvergiert gegen ein } x \in E, \end{aligned}$$

denn die Menge $I^{-1}((-\infty, a+1])$ ist kompakt. Da $T$ stetig ist, ist $T(x) = \lim_{n\to\infty} T(x'_n) = y$, und da $I$ unterhalbstetig ist, ist $I(x) \leqslant \limsup_{n\to\infty} I(x_n) \leqslant a$. Also ist

$$y \in J^{-1}((-\infty, a]) \quad \Longleftrightarrow \quad \exists x \in I^{-1}((-\infty, a]) \text{ mit } T(x) = y,$$

so dass $J^{-1}((-\infty, a]) = T(I^{-1}((-\infty, a]))$. Da $I^{-1}((-\infty, a])$ nach Voraussetzung kompakt und $T$ stetig ist, ist auch $J^{-1}((-\infty, a])$ kompakt.
- Untere Abschätzung: Sei $G \subseteq F$ offen. Dann ist

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n \circ T^{-1}(G) \quad \geqslant \quad -\inf_{x\in T^{-1}(G)} I(x) = -\inf_{y\in G} \inf_{x\in T^{-1}\{y\}} I(x) = -\inf_{y\in G} J(y).$$

- Obere Abschätzung: Sei $A \subseteq F$ abgeschlossen. Dann ist

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n \circ T^{-1}(A) \leqslant -\inf_{x\in T^{-1}(A)} I(x) = -\inf_{y\in A} \inf_{x\in T^{-1}\{y\}} I(x) = -\inf_{y\in A} J(y).$$

□

**Bemerkung 10.2** Die obere und die untere Abschätzung bleiben richtig, auch wenn die Ratenfunktion $I$ nicht gut ist. Dann muss $J$ aber keine Ratenfunktion sein.

**Aufgabe* 10.1** Geben Sie ein Beispiel an, wo $E = F = \mathbb{R}$, $T : \mathbb{R} \to \mathbb{R}$ stetig, $I$ eine nicht gute Ratenfunktion auf $\mathbb{R}$ und $J$ gar keine Ratenfunktion (d.h. nicht unterhalb stetig) ist.

**Beispiel 10.3** Seien $X_1, X_2, \ldots$ u.i.v. $\mathbb{R}$-wertige ZVn, $\mu := P_{X_i}$. Wir fassen die empirischen Maße $\varepsilon_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ wieder als $\mathcal{P}$-wertige ZVn auf.

- Aus dem Satz von Sanov folgt: $(P_{\varepsilon_n})_{n\geqslant 1}$ erfüllt ein volles LDP mit Ratenfunktion $I : \mathcal{P} \to [0, \infty]$, $I(\nu) = D(\nu\|\mu)$.
- Aus Proposition 9.4 folgt, dass $I$ gut ist.

Sei nun $\varphi \in C_b(\mathbb{R}; \mathbb{R})$ fest gewählt. Wir definieren $h_1, h_2 : \mathcal{P} \to \mathbb{R}$ durch

$$\begin{aligned} h_1(\nu) &= \langle \varphi, \nu\rangle = \int \varphi d\nu \\ h_2(\nu) &= \int (\varphi - \langle \varphi, \nu\rangle)^2 d\nu = \langle \varphi^2, \nu\rangle - \langle \varphi, \nu\rangle^2 \end{aligned}$$

Da mit $\varphi$ auch $\varphi^2$ stetig und beschränkt ist, sind $h_1$ und $h_2$ stetig bzgl. der schwachen Topologie auf $\mathcal{P}$. Wir halten fest:

- $h_1(\varepsilon_n) = \frac{1}{n}\sum_{i=1}^n \varphi(X_i)$ ist das empirische Mittel der Beobachtungen $\varphi(X_1), \ldots, \varphi(X_n)$,

– $h_2(\varepsilon_n) = \frac{1}{n}\sum_{i=1}^n (\varphi(X_i))^2 - \left(\frac{1}{n}\sum_{i=1}^n \varphi(X_i)\right)^2$ ist bis auf einen Vorfaktor $\frac{n}{n-1}$ die empirische Varianz der Beobachtungen $\varphi(X_1), \ldots, \varphi(X_n)$.

Aus dem LDP für $(P_{\varepsilon_n})_{n\geqslant 1}$ und dem Kontraktionsprinzip folgt nun wegen $P_{h(\varepsilon_n)} = P_{\varepsilon_n} \circ h^{-1}$:
Die Familien $(P_{h_1(\varepsilon_n)})_{n\geqslant 1}$ und $(P_{h_2(\varepsilon_n)})_{n\geqslant 1}$ erfüllen ein volles LDP mit Ratenfunktionen

$$J_1(y) = \inf\{I(\nu) : h_1(\nu) = y\} = \inf\{D(\nu\|\mu) : \nu \in \mathcal{P}, \langle\varphi, \nu\rangle = y\}$$

und

$$J_2(y) = \inf\{I(\nu) : h_2(\nu) = y\} = \inf\{D(\nu\|\mu) : \nu \in \mathcal{P}, \langle\varphi^2, \nu\rangle - \langle\varphi, \nu\rangle^2 = y\}.$$

**Zu $J_1$:** Aus Satz 3.6 folgt: $J_1(y) = D(f_{\vartheta_y}\|\mu)$, wo $\vartheta_y$ so gewählt ist, dass $\int \varphi \cdot f_{\vartheta_y} d\mu = y$, falls das möglich ist. Aus Bemerkung 5.7 folgt weiter $J_1(y) = \psi^*(y)$ für $\psi(\vartheta) = \log \int e^{\vartheta\cdot\varphi(x)} d\mu(x)$, so dass man für $\frac{1}{n}\sum_{i=1}^n \varphi(X_i)$ wieder den Satz von Cramer erhält.
**Zu $J_2$:** Hierfür ist mir keine einfachere, explizitere Form bekannt.

**Beispiel 10.4** Seien $X_1, X_2, \ldots$ u.i.v. $\mathbb{R}^d$-wertig und nach $\mathcal{N}(0, V)$ verteilt. Folgende Identität werden wir zweimal benutzen:

$$\begin{aligned} -\frac{1}{2}\langle x - V\vartheta, V^{-1}(x - V\vartheta)\rangle &= -\frac{1}{2}\langle x, V^{-1}x\rangle + \frac{1}{2}\langle x, \vartheta\rangle + \frac{1}{2}\langle V\vartheta, V^{-1}x\rangle - \frac{1}{2}\langle V\vartheta, \vartheta\rangle \\ &= \langle\vartheta, x\rangle - \frac{1}{2}\langle x, V^{-1}x\rangle - \frac{1}{2}\langle\vartheta, V\vartheta\rangle. \end{aligned}$$

Dann ist

$$\begin{aligned} \psi_X(\vartheta) &= \log\left(\frac{1}{(2\pi)^{d/2}\det(V)^{1/2}}\int_{\mathbb{R}^d} e^{\langle\vartheta,x\rangle - \frac{1}{2}\langle x, V^{-1}x\rangle} dx\right) \\ &= \log\left(\frac{1}{(2\pi)^{d/2}\det(V)^{1/2}}\int_{\mathbb{R}^d} e^{-\frac{1}{2}\langle x - V\vartheta, V^{-1}(x-V\vartheta)\rangle + \frac{1}{2}\langle\vartheta, V\vartheta\rangle} dx\right) \\ &= \frac{1}{2}\langle\vartheta, V\vartheta\rangle \end{aligned}$$

und daher (durch Wahl von $\vartheta = V^{-1}z$)

$$\begin{aligned} \psi_X^*(z) &= \sup_{\vartheta\in\mathbb{R}^d}\left(\langle\vartheta, z\rangle - \frac{1}{2}\langle\vartheta, V\vartheta\rangle\right) \\ &= \sup_{\vartheta\in\mathbb{R}^d}\left(\frac{1}{2}\langle z, V^{-1}z\rangle - \frac{1}{2}\langle z - V\vartheta, V^{-1}(z - V\vartheta)\rangle\right) \\ &= \frac{1}{2}\langle z, V^{-1}z\rangle. \end{aligned}$$

Sei nun $\lambda_{\max}$ der größte Eigenwert von $V$. Dann folgt aus dem Satz von Cramer für $\mathbb{R}^d$-wertige Zufallsvariablen und dem Kontraktionsprinzip

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{n}\log P\{\|\bar{S}_n\| \geqslant r\} &= -\inf_{t\geqslant r}(\inf\{\psi_X^*(z) : \|z\| = t\}) \\ &= -\inf_{t\geqslant r}\left(\inf\{\frac{1}{2}\langle z, V^{-1}z\rangle : \|z\| = t\}\right) \\ &= -\inf_{t\geqslant r}\frac{t^2}{2}(\inf\{\langle z, V^{-1}z\rangle : \|z\| = 1\}) \\ &= -\frac{r^2}{2}\cdot\frac{1}{\lambda_{\max}}. \end{aligned}$$

Für die letzte Gleichung wurde benutzt, dass $V^{-1}$ eine positiv definite Matrix mit kleinstem Eigenwert $1/\lambda_{\max}$ ist.

# 11 Das Lemma von Varadhan und seine Umkehrung

**Satz 11.1 (Lemma von Varadhan)**
*Sei $(\mu_n)_{n\geqslant 1}$ eine Folge von W'verteilungen auf dem metrischen Raum $(M,d)$, die ein volles LDP mit guter Ratenfunktion $I$ erfüllt. Dann existiert für jedes stetige beschränkte $h : M \to \mathbb{R}$ der Limes*

$$\Lambda(h) := \lim_{n\to\infty} \frac{1}{n} \log \int_M e^{nh} d\mu_n \tag{31}$$

*und es gilt*

$$\Lambda(h) = \sup_{x\in M} (h(x) - I(x)). \tag{32}$$

*(Die Voraussetzung, dass h beschränkt ist, kann abgeschwächt werden; siehe z.B. [18, Satz 23.17].)*

**Beispiel 11.2**

- $X_1, X_2, \ldots$ u.i.v., $|X_i| \leqslant a$. Betrachte $M = [-a, a]$.
- $S_n = X_1 + \ldots + X_n$, $\overline{S}_n = S_n/n$, $\mu_n = P_{\overline{S}_n}$ und $h(x) = \vartheta \cdot x$.
- Dann ist $\frac{1}{n} \log \int e^{nh} d\mu_n = \frac{1}{n} \log \int e^{\vartheta \cdot S_n} dP = \log \int e^{\vartheta \cdot X_1} dP = \psi_{X_1}(\vartheta)$.
- Es ist $\sup_{x\in M}(h(x) - I(x)) = \sup_{x\in M}(\vartheta \cdot x - \psi^*(x)) = \psi(\vartheta)$ (bei Wahl von $x = \Gamma(\vartheta)$).

*Beweis:* „$\geqslant$“: Sei $x \in M$ und $\delta > 0$. $U_\delta(x)$ sei die $\delta$-Umgebung von $x$. Dann ist für jedes $\delta > 0$

$$\begin{aligned}
\frac{1}{n} \log \int_M e^{nh} d\mu_n &\geqslant \frac{1}{n} \log \int_{U_\delta(x)} e^{nh} d\mu_n \\
&\geqslant \frac{1}{n} \log \big( \exp[n \cdot \inf_{y\in U_\delta(x)} h(y)] \cdot \mu_n(U_\delta(x)) \big) \\
&= \inf_{y\in U_\delta(x)} h(y) + \frac{1}{n} \log \mu_n(U_\delta(x)).
\end{aligned}$$

Aus dem LDP folgt:

$$\liminf_{n\to\infty} \frac{1}{n} \log \int_M e^{nh} d\mu_n \geqslant \inf_{y\in U_\delta(x)} h(y) - \inf_{y\in U_\delta(x)} I(y) \geqslant \inf_{y\in U_\delta(x)} h(y) - I(x).$$

Da $h$ stetig ist, konvergiert das Infimum für $\delta \to 0$ gegen $h(x)$. Also ist

$$\liminf_{n\to\infty} \frac{1}{n} \log \int_M e^{nh} d\mu_n \geqslant h(x) - I(x),$$

und da $x \in M$ beliebig gewählt war, folgt die $\geqslant$-Richtung von (32).

„$\leqslant$“: Seien $\lambda, \eta > 0$. Da $I$ „gut“ ist, ist die Menge $K := I^{-1}([0, \lambda])$ kompakt. Für jedes $x \in K$ gibt es ein $\delta(x) > 0$, so dass für die Umgebung $V_x = U_{\delta(x)}(x)$ von $x$ gilt:

$$\sup\{h(y) : y \in V_x\} \leqslant h(x) + \eta \quad \text{und} \quad \inf\{I(y) : y \in \overline{V_x}\} \geqslant I(x) - \eta.$$

Dabei wurde die Stetigkeit von $h$ und die Unterhalbstetigkeit von $I$ benutzt. Die $V_x$, $x \in K$, bilden eine offene Überdeckung der kompakten Menge $K$. Daher gibt es $x_1, \ldots, x_N \in K$ so dass $K \subseteq G := \bigcup_{j=1}^N V_{x_j}$. Es folgt:

$$\begin{aligned}
\int_M e^{nh} d\mu_n &= \int_G e^{nh} d\mu_n + \int_{M\setminus G} e^{nh} d\mu_n \\
&\leqslant \sum_{j=1}^N \int_{V_{x_j}} e^{nh} d\mu_n + \int_{M\setminus G} e^{nh} d\mu_n \\
&\leqslant \sum_{j=1}^N e^{n(h(x_j)+\eta)} \mu_n(V_{x_j}) + e^{n\cdot\sup\{h(y):y\in M\setminus G\}} \mu_n(M \setminus G).
\end{aligned}$$

Daher folgt aus dem LDP und Lemma 6.2:

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\int_M e^{nh}d\mu_n &\leqslant \max\left\{\max_{j=1,\dots,N}\left(h(x_j)+\eta-\inf_{\overline{V_{x_j}}}I\right),\sup_{x\in M}h(x)-\inf_{x\in M\setminus G}I(x)\right\}\\ &\leqslant \max\left\{\max_{j=1,\dots,N}\left(h(x_j)+\eta-(I(x_j)-\eta)\right),\sup_{x\in M}h(x)-\lambda\right\}\\ &\leqslant \max\left\{\sup_{x\in M}(h(x)-I(x)+2\eta),\sup_{x\in M}h(x)-\lambda\right\}.\end{aligned}$$

Lässt man für festes $\lambda$ zunächst $\eta\to 0$ gehen und dann erst $\lambda\to\infty$, so folgt die $\leqslant$-Richtung von (32). □

Das Lemma von Varadhan hat auch eine Umkehrung, die auf Bryc zurückgeht. Wir formulieren und beweisen sie hier nur für kompaktes $M$. Die allgemeine Version findet man in [17, Satz 27.10 und Bemerkungen dazu auf S. 594].

**Satz 11.3 (Satz von Bryc)** *Sei $(\mu_n)_{n\geqslant 1}$ eine Folge von W'verteilungen auf dem kompakten metrischen Raum $M$. Existiert $\Lambda(h)$ aus (31) für jedes $h\in C(M,\mathbb{R})$, so erfüllt $(\mu_n)_{n\geqslant 1}$ ein LDP mit Ratenfunktion*

$$J(x):=\sup_{h\in C(M,\mathbb{R})}(h(x)-\Lambda(h)). \tag{33}$$

*(Es ist also $J(x)=\Lambda^*(\delta_x)$ im Sinne der Legendre-Fenchel Transformierten von $\Lambda$. Außerdem: Da $M$ kompakt ist, ist jede Ratenfunktion gut, also auch $J$.)*

*Beweis:* Als Supremum stetiger Funktionen ist $J$ unterhalbstetig, und da $\Lambda(0)=0$, ist $J\geqslant 0$. Also ist $J$ eine Ratenfunktion.

Sei zunächst $\delta>0$. Zu $x\in M$ gibt es ein $h_x\in C(M,\mathbb{R})$ mit

$$h_x(x)-\Lambda(h_x)>(J(x)-\delta)\wedge\delta^{-1},$$

und da die $h_x$ stetig sind, gibt es zu jedem $x$ eine Umgebung $U_x$, so dass

$$h_x(y)>\Lambda(h_x)+(J(x)-\delta)\wedge\delta^{-1}\text{ für alle }y\in U_x.$$

Aus der Markov-Ungleichung folgt

$$\begin{aligned}\mu_n(U_x) &\leqslant \int_{U_x}\exp(n(h_x-\inf\{h_x(y):y\in U_x\}))d\mu_n\\ &\leqslant \int\exp(n(h_x-(\Lambda(h_x)+(J(x)-\delta)\wedge\delta^{-1})))d\mu_n\\ &= \int e^{nh_x}d\mu_n\cdot e^{-n(\Lambda(h_x)+(J(x)-\delta)\wedge\delta^{-1})},\end{aligned}$$

so dass

$$\limsup_{n\to\infty}\frac{1}{n}\log\mu_n(U_x)\leqslant\Lambda(h_x)-(\Lambda(h_x)+(J(x)-\delta)\wedge\delta^{-1})=-((J(x)-\delta)\wedge\delta^{-1})$$

Sei nun $K\subseteq M$ abgeschlossen, also kompakt. Es gibt $x_1,\dots,x_m\in K$ so dass $K\subseteq\bigcup_{i=1}^m U_{x_i}$. Also:

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\mu_n(K) &\leqslant \max_{i=1,\dots,m}\limsup_{n\to\infty}\frac{1}{n}\log\mu_n(U_{x_i})\\ &\leqslant \max_{i=1,\dots,m}-((J(x_i)-\delta)\wedge\delta^{-1})\\ &\leqslant -\inf_{x\in K}(J(x)-\delta)\wedge\delta^{-1}.\end{aligned}$$

Im Limes $\delta \to 0$ folgt die obere LDP-Abschätzung.

Sei nun $G \subseteq M$ offen und $x \in G$. Wähle ein $f \in C(M, \mathbb{R})$ mit $-1 \leqslant f \leqslant 0$, $f(x) = 0$ und $f(y) = -1$ für $y \in M \setminus G$. (Das geht nach dem Satz von Urysohn.) Dann ist für jedes $\alpha > 0$

$$\begin{aligned} -J(x) &= \inf_{h \in C(M,\mathbb{R})} (\Lambda(h) - h(x)) \\ &\leqslant \Lambda(\alpha f) - \alpha f(x) = \Lambda(\alpha f) \\ &= \lim_{n\to\infty} \frac{1}{n} \log \int e^{n\alpha f} d\mu_n \\ &\leqslant \lim_{n\to\infty} \frac{1}{n} \log \left( \int_G e^0 d\mu_n + \int_{M\setminus G} e^{-n\alpha} d\mu_n \right) \\ &\leqslant \liminf_{n\to\infty} \frac{1}{n} \log(\mu_n(G) + e^{-n\alpha}) \\ &= \max \left\{ \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G), -\alpha \right\}. \end{aligned}$$

Für $\alpha > J(x)$, d.h. für $-J(x) > -\alpha$, folgt

$$-J(x) \leqslant \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G),$$

und da $x \in G$ beliebig gewählt war, folgt daraus die untere LDP-Abschätzung. □

**Korollar 11.4** *Sei $(\mu_n)_{n>0}$ eine Folge von W'verteilungen auf dem kompakten metrischen Raum $M$, die ein LDP mit Ratenfunktion $I$ erfüllt. Dann ist $I(x) = \Lambda^*(\delta_x) = \sup_{g\in C(M,\mathbb{R})}(g(x) - \Lambda(g))$, wo $\Lambda(g) = \sup_{y\in M}(g(y) - I(y))$.*

*Beweis:* Wegen des Lemmas von Varadhan kann man den Satz von Bryc anwenden, so dass $(\mu_n)_{n>0}$ auch ein LDP mit Ratenfunktion $J(x) = \Lambda^*(\delta_x)$ erfüllt. Das nachfolgende Lemma zeigt, dass dann $I = J$ sein muss. □

**Lemma 11.1** *Sei $(\mu_n)_{n>0}$ eine Folge von W'verteilungen, die sowohl ein LDP mit Ratenfunktionen $I$ als auch eines mit Ratenfunktion $J$ erfüllt. Dann ist $I = J$.*

*Beweis:* Da sowohl $I$ als auch $J$ Ratenfunktionen für $(\mu_n)_n$ sind, ist für alle $x \in M$ und alle $\delta > 0$

$$-\inf_{U_\delta(x)} I \leqslant \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(U_\delta(x)) \leqslant \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(\overline{U_\delta(x)}) \leqslant -\inf_{\overline{U_\delta(x)}} J \leqslant -\inf_{U_{2\delta}(x)} J.$$

Da sowohl $I$ als auch $J$ unterhalb stetig sind, folgt aus Aufgabe 5.1(iii), dass für alle $x \in M$ gilt:

$$I(x) = \lim_{\delta\downarrow 0} \inf_{U_\delta(x)} I \geqslant \lim_{\delta\downarrow 0} \inf_{U_{2\delta}(x)} J = J(x).$$

Durch Vertauschung der Rollen von $I$ und $J$ erhält man die Gleicheit $I = J$. □

**Satz 11.5** *Sei $M$ ein kompakter metrischer Raum, $(\mu_n)_{n\geqslant 1}$ eine Folge von W'verteilungen auf $M$, die ein LDP mit Ratenfunktion $I_\mu$ erfüllt. Sei außerdem $f : M \to \mathbb{R}$ stetig, und seien $\nu_n$ W'verteilungen auf $M$ mit Dichten $\frac{d\nu_n}{d\mu_n} = e^{-\psi_n(f)+nf}$, wobei $\psi_n(f) := \log \int e^{nf} d\mu_n$ ist.*

*Dann erfüllen die $(\nu_n)_{n\geqslant 1}$ ein LDP mit Ratenfunktion*

$$I_\nu(x) := I_\mu(x) - (f(x) - \Lambda_\mu(f)) = \sup_{y\in M} (f(y) - I_\mu(y)) - (f(x) - I_\mu(x))$$

*Beweis:* Wir zeigen zunächst, dass $\Lambda_\nu(h) = \lim_{n\to\infty} \frac{1}{n} \log \int e^{nh} d\nu_n$ für jedes $h \in C(M, \mathbb{R})$ existiert und dass

$$\Lambda_\nu(h) = \Lambda_\mu(h+f) - \Lambda_\mu(f).$$

Aus dem Lemma von Varadhan, angewandt auf die Folge $(\mu_n)_{n\geqslant 1}$, folgt, dass die folgenden Limiten existieren:

$$\begin{aligned}
\Lambda_\mu(h+f) - \Lambda_\mu(f) &= \lim_{n\to\infty} \frac{1}{n} \log \int e^{n(h+f)} d\mu_n - \lim_{n\to\infty} \frac{1}{n} \log \underbrace{\int e^{nf} d\mu_n}_{e^{\psi_n(f)}} \\
&= \lim_{n\to\infty} \frac{1}{n} \log \int e^{n(h+f)-\psi_n(f)} d\mu_n \\
&= \lim_{n\to\infty} \frac{1}{n} \log \int e^{nh} d\nu_n \\
&= \Lambda_\nu(h)
\end{aligned}$$

Also erfüllt $(\nu_n)_{n\geqslant 1}$ nach dem Satz 11.3 von Bryc ein LDP mit Ratenfunktion

$$\begin{aligned}
I_\nu(x) &= \sup_{h\in C(M,\mathbb{R})} (h(x) - \Lambda_\nu(h)) \\
&= \sup_{h\in C(M,\mathbb{R})} (h(x) + \Lambda_\mu(f) - \Lambda_\mu(h+f)) \\
&= \sup_{g\in C(M,\mathbb{R})} (g(x) - f(x) + \Lambda_\mu(f) - \Lambda_\mu(g)) \\
&= \sup_{g\in C(M,\mathbb{R})} (g(x) - \Lambda_\mu(g)) - f(x) + \Lambda_\mu(f) \\
&= I_\mu(x) - f(x) + \Lambda_\mu(f) \\
&= \sup_{y\in M} (f(y) - I_\mu(y)) - (f(x) - I_\mu(x)),
\end{aligned}$$

wobei für die vorletzte Gleichheit der Satz von Bryc (angewandt auf die Folge $(\mu_n)_{n\geqslant 1}$) und Korollar 11.4 benutzt wurden und für die letzte das Lemma von Varadhan. □

# 12 Das Curie-Weiss-Modell

Die Arbeit, in der die großen Abweichungen für dieses Modell geklärt wurden, ist [21].

Das Curie-Weiss-Modell ist eigentlich ein Modell der elementaren diskreten Stochastik. Man betrachtet $N$ binäre Zufallsvariablen $X = (X_1, \ldots, X_N)$, die die Werte $+1$ und $-1$ annehmen können und untersucht die gemeinsame Verteilung $P_X = P_{X_1,\ldots,X_N}$ auf $\{-1,1\}^N$ unter speziellen Annahmen an die Abhängigkeit der $X_i$.

Man kann sich die $X_i$ im Sinne der statischen Mechanik als die Ausrichtungen von $N$ Elementarmagneten vorstellen, oder z.B. auch als die Meinungen oder Einstellungen (Zustimmung/Ablehnung) von $N$ Individuen zu einer politischen oder ökonomischen Frage. Modelliert werden soll, dass jedes Individuum seine Meinung an der ,,durchschnittlichen Meinung" $\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i$ aller Individuen ausrichtet und dass wegen äußerer Einflüsse eine gewisse Meinungstendenz vorherrscht. Daher betrachten wir die Observablen $T_i : \{-1,1\}^N \to \mathbb{R}$,

$$T_0(x) = N\bar{x} \text{ und } T_i(x) = x_i\bar{x}$$

und geben die Erwartungswerte

$$E[\bar{X}] = \gamma_0 \in (-1,1), \text{ also } \int T_0 dP_X = N\gamma_0, \text{ und } \int T_i dP_X = E[X_i\bar{X}] = \gamma_1 \geqslant 0 \ (i = 1,\ldots,d)$$

vor[1].

Die Verteilung, die unter diesen Nebenbedingungen die Entropie maximiert bzw. die relative Entropie zur Bernoulli-$(\frac{1}{2},\frac{1}{2})$-Verteilung $P_N$ (d.h. zur Gleichverteilung auf $\{-1,1\}^N$) minimiert hat zu $P_N$ eine Dichte

$$f_{\vartheta,N}(x) = e^{-\psi_N(\vartheta)} \exp\left(\vartheta_0 N\bar{x} + \sum_{i=1}^N \vartheta_i x_i \bar{x}\right).$$

Zur Vermeidung unnötiger Komplikationen nehmen wir an, dass $\Omega = \{-1,1\}^N$, $P = P_N$ und dass die $X_i : \Omega \to \{-1,1\}$ durch $X_i(x) = x_i$ definiert sind.

Ist nun $\Gamma(\vartheta) = D\psi_N(\vartheta) = (\gamma_0, \gamma_1, \ldots, \gamma_1) =: \gamma$, so kann man zeigen, dass aus der Symmetrie des Modells in den $x_1, \ldots, x_N$ und der Permutationsinvarianz von $\gamma$ in den letzten $N$ Variablen auch die Permutationsinvarianz von $\vartheta$ in den letzten $N$ Variablen folgt.[2] Deshalb ist $\vartheta$ von der Form $\vartheta = (\vartheta_0, \vartheta_1, \ldots, \vartheta_1)^T$ und $f_{\vartheta,N}$ hat die Form

$$f_{\vartheta,N}(x) = f_{(\vartheta_0,\vartheta_1),N}(x) = e^{-\psi_N(\vartheta)} \exp(N(\vartheta_0\bar{x} + \vartheta_1\bar{x}^2)) =: \tilde{f}_{\vartheta,N}(\bar{x}).$$

[1] Es muss gelten $\gamma_1 \geqslant 0$, da $\gamma_1 = N^{-1}\sum_{i=1}^N E[X_i\bar{X}] = E[\bar{X}^2]$.

[2] Sei $\sigma$ eine Permutation von $\{1,\ldots,N\}$. Definiere $\hat{\sigma} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}, \vartheta = (\vartheta_0,\ldots,\vartheta_N)^T \mapsto (\vartheta_0, \vartheta_{\sigma(1)},\ldots,\vartheta_{\sigma(N)})^T$. Dann ist $D\hat{\sigma}$ eine Permutationsmatrix, also $(D\hat{\sigma})^{-1} = (D\hat{\sigma})^T$ und es gilt $\hat{\sigma}(\vartheta) = D\hat{\sigma}\cdot\vartheta$. Für $x, \vartheta \in \mathbb{R}^{N+1}$ ist $\overline{\hat{\sigma}^{-1}(x)} = \overline{x}$, wobei $\overline{x} = N^{-1}\sum_{i=1}^N x_i$ , und

$$\langle\hat{\sigma}(\vartheta), x\rangle = \langle D\hat{\sigma}\cdot\vartheta, x\rangle = \langle\vartheta, (D\hat{\sigma})^{-1}x\rangle = \langle\vartheta, \hat{\sigma}^{-1}(x)\rangle,$$

so dass

$$\begin{aligned}
\psi_N(\hat{\sigma}(\vartheta)) &= 2^{-N} \sum_{x\in\{0\}\times\{-1,1\}^N} \exp(\vartheta_0 N\bar{x} + \langle\hat{\sigma}(\vartheta), x\rangle\cdot\bar{x}) \\
&= 2^{-N} \sum_{x\in\{0\}\times\{-1,1\}^N} \exp\left(\vartheta_0 N\overline{\hat{\sigma}^{-1}(x)} + \langle\vartheta, \hat{\sigma}^{-1}(x)\rangle\cdot\overline{\hat{\sigma}^{-1}(x)}\right) \\
&= 2^{-N} \sum_{x\in\{0\}\times\{-1,1\}^N} \exp(\vartheta_0 N\bar{x} + \langle\vartheta, x\rangle\cdot\bar{x}) \\
&= \psi_N(\vartheta).
\end{aligned}$$

Es folgt:

$$\Gamma^T\circ\hat{\sigma} = (D\psi_N\circ\hat{\sigma})^T = (D(\psi_N\circ\hat{\sigma})\cdot D\hat{\sigma}^{-1})^T = D\hat{\sigma}\cdot D\psi_N^T = \hat{\sigma}\circ\Gamma^T.$$

Ist nun $\Gamma(\vartheta) = (\gamma_0,\gamma_1,\ldots,\gamma_1) =: \gamma$, so folgt $\Gamma(\hat{\sigma}(\vartheta))^T = \hat{\sigma}(\Gamma(\vartheta)^T) = \hat{\sigma}(\gamma^T) = \gamma^T = \Gamma(\vartheta)^T$. Da $\Gamma$ ein Diffeomorphismus ist, folgt $\hat{\sigma}(\vartheta) = \vartheta$.

Wir bezeichnen $P_{\vartheta,N} = f_{\vartheta,N} P_N$.

Sei $\mu_N$ die Verteilung von $\bar{X}$ unter $P_N$ und $\nu_N$ die unter $P_{\vartheta,N}$. Dann sind $\mu_N$ und $\nu_N$ konzentriert auf dem kompakten Raum $M = [-1, 1]$, und es ist

$$\frac{d\nu_N}{d\mu_N}(v) = \tilde{f}_{\vartheta,N}(v) = \exp(-\psi_N(\vartheta) + N(\vartheta_0 v + \vartheta_1 v^2)),$$

denn für messbare $A \subseteq \mathbb{R}$ gilt:

$$\nu_N(A) = P_{\vartheta,N}\{\bar{X} \in A\} = \int_{\{\bar{x}\in A\}} \tilde{f}_{\vartheta,N}(\bar{x}) dP_N(x) = \int_{\{v\in A\}} \tilde{f}_{\vartheta,N}(v) d\mu_N(v).$$

Nach Beispiel 2.8 erfüllt die Folge $(\mu_N)_{N\geqslant 1}$ von skalierten Binomialverteilungen ein LDP mit guter Ratenfunktion

$$I(v) = D((\frac{1+v}{2}, \frac{1-v}{2}) \| \,\mathrm{GV}_2) = H(\frac{1}{2}) - H(\frac{1+v}{2}),$$

wo $H(x) = -x \log x - (1-x)\log(1-x)$.[3] Also erfüllt die Folge $(\nu_N)_{N\geqslant 1}$ nach Satz 11.5 ein LDP mit guter Ratenfunktion

$$I_\vartheta(v) \quad = \quad \sup_{-1\leqslant u\leqslant 1} \big(\vartheta_0 u + \vartheta_1 u^2 - I(u)\big) - \underbrace{\big(\vartheta_0 v + \vartheta_1 v^2 - I(v)\big)}_{=:G_\vartheta(v)}.$$

Man sieht, dass auf jeden Fall $\inf_{-1\leqslant v\leqslant 1} I_\vartheta(v) = 0$. (Betrachte $v = \arg\max G_\vartheta(u)$). Wir untersuchen, wie die Ratenfunktion $I_\vartheta$ vom Parameter $\vartheta$ abhängt:

Da $H'(x) = \log\frac{1-x}{x}$, ist $I'(v) = -\frac{1}{2}H'(\frac{1+v}{2}) = -\frac{1}{2}\log\frac{1-v}{1+v}$, also

$$I'_\vartheta(v) = -G'_\vartheta(v) = I'(v) - \vartheta_0 - 2\vartheta_1 v = -\frac{1}{2}\log\frac{1-v}{1+v} - \vartheta_0 - 2\vartheta_1 v$$

und $I''_\vartheta(v) = \frac{1}{1-v^2} - 2\vartheta_1$.

Insbesondere ist $I'_\vartheta(v) = 0 = G'_\vartheta(v)$ gdw. $\frac{1-v}{1+v} = e^{-2\vartheta_0 - 4\vartheta_1 v}$ gdw.

$$v = \frac{1 - e^{-2\vartheta_0 - 4\vartheta_1 v}}{1 + e^{-2\vartheta_0 - 4\vartheta_1 v}} = \frac{e^{\vartheta_0 + 2\vartheta_1 v} - e^{-(\vartheta_0 + 2\vartheta_1 v)}}{e^{\vartheta_0 + 2\vartheta_1 v} + e^{-(\vartheta_0 + 2\vartheta_1 v)}} = \tanh(\vartheta_0 + 2\vartheta_1 v) =: h(v),$$

d.h. wenn $v$ ein Fixpunkt der Abbildung $h$ ist.[4] Beachte, dass $h'(v) = \frac{2\vartheta_1}{\cosh(\vartheta_0 + 2\vartheta_1 v)^2} \leqslant 2\vartheta_1$.

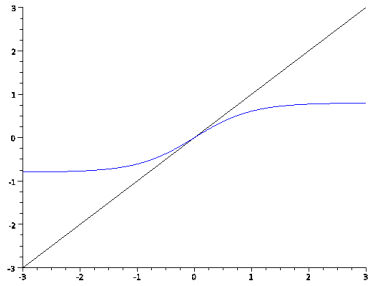

Abbildung 1: $\vartheta_0 = 0, h'(0) < 1$

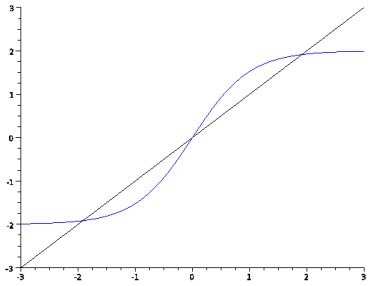

Abbildung 2: $\vartheta_0 = 0, h'(0) > 1$

[3] Formal folgt das aus dem Kontraktionsprinzip: Seien $Z_1, \ldots, Z_N$ unabhängig mit $P\{Z_i = 0\} = P\{Z_i = 1\} = \frac{1}{2}$. Dann erfüllt $(\bar{Z}_N)_{N>0}$ ein LDP mit guter Ratenfunktion $I_Z(u) = D((u, 1-u) \| \,\mathrm{GV}_2)$. Dann hat $X = (X_1, \ldots, X_N)$ mit $X_i = 2Z_i - 1$ die Gleichverteilung $P_N$ auf $\{-1, 1\}^N$, und es ist $\bar{X}_N = 2\bar{Z}_N - 1$. Aus dem Kontraktionsprinzip folgt also, dass $(\bar{X}_N)_{N>0}$ ein LDP mit guter Ratenfunktion $I(v) = \inf\{I_Z(u) : v = 2u - 1\} = I_Z\left(\frac{1+v}{2}\right) = D\left(\left(\frac{1+v}{2}, \frac{1-v}{2}\right) \| \,\mathrm{GV}_2\right) = H\left(\frac{1}{2}\right) - H\left(\frac{1+v}{2}\right)$ erfüllt.

[4] Zur Erinnerung: $\sinh(x) = \frac{e^x - e^{-x}}{2}$, $\cosh(x) = \frac{e^x + e^{-x}}{2} \geqslant 1$, $\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Ist $\vartheta_0 = 0$, so ist $h(0) = 0$, und ob weitere Fixpunkte von $h$ existieren, hängt von $\vartheta_1$ ab, siehe Abbildungen 1 und 2. In diesem Fall ist $I_\vartheta$ symmetrisch, d.h. es ist $I_\vartheta(-v) = I_\vartheta(v)$, und es sind die folgenden Fälle zu unterscheiden, die sich durch die Fixpunkteigenschaften der Abbildung $h$ charakterisieren lassen:

1. $\vartheta_1 \leqslant \frac{1}{2}$. Dann ist $h'(v) \leqslant 2\vartheta_1 \leqslant 1$ mit Gleichheit genau dann, wenn $v = 0$. Also ist $h(v) = v$ gdw. $v = 0$, und es ist $I''_\vartheta(v) \geqslant -1 + \frac{1}{1-v^2} \geqslant 0$ mit Gleichheit nur für $v = 0$, so dass $I_\vartheta$ bei $v = 0$ ein eindeutiges Minimum $I_\vartheta(0) = 0$ hat, siehe Abbildung 3.

2. $\vartheta_1 > \frac{1}{2}$. Dann ist $h'(0) = 2\vartheta_1 \cdot \tanh'(0) > 1$, und es ist nicht nur $h(0) = 0$, sondern es gibt genau zwei weitere Fixpunkte $v_- < 0 < v_+$. Offensichtlich ist $I''_\vartheta(0) < 0$, d.h. bei $v = 0$ hat $I_\vartheta$ ein lokales Maximum. Daher müssen bei $v = v_+, v_-$ nun lokale Minima vorliegen, und da $I_\vartheta$ symmetrisch ist, sind die beiden Minimalwerte $I_\vartheta(v_\pm)$ identisch gleich 0, siehe Abbildung 4.

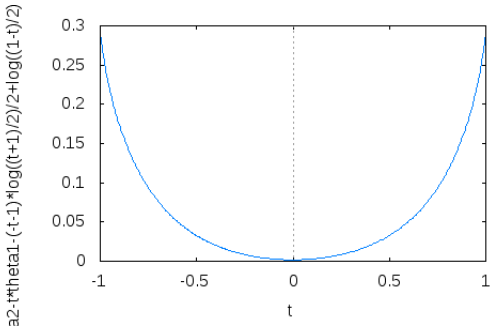


Abbildung 3: $I_\vartheta(t)$ für $\vartheta_0 = 0$, $\vartheta_1 = 0.4$

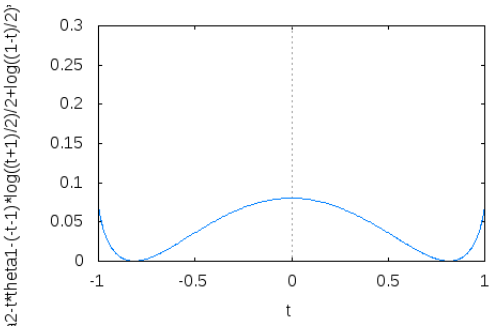


Abbildung 4: $I_\vartheta(t)$ für $\vartheta_0 = 0$, $\vartheta_1 = 0.7$

Zur korrekten Formulierung des folgenden Theorems setzen wir noch $I_\vartheta(x) := \infty$ für $x \not\in [-1, 1]$.

**Satz 12.1** *Sei $\gamma_0 = 0$. Dann ist auch $\vartheta_0 = 0$.*

*a) Ist $\vartheta_1 < \frac{1}{2}$, so ist für jedes $\alpha > 0$*

$$\lim_{N\to\infty} \tfrac{1}{N} \log P_{\vartheta,N}\{|\bar{X}| > \alpha\} \quad = \quad -I_\vartheta(\alpha)$$

*b) Ist $\vartheta_1 > \frac{1}{2}$, so ist für jedes $\alpha > 0$*

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\vartheta,N}\{\bar{X} > 0 \,\text{und}\, |\bar{X} - v_+| > \alpha\} = \begin{cases} -\min\{I_\vartheta(v_+ - \alpha), I_\vartheta(v_+ + \alpha)\} & \text{für} \quad \alpha < v_+ \\ -I_\vartheta(v_+ + \alpha) & \text{für} \quad \alpha \geqslant v_+ \end{cases}$$

*und eine entsprechende Gleichheit gilt für $P_{\vartheta,N}\{\bar{X} < 0 \,\text{und}\, |\bar{X} - v_-| > \alpha\}$.*

*Beweis:* Ist $\vartheta_0 = 0$, so ist $f_{\vartheta,N}(-x) = f_{\vartheta,N}(x)$, so dass $\gamma_0 = E_{P_{\vartheta,N}}[\bar{X}] = E_{P_{\vartheta,N}}[-\bar{X}] = -\gamma_0 = 0$. Tatsächlich ist $\gamma_0 = 0$ genau dann wenn $\vartheta_0 = 0$, weil $\frac{\partial}{\partial\vartheta_0} E_{P_{\vartheta,N}}[\bar{X}] > 0$. Das folgt so:

$$\frac{\partial}{\partial\vartheta_0} E_{P_{\vartheta,N}}[\bar{X}] \quad = \quad N^{-1} \frac{\partial}{\partial\vartheta_0} E_{P_{\vartheta,N}}[T_0] \quad = \quad N^{-1} \frac{\partial^2}{\partial\vartheta_0^2} \psi_N(\vartheta) = N^{-1} \operatorname{Var}_\vartheta[T_0] \quad > \quad 0,$$

da $T_0(x_1, \ldots, x_N) = x_1 + \ldots + x_N$ auf $\{-1, 1\}^N$ nicht konstant ist. Der Rest folgt aus dem LDP und den Skizzen von $I_\vartheta$. □

Ist $\vartheta_0 \neq 0$, so liegt die in Abbildungen 5 und 6 skizzierte Situation vor: $I_\vartheta$ hat ein eindeutiges Minimum, das dasselbe Vorzeichen wie $\vartheta_0$ hat. Für große $\vartheta_1$ kann ein weiteres lokales Minimum

vorliegen, siehe Abbildung 6. In der dort skizzierten Situation bezeichne $v_+ > 0$ das globale Minimum von $I_\vartheta$ und $v_- < 0$ das lokale. Dann ist

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\vartheta,N}\{|\bar{X} - v_+| > \alpha\} = \begin{cases} \min\{I_\vartheta(v_+ - \alpha), I_\vartheta(v_+ + \alpha), I_\vartheta(v_-)\} & \text{für} \alpha \leqslant v_+ - v_- \\ \min\{I_\vartheta(v_+ - \alpha), I_\vartheta(v_+ + \alpha)\} & \text{für} \alpha > v_+ - v_- \end{cases} .$$

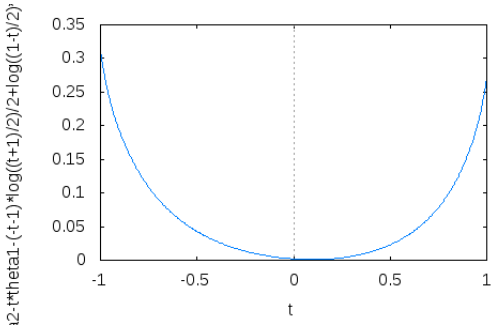


Abbildung 5: $I_\vartheta(t)$ für $\vartheta_0 = 0.1$, $\vartheta_1 = 0.4$

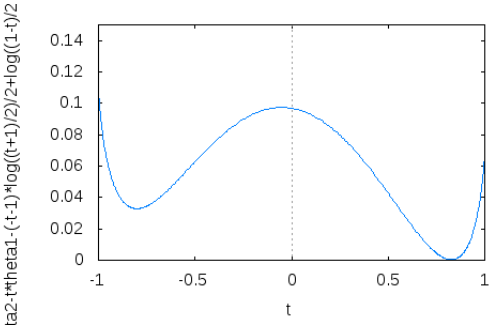


Abbildung 6: $I_\vartheta(t)$ für $\vartheta_0 = 0.1$, $\vartheta_1 = 0.7$

# 13 Große Abweichungen in dynamischen Systemen

In diesem Abschnitt stelle ich eine vereinfachte Version des Zugangs von Wentzell und Freidlin zu großen Abweichungen in dynamischen Systemen vor.

**Beispiel 13.1** Um technische Probleme so gering wie möglich zu halten, betrachten wir zunächst nur dynamische Systeme, die durch eine Abbildung $f : \mathbb{R}^d \to \mathbb{R}^d$ beschrieben werden, d.h. Dynamiken, die ausgehend von einem Anfangszustand $x_0 \in \mathbb{R}^d$ während des (diskreten) Zeitintervalls $\{0, \dots, T\}$ die Folge von Zuständen $x_1 = f(x_0), x_2 = f(x_1), \dots, x_T = f(x_{T-1})$ durchlaufen. Dabei ist $T$ groß aber endlich.

Nun fügen wir zufällige Störungen zu einem solchen dynamischen System hinzu: Seien $\xi_1, \dots, \xi_T$ u.i.v. mit Erwartungswert 0 und Varianz 1. Für $\sigma > 0$ definieren wir den Prozess $(X_t^\sigma)_{t=0,\dots,T}$ durch

$$X_0^\sigma = x_0, X_t^\sigma = f(X_{t-1}^\sigma) + \sigma\xi_t \quad (t = 1, \dots, T). \tag{34}$$

Die Frage lautet:

> Wie weit entfernt sich die zufällige Trajektorie $(X_1^\sigma, \dots, X_T^\sigma)$ von der ungestörten $(x_1, \dots, x_T)$? Insbesondere: wie groß/klein ist die Wahrscheinlichkeit, dass die zufällige Trajektorie ein qualitativ gänzlich anderes Verhalten zeigt als die ungestörte, also z.B. nur negative Werte annimmt, während die ungestörte gegen einen positiven Fixpunkt konvergiert?

Solche Fragen können durch ein LDP im Limes $\sigma \to 0$ beantwortet werden. Wir beschränken uns auf den Fall, dass die $\xi_i$ nach $\mathcal{N}(0, 1)$ verteilt sind, also insbesondere auf $d = 1$.

Sei $\xi = (\xi_1, \dots, \xi_T)$. Die Familie der Verteilungen $(P_{\sigma\xi})_{\sigma>0}$ erfüllt im Limes $\sigma \to 0$ ein volles LDP *mit Skala* $\sigma^{-2}$ und guter Ratenfunktion $I(z) = \frac{1}{2}\|z\|^2$:

$$\begin{aligned}
\liminf_{\sigma\to 0} \sigma^2 \cdot \log P\{\sigma\xi \in G\} &\geqslant -\inf_{z\in G} \frac{1}{2}\|z\|^2 \text{für alle offenen Mengen } G \subseteq \mathbb{R}^T, \\
\limsup_{\sigma\to 0} \sigma^2 \cdot \log P\{\sigma\xi \in A\} &\leqslant -\inf_{z\in A} \frac{1}{2}\|z\|^2 \text{für alle abgeschlossenen Mengen } A \subseteq \mathbb{R}^T.
\end{aligned}$$

Das wurde in Problem 6.2 gezeigt.[5] Nun wird (34) für gegebenes $x_0 \in \mathbb{R}$ durch eine stetige Abbildung $F_{x_0} : \mathbb{R}^T \to \mathbb{R}^T$ beschrieben, die jedem $(\sigma\xi_1, \dots, \sigma\xi_T)$ ein $(X_1^\sigma, \dots, X_T^\sigma)$ zuordnet:

$$F_{x_0}(z_1, \dots, z_T) = (x_1, \dots, x_T) \text{ mit } x_t = f(x_{t-1}) + z_t \text{ für } t = 1, \dots, T.$$

Nach dem Kontraktionsprinzip (genauer: einer Variante für die hiesige Situation mit $\sigma \to 0$) erfüllt deshalb die Familie $(\mu_\sigma)_{\sigma>0}$ von Verteilungen der $(X_1^\sigma, \dots, X_T^\sigma) = F_{x_0}(\sigma\xi_1, \dots, \sigma\xi_T)$, bei gegebenem $x_0$, ein volles LDP mit Skala $\sigma^{-2}$ und mit guter Ratenfunktion

$$J(x) = J(x_1, \dots, x_T) = \inf\left\{\frac{1}{2}\|z\|^2 : F_{x_0}(z) = x\right\} = \inf\left\{\frac{1}{2}\sum_{t=1}^T z_t^2 : F_{x_0}(z) = x\right\} = \frac{1}{2}\sum_{t=1}^T (x_t - f(x_{t-1}))^2.$$

Die Auswertung von Ausdrücken der Form $\inf_{x\in V} J(x)$, wie sie für LDP-Abschätzungen nötig ist, kann ein sehr schwieriges Optimierungsproblem sein. Bevor wir ein einfaches Beispielsystem im Detail diskutieren, betrachten wir zunächst eine Variante der allgemeinen Vorgehensweise für Systeme mit stetiger Zeit:

---

[5] Es ist $\psi_{\sigma\xi_i}(t) = \log\left(\frac{1}{(2\pi\sigma^2)^{1/2}} \int \exp(tx - \frac{x^2}{2\sigma^2})dx\right) = \log \frac{\sigma^2 t^2}{2}$, also $\psi_{\sigma\xi}(\vartheta) = \frac{\sigma^2\vartheta_1^2}{2} + \dots + \frac{\sigma^2\vartheta_T^2}{2} = \frac{\sigma^2}{2}\|\vartheta\|^2$. Daher ist $\psi^*_{\sigma\xi}(z) = \sup_{\vartheta\in\mathbb{R}^T}(\langle\vartheta, z\rangle - \frac{\sigma^2}{2}\|\vartheta\|^2) = \left\langle \frac{z}{\sigma^2}, z\right\rangle - \frac{\sigma^2}{2}\|\frac{z}{\sigma^2}\|^2 = \frac{1}{2\sigma^2}\|z\|^2$, und man überzeugt sich leicht, dass $\lim_{\sigma\to 0} \sigma^2\psi^*_{\sigma\xi}(z) = \frac{1}{2}\|z\|^2$ die richtige Ratenfunktion ist.

**Beispiel 13.2** Ähnliche Überlegungen kann man für dynamische Systeme in stetiger Zeit anstellen. Das ungestörte System wird durch eine $\mathbb{R}^d$-wertige gewöhnliche Differentialgleichung

$$\dot{x}(t) \quad = \quad b(x(t)), \quad x(0) = x_0, \qquad \text{im Intervall } 0 \leqslant t \leqslant 1$$

gegeben, das stochastisch gestörte entsprechend durch

$$dX_t^{(\sigma)} = b(X_t^{(\sigma)})dt + \sigma dW_t, \quad X_0^{(\sigma)} = x_0$$

wobei $(W_t)_{0\leqslant t\leqslant 1}$ eine $\mathbb{R}^d$-wertige Brownsche Bewegung ist. (Für $\sigma = 0$ ist das also nichts anderes als die ungestörte gewöhnliche Differentialgleichung.) Diese einfache stochastische Differentialgleichung ist äquivalent zu folgender stochastischen Integralgleichung

$$X_t^{(\sigma)} = x_0 + \int_0^t b(X^{(\sigma)}(s))ds + \sigma W_t.$$

In Beispiel 3.1.7 des Skripts von König [19] wird diese Situation beschrieben, und es wird folgender Satz über große Abweichungen aus einem LDP für die Brownsche Bewegung (Satz von Schilder) und dem Kontraktionsprinzip hergeleitet. Hier ist zunächst der Satz von Schilder, siehe auch [24, Satz 4.9]:

**Satz 13.3** *Die Familie $(P_{\sigma W})_{\sigma>0}$ von Verteilungen der skalierten Brownschen Bewegungen auf $[0,1]$ erfüllt ein volles LDP mit Skala $\sigma^{-2}$ und guter Ratenfunktion $I : C([0,1],\mathbb{R}) \to [0,+\infty]$,*

$$I(\psi) \quad = \quad \begin{cases} \frac{1}{2}\int_0^1 |\psi'(t)|^2 dt & \text{für absolut stetige } \psi \text{ mit } \psi(0) = 0 \\ +\infty & \text{sonst.} \end{cases}$$

Dabei heißt $\psi : [0,1] \to \mathbb{R}$ *absolut stetig*, falls für jede Zahl $\epsilon > 0$ eine Zahl $\delta > 0$ existiert, so dass für jede endliche oder unendliche Folge paarweisse disjunkter Intervalle $[x_k, y_k] \subseteq [0,1]$, die der Bedingung $\sum_k (y_k - x_k) < \delta$ genügen, gilt: $\sum_k |\psi(y_k) - \psi(x_k)| < \epsilon$. Jede absolut stetige Funktion ist gleichmäßig stetig. Andererseits ist jede Lipschitz-stetige Funktion auch absolut stetig. Die Cantor-Funktion („Teufelstreppe“) ist ein Beispiel für eine überall stetige, aber nicht absolut stetige Funktion. Absolut stetige Funktionen sind fast überall differenzierbar und diese Ableitung stimmt mit der schwachen Ableitung überein, d.h. für jede stetig differenzierbare Testfunktion $f : [0,1] \to \mathbb{R}$ mit $f(0) = f(1) = 0$ gilt: $\int_0^1 \psi'(x) f(x) dx = -\int_0^1 \psi(x) f'(x) dx$.

Daraus folgert man z.B. den folgenden Satz (siehe auch [24, Satz 6.7]):

**Satz 13.4** *Die Familie $(P_{X^{(\sigma)}})_{\sigma>0}$ von Verteilungen der Prozesse $(X_t^{(\sigma)})_{0\leqslant t\leqslant 1}$ erfüllt ein volles LDP mit Skala $\sigma^{-2}$ und guter Ratenfunktion $J : C([0,1],\mathbb{R}) \to [0,+\infty]$,*

$$J(\psi) \quad = \quad \begin{cases} \frac{1}{2}\int_0^1 |\psi'(t) - b(\psi(t))|^2 dt & \text{für absolut stetige } \psi \text{ mit } \psi(0) = x_0 \\ +\infty & \text{sonst.} \end{cases}$$

*Beweis:* (Herleitung aus Satz 13.3) Nach dem Satz von Picard-Lindelöf (bekannt aus dem Modul Gewöhnliche Differentialgleichungen) hat die Integralgleichung

$$\psi(t) \;=\; x_0 + \int_0^t b(\psi(s))ds + \varphi(t)\,, \qquad t \in [0,1]$$

für jedes $\varphi \in C([0,1],\mathbb{R})$ eine eindeutige Lösung $\psi = F(\varphi) \in C([0,1],\mathbb{R})$. Die Abbildung $F : C([0,1],\mathbb{R}) \to C([0,1],\mathbb{R})$ ist stetig. Das kann man direkt aus dem Beweis des Satzes von Picard-Lindelöf folgern, denn der beruht auf dem Banachschen Fixpunktsatz (Analysis II), und der dadurch bestimmte Fixpunkt $\psi$ hängt stetig von den „Zutaten“ der Integralgleichung ab. Alternativ kann man, wie im Winter-Skript oder im König-Skript auch mit dem Lemma von Gronwal argumentieren.

Insbesondere ist daher $X^{(\sigma)}(\omega) = F(\sigma W(\omega))$ für jeden Brownschen Pfad $W(\omega)$, und da der Satz von Schilder ein LDP mit guter Ratenfunktion für die Familie $(P_{\sigma W})_{\sigma>0}$ bereitstellt, folgt aus dem Kontraktionsprinzip ein LDP für die Familie $(P_{X^{(\sigma)}})_{\sigma>0}$ mit guter Ratenfunktion

$$\begin{aligned} J(\psi) &= \inf\{I(\varphi) : F(\varphi) = \psi\} = I\left(\psi - x_0 - \int_0^{\cdot} b(\psi(s))ds\right) \\ &= \begin{cases} \frac{1}{2}\int_0^1 |\psi'(t) - b(\psi(t))|^2 dt & \text{für absolut stetige } \psi \text{ mit } \psi(0) = x_0 \\ +\infty & \text{sonst.} \end{cases} \end{aligned}$$

□

Nun wenden wir uns wieder Systemen in diskreter Zeit zu und untersuchen eine sehr einfache Beispielfamilie genauer.

**Beispiel 13.5** $[d = 1,\ f(x) = mx + b,\ m > 0,\ m \neq 1]$ Dieses $f$ hat den Fixpunkt $x_* = \frac{b}{1-m}$, und da $x_t - x_* = f(x_{t-1}) - f(x_*) = m(x_{t-1} - x_*)$, ist $x_t - x_* = m^t \cdot (x_0 - x_*)$. Also:

- Für $m < 1$ nähern sich die $x_t$ exponentiell schnell dem Fixpunkt $x_*$,
- für $m > 1$ entfernen sich die $x_t$ exponentiell schnell von $x_*$.

Wir betrachten nur den Fall $b = 0$, den allgemeinen Fall kann man leicht darauf zurückführen. Dann ist $J(x) = \frac{1}{2}\sum_{t=1}^{T}(x_t - mx_{t-1})^2$. Will man $J(x)$ z.B. für gebenes $x_T$ minimieren, so betrachtet man

$$D_{x_1,\ldots,x_{T-1}} J(x) = (x_1 - mx_0 - m(x_2 - mx_1), \ldots, x_{T-1} - mx_{T-2} - m(x_T - mx_{T-1})).$$

$DJ(x) = 0$ ist daher äquivalent zu

$$(1 + m^2)x_t = m(x_{t-1} + x_{t+1}) \qquad \text{für } t = 1, \ldots, T-1.$$

Das ist eine lineare Differenzengleichung mit Randbedingungen. Ihr charakteristisches Polynom ist

$$m\lambda^2 - (1 + m^2)\lambda + m = m(\lambda - m)(\lambda - m^{-1}),$$

und man erhält als Lösung

$$x_t^* = A \cdot m^{-t} - B \cdot m^t \ (t = 0, \ldots, T) \ \text{ mit } \ A = \frac{x_0 m^T - x_T}{m^T - m^{-T}} \ \text{ und } \ B = \frac{x_0 m^{-T} - x_T}{m^T - m^{-T}}. \tag{35}$$

Die Hessematrix $HJ$ von $J$ ist eine Tridiagonalmatrix mit Diagonaleinträgen $1 + m^2$ und Nebendiagonaleinträgen $-m$. Da $1 + m^2 - |m| - |m| = (1 - m)^2 > 0$ für $m \neq 1$, folgt aus dem Satz von Gershgorin[6], dass $HJ$ positiv definit ist und damit (35) ein Minimum ist. Für $m < 1$ und großes $T$ ist $A \approx m^T x_T$ und $B \approx -x_0$, so dass $x_t^* \approx x_T m^{T-t} + x_0 m^t$. Dadurch wird derjenige Pfad charakterisiert, entlang dem es am wahrscheinlichsten ist, dass man von $x_0$ nach $x_T$ gelangt, wenn man überhaupt (mit oft nur kleiner Wahrscheinlichkeit) dorthin kommt.[7] Für $m = 0.9$, $x_0 = 0$

[6] Der Satz besagt, dass für jede komplexe $n \times n$-Matrix $A$ gilt: $\mathrm{EW}(A) \subset \bigcup_{i=1}^n B_{r_i}(a_{ii})$ mit $r_i = \sum_{j=1, j\neq i}^n |a_{ij}|$. Im obigen Fall folgt daraus $\mathrm{EW}(A) \subset B_{2m}(1 + m^2) \subset \{z \in \mathbb{C} : \mathrm{Re}(z) > 0\}$.

[7] Als Formel sieht diese Aussage folgendermaßen aus: Für jedes $\delta > 0$ und jedes $x \in \mathbb{R}^T$ ist

$$\begin{aligned} & \lim_{\sigma\to 0} \sigma^2 \log P(|X_t^\sigma - x_t| < \delta \ (t = 1, \ldots, T) \mid |X_T^\sigma - x_T| < \delta) \\ = \ & \lim_{\sigma\to 0} \sigma^2 \log \frac{P(|X_t^\sigma - x_t| < \delta \ (t = 1, \ldots, T))}{P(|X_T^\sigma - x_T| < \delta)} \\ = \ & -(\inf\{J(y) : |y_t - x_t| < \delta \ (t = 1, \ldots, T)\} - \inf\{J(y) : |y_T - x_T| < \delta\}), \end{aligned}$$

und im Limes $\delta \to 0$ ist erhält man

$$-\left(J(x) - \inf\{J(y) : y_T = x_T\}\right) = -\left(J(x) - J(x^*)\right).$$

Das ist $= 0$ genau dann, wenn $x = x^*$.

und $x_T = 1$ ist dieser Pfad in Abb. 7 dargestellt.

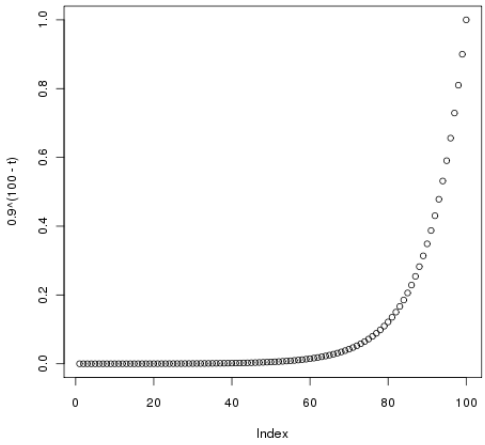


Abbildung 7: $T = 100$, $m = 0.9$, $x_0 = 0$, $x_T = 1$

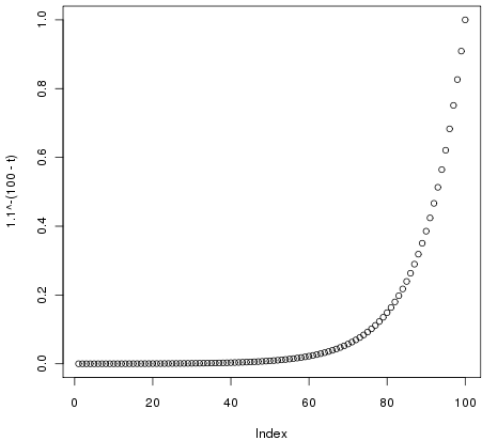


Abbildung 8: $T = 100$, $m = 1.1$, $x_0 = 0$, $x_T = 1$

Der wahrscheinlichste aller Pfade, die einen bei kleinem $\sigma$ von 0 nach 1 führen (was an sich ja unwahrscheinlich ist), bleibt also lange bei 0 und geht erst gegen Ende nach 1. Ähnliche Überlegungen kann man für $m > 1$ anstellen. Dann ist $x_t^* \approx x_0 m^{-t} + x_T m^{-(T-t)}$, siehe Abb. 8.

Wir bestimmen noch das Infimum $J(x^*)$ aller $J(x) = \frac{1}{2}\sum_{t=1}^{T}(x_t - m x_{t-1})^2$ bei gegebenen $x_0$ und $x_T$:

$$x_t^* - m x_{t-1}^* = A(m^{-t} - m^{-t+2}) - B(m^t - m^t) = A(1-m^2)m^{-t},$$

so dass

$$\begin{aligned}\inf J(x) &= \frac{1}{2}A^2(1-m^2)^2\sum_{t=1}^{T} m^{-2t} = \frac{1}{2}\left(\frac{x_0 m^T - x_T}{m^T - m^{-T}}\right)^2 (1-m^2)^2 m^{-2}\frac{m^{-2T}-1}{m^{-2}-1} \\ &= \frac{1}{2}\left(\frac{x_0 m^T - x_T}{m^T - m^{-T}}\right)^2 (1-m^2)(m^{-2T}-1).\end{aligned}$$

Im Limes $T \to \infty$ erhält man bei festgehaltenem $x_T \neq 0$:

$$m < 1 : \inf J(x) \approx \frac{1}{2}x_T^2(1-m^2)$$

$$m > 1 : \inf J(x) \approx \frac{1}{2}x_0^2(m^2-1).$$

Für $m = 1 + \delta$, $|\delta| \to 0$, erhält man $\inf J(x) \approx |\delta| x_T^2$ bzw. $\approx |\delta| x_0^2$.

Im Fall $x_T = 0$ und $x_0 \neq 0$ erhält man für $m > 1$ das selbe Ergebnis, für $m < 1$ aber:

$$\inf J(x) \approx \frac{1}{2}x_0^2(1-m^2)\cdot m^{2T}$$

Das ist zu erwarten, da in diesem Fall auch die deterministische Dynamik nach langer Zeit von $x_0$ schon fast nach $x_T = 0$ führt.