Große Abweichungen

Matthias Löwe

1 Einleitung

Die Theorie der großen Abweichungen ist ein Zweig der Wahrscheinlichkeitstheorie. Ihr Ausgangspunkt sind Konvergenzsätze wie etwa das Gesetz der großen Zahlen, der Satz von Glivenko-Cantelli oder die Ergodensätze. Diese besagen ganz grob gesprochen, dass Zufallsvariable im Schnitt über große Populationen annähernd konstant gleich ihrem Erwartungswert sind und dass große Fluktuationen um dieses typische Verhalten unwahrscheinlich sind. Die Kunst der Theorie der großen Abweichungen ist es, die Wahrscheinlichkeit solch untypischen Verhaltens zu qualifizieren.

Die Wurzeln für diese Theorie sind vielfältig:

1. Schon Boltzmanns Untersuchungen zur statistischen Mechanik, die die Wahrscheinlichkeit, ein großes System von Teilchen in einem untypischen Zustand zu finden, ins Verhältnis zu seiner Entropie setzen, und die in dem Boltzmannschen Gesetz
$$S = k \log W$$
(wobei S die Entropie und W die Wahrscheinlichkeit des Systems ist) gipfeln, können als ein erstes Prinzip der großen Abweichungen betrachtet werden.

2. In der Statistik ist es naheliegend, den empirischen Mittelwert $\frac{1}{n}\sum_{i=1}^{n} X_i$ als guten Schätzer für den Erwartungswert $\mathbb{E}X_1$, einer Zufallsvariablen X_1 zu nehmen (dabei seien $X_2, X_3, \ldots$ i.i.d. Kopien von X_1); ebenso lässt sich das empirische Maß $L_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ (S das Dirac-Maß) als guter Schätzer für die Verteilung einer Zufallsvariablen verstehen. Will man diese Schätzer tatsächlich benutzen, so stellt sich die Frage, wie schnell diese gegen den wahren Wert konvergieren, d. h. wie groß die Wahrscheinlichkeit eines abweichenden Verhaltens ist. Dies war die Fragestellung von Cramér und Sanov in den 1930er bzw. 1950er Jahren.

3. Varadhan stieß bei der Analyse der Lösung gewisser partieller Differentialgleichungen auf Integrale der Form
$$\int e^{nF(x)} d\mathbb{P}_n(x),$$
wobei die Maße $\mathbb{P}_n$ sich ebenfalls auf einer e^{-n}-Skala konzentrieren. Diese erinnern an Integrale deren asymptotisches Verhalten schon Laplace studierte (er zeigte, dass für die Integrale der Form $\int e^{nF(x)} dx$ im Wesentlichen der Maximalwert zählt). Die Frage ist, wie sich diese auswerten lassen.

Heutzutage werden Prinzipien großer Abweichungen als eine natürliche Analyse des Fluktuationsverhaltens von Zufallsvariablen betrachtet, vergleichbar etwa mit dem Zentralen Grenzwertsatz.

Wir wollen in den vorliegenden Notizen zunächst einige prinzipielle große Abweichungsresultate sammeln. Einige theoretische Resultate erlauben es dann, diese Resultate auf andere Situationen zu übertragen. Diese sollen vorgestellt werden.

Ein besonderes Augenmerk liegt auf Anwendungen, in denen eine Theorie der großen Abweichungen einen entscheidenden Beitrag zum Verständnis der Modelle liefert.

Die Theorie der großen Abweichungen ist im Laufe der vergangenen 40 Jahre so umfangreich geworden, dass es unmöglich ist, im Rahmen eines Kurses einen umfassenden Überblick über diese Theorie zu geben. Interessierte Leser seien auf die Bücher von Dembo und Zeitouni [DZ], den Hollander [dH] oder auf die ältere Übersichtsarbeit von Varadhan [V] für eine tiefergehende Lektüre verwiesen.

2 Der Satz von Cramér

Bereits in den Vorlesungen über Wahrscheinlichkeitstheorie haben wir ein wichtiges Prinzip der großen Abweichungen kennengelernt. Um dies in den formalen Rahmen der Vorlesung einzubetten, definieren wir zunächst, was wir unter dem Prinzip der großen Abweichungen verstehen wollen.

Definition 2.1 *Es sei $(X, \mathcal{X})$ ein metrischer Raum mit seiner Borelschen σ-Algebra. Wir sagen, dass eine Folge von Wahrscheinlichkeitsmaßen $(\mathbb{P}_n)_n$ auf $(X, \mathcal{X})$ einem Prinzip der großen Abweichungen (kurz: LDP) mit Ratenfunktion $I : X \to [0, \infty]$ und Geschwindigkeit a_n genügt, falls*

1. *Für jedes $L < \infty$ die Niveaumengen*

$$N_L = \{x : I(x) \leq L\}$$

   kompakt sind.

2. *Für alle offenen Mengen $G \subseteq X$ gilt*

$$\liminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}_n(G) \geq - \inf_{x\in G} I(x).$$

3. *Für alle abgeschlossenen Mengen $A \subseteq X$ gilt*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}_n(A) \leq - \inf_{x\in A} I(x).$$

Wir sagen, eine Folge $(X_n)_n$ von Zufallsvariablen genügt einem LDP, falls $(\mathbb{P}^{X_n})$ dies tut.

Bemerkung 2.2 *Oftmals unterscheidet man in der Literatur zwischen guten Ratenfunktionen, für die N_L wie in Definition 2.1 für alle $L > 0$ kompakt ist, und gewöhnlichen Ratenfunktionen, für die N_L nur abgeschlossen ist (was der unteren Halbstetigkeit von I entspricht).*

Nun kommen wir zu einem ersten LDP, das wir schon in der WT II kennengelernt haben.

Theorem 2.3 *Es sei (X_i) eine Folge von i.i.d. Zufallsvariablen , die*

$$\varphi(t) := \mathbb{E}e^{tX_1} < \infty \quad \forall\, t$$

erfüllt und die zudem nicht Dirac-verteilt seien. Sei $S_n := \sum_{i=1}^n X_i$. Dann gilt für alle $a > \mathbb{E}X_1$ die folgende Gleichheit:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na) = -I(a), \tag{1}$$

wobei

$$I(a) := \sup_{t\in\mathbb{R}} [ta - \psi(t)] \tag{2}$$

und

$$\psi(t) = \log \varphi(t) \tag{3}$$

gilt.

Beweis: Ohne Beschränkung der Allgemeinheit nehmen wir an, dass $a = 0$ und $\mathbb{E}X_1 < 0$ gilt (substituiert man nämlich $X_1 \to X_1 + a$, so ersetzt man auch $\varphi(t)$ durch $e^{at}\varphi(t)$. Mit $I(\cdot)$ – definiert wie in (2) – verschiebt sich dann auch $I(a)$ zu $I(0)$). Wir schreiben in der Folge

$$g := \inf_{t\in\mathbb{R}} \varphi(t)$$

und bemerken, dass

$$I(0) = -\log g \quad \text{mit } I(0) = \infty \quad \text{falls } g = 0$$

gilt.

Nun folgt mithilfe der exponentiellen Chebyschev–Ungleichung für alle positiven t

$$\mathbb{P}(S_n \geq na) \leq e^{-n(ta-\psi(t))} \tag{4}$$

und somit

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na) \leq -\sup_{t\in\mathbb{R}^+} [ta - \psi(t)]\,. \tag{5}$$

Um das Supremum über die ganze reelle Achse auszudehnen, überlegen wir, dass φ eine strikt konvexe Funktion ist. Es ist offenbar $\varphi'(0) = \mathbb{E}X_1 < 0$ (nach Annahme). Wir unterscheiden drei Fälle, je nachdem, wo $\mathbb{P}$ seine Masse hat.

- $\mathbb{P}(X_1 < 0) = 1$.
  Dann ist $\varphi'(t) = \int x e^{tx} d\mathbb{P}^X(x) < 0$ für alle $t \in \mathbb{R}$. Somit ist φ strikt fallend. Es ist somit
  $$g = \lim_{t\to\infty} \varphi(t) = \mathbb{P}(X_1 = 0) = 0.$$
  Da auch
  $$\mathbb{P}(S_n \geq 0) = 0$$
  gilt, haben wir in diesem Fall schon (1).

- $\mathbb{P}(X_1 \leq 0) = 1$ und $1 \neq \mathbb{P}(X_1 = 0) > 0$.
  Wie oben zeigt man, dass φ strikt fallend ist und

$$\lim_{t\to\infty} \varphi(t) = g = \mathbb{P}(X_1 = 0) > 0.$$

  Da in diesem Falle

$$\mathbb{P}(S_n \geq 0) = \mathbb{P}(X_1 = \ldots = X_n = 0) = g^n$$

  gilt, folgt auch hier (1).

- $\mathbb{P}(X_1 < 0) > 1$ und $\mathbb{P}(X_1 > 0) > 0$.
  Dann gilt offenbar $\lim_{t\to\pm\infty} \varphi(t) = \infty$ und da φ wie oben bemerkt strikt konvex ist, gibt es ein eindeutiges τ, so dass φ in τ minimal wird. Für diese τ gilt natürlich $\varphi'(\tau) = 0$ und $\tau > 0$, denn die Ableitung von φ ist in 0 negativ. Somit gehört τ zu den in (4) zulässigen t und es gilt daher

$$\mathbb{P}(S_n \geq 0) \leq \mathbb{E}e^{\tau S_n} = (\varphi(\tau))^n = g^n,$$

  also

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) \leq \log g.$$

  Um zu zeigen, dass $\log g$ auch eine untere Schranke ist, verwenden wir eine Technik, die als Tilten oder exponentielle Maßtransformation bekannt ist. Die Idee hierbei ist es, die zugrunde liegende Verteilung der X_i so zu verschieben, dass der Erwartungswert 0 (also unser a) ist. Dann wissen wir aus den Gesetzen der großen Zahlen, dass sich S_n so wie na verhalten wird. Wir kassieren aber einen "Strafterm" dafür, dass wir die Verteilung geändert haben.

  Genauer führen wir eine neue Folge (Y_i) von i.i.d. Zufallsvariablen ein, die gemäß Q verteilt sind, wobei

$$\frac{dQ}{d\mathbb{P}}(x) = \frac{1}{g} e^{\tau x}$$

  besitzen. Q heißt auch die Cramér-Transformierte von $\mathbb{P}$. Bemerke, dass

$$g = \varphi(\tau) = \int_{-\infty}^{\infty} e^{\tau y} d\mathbb{P}^X(y).$$

  Wir benötigen nun die folgenden drei Lemmata.

Lemma 2.4 *Es gilt* $\mathbb{E}Y = 0$ *und* $\mathbb{V}Y \in (0, \infty)$.

Beweis: Wir bezeichnen mit $\hat{\varphi}(t) = \mathbb{E}e^{tY}$. Dann erhalten wir für alle $t \in \mathbb{R}$

$$\hat{\varphi}(t) = \int_{\mathbb{R}} e^{tx} dQ^Y(x) = \frac{1}{g} \int_{\mathbb{R}} e^{tx} e^{\tau x} d\mathbb{P}^X(x) = \frac{1}{g} \varphi(t + \tau) < \infty.$$

Dies impliziert, dass mit φ auch $\hat{\varphi}$ eine C^∞–Funktion ist. Damit ergibt sich

$$\mathbb{E}Y = \hat{\varphi}'(0) = \frac{1}{g}\varphi'(\tau) = 0 \text{ und}$$
$$\mathbb{V}Y = \hat{\varphi}''(0) = \frac{1}{g}\varphi''(\tau) \in (0, \infty).$$

□

Lemma 2.5 *Es sei $T_n = \sum_{i=1}^n Y_i$. Dann gilt*

$$\mathbb{P}(S_n \geq 0) = g^n \mathbb{E}(e^{-\tau T_n} 1_{\{T_n \geq 0\}}).$$

Beweis: Beachtet man, dass

$$\begin{aligned}\mathbb{P}(S_n \geq 0) &= \int_{\sum_{i=1}^n x_i \geq 0} d\mathbb{P}^X(x_1) \dots d\mathbb{P}^X(x_n) \\ &= \int_{\sum_{i=1}^n x_i \geq 0} [ge^{-\tau x_1} dQ^Y(x_1)] \dots [ge^{-\tau x_n} dQ^Y(x_n)],\end{aligned}$$

so folgt die Behauptung. □

Lemma 2.6 *Es gilt*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}(e^{-\tau T_n} 1_{\{T_n \geq 0\}}) \geq 0.$$

Beweis: Aufgrund von Lemma 2.4 kann man den Zentralen Grenzwertsatz auf T_n anwenden. Wählen wir nun eine Zahl $C > 0$ so, dass

$$\frac{1}{\sqrt{2\pi}} \int_0^C e^{\frac{-x^2}{2}} dx > \frac{1}{4}$$

gilt, erhalten wir die folgende Schranke

$$\mathbb{E}(e^{-\tau T_n} 1_{\{T_n \geq 0\}}) \geq e^{-\tau C \sqrt{\mathbb{V}Y_1}\sqrt{n}} \mathbb{P}\left(\frac{T_n}{\sqrt{\mathbb{V}Y_1}\sqrt{n}} \in [0, C) \right).$$

Da die Wahrscheinlichkeit rechts für n gegen unendlich gegen eine Zahl $\geq \frac{1}{4}$ konvergiert, folgt die Behauptung. □

Der Beweis des Theorems ergibt sich nun, da aus Lemma 2.5 zusammen mit Lemma 2.6 folgt, dass

$$\liminf_{n\to\infty} \frac{1}{n} \mathbb{P}(S_n \geq 0) = \log g + \liminf_{n\to\infty} \frac{1}{n} \mathbb{E}(e^{-\tau T_n} 1_{\{T_n \geq 0\}}) \geq \log g.$$

Dies ist die Aussage des Theorems.

□

Bemerkung 2.7 *Durch Übergang von X_1 auf $-X_1$ und der Beobachtung, dass dies die Ratenfunktion nicht ändert, erhalten wir aus Satz 2.3 auch*

$$\lim \frac{1}{n} \log \mathbb{P}(S_n \leq na) = -I(a)$$

für alle $a < \mathbb{E}X_1$.

Da nun $I(x) = 0 \Leftrightarrow x = \mathbb{E}X_1$, denn

- $I(x) = \sup_t [tx - \psi(x)] \geq -\psi(0) = 0, \quad \forall x \qquad$ *und*
- $I(\mathbb{E}X_1) = \sup_t [t\mathbb{E}X_1 - \psi(t)] \underset{Jensen}{\leq} \sup_t [t\mathbb{E}X_1 - \log \exp t\mathbb{E}X_1] = 0$

und außerdem I strikt konvex ist auf

$$D_I := \{x : I(x) \neq \infty\},$$

folgt

$$\lim \frac{1}{n} \log \mathbb{P}(S_n \leq na) = - \inf_{x \in (-\infty, a]} I(x)$$

für alle $c < \mathbb{E}X_1$ und

$$\lim \frac{1}{n} \log \mathbb{P}(S_n \geq na) = - \inf_{x \in [a, \infty)} I(x).$$

Dies impliziert aber auch

$$\lim \frac{1}{n} \log \mathbb{P}(\frac{S_n}{n} \in (a, b)) = - \inf_{x \in (a,b)} I(x).$$

Z. B. überlegt man für $\mathbb{E}X_1 < a < b < \infty$

$$\begin{aligned} \lim \frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} \in (a, b)\right) &= \lim \frac{1}{n} \log \left(\mathbb{P}(S_n > na, -\mathbb{P}(S_n \geq nb)\right) \\ &\leq \lim \frac{1}{n} \log \mathbb{P}(S_n > na) = -I(a) = \inf_{x \in (a,b)} I(x) \end{aligned}$$

und

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} \in (a, b)\right) \geq \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n > na)(1 - \varepsilon) = -I(a) \quad \forall\, \varepsilon > 0.$$

Schließlich erhält man hieraus das LDP für $\frac{S_n}{n}$ mit Geschwindigkeit n und Rate I, indem man beliebige Mengen durch Intervalle approximiert.

Beispiel 2.8 *Sind die (X_i) i.i.d. $\mathcal{N}(0,\sigma^2)$-verteilt, so ist*

$$I(x) = \frac{x^2}{2\sigma^2},$$

wie man leicht nachrechnet.

Beispiel 2.9 *Wir haben schon in der Vorlesung über Wahrscheinlichkeitstheorie gesehen, dass für (X_i) die i.i.d. verteilt sind*

$$1 - p = \mathbb{P}(X_i = 0) = 1 - \mathbb{P}(X_i = 1)$$

die Ratenfunktion durch

$$I(x) = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p} =: H(x|p)$$

gegeben ist.

Bemerkung 2.10 *Aufgrund des Borel-Cantelli-Lemmas impliziert ein LDP wieder das Gesetz der großen Zahlen für $\frac{S_n}{n}$.*

3 Allgemeine Techniken und Prinzipien

In diesem Abschnitt sollen Maße studiert werden, die einem LDP genügen. Es werden sich einige Eigenschaften herauskristallisieren, die ganz allgemein gelten. Desweiteren werden wir gewisse Übertragungsmechanismen aufzeigen, die uns erlauben, wieder neue Prinzipien der großen Abweichungen zu erhalten. Wir beginnen damit, die exponentielle Konvergenz, die die Grundlage von Bemerkung **??** ist, im allgemeinen Rahmen zu studieren. Hier gibt es keinen Grund, dass eine Ratenfunktion genau eine Nullstelle besitzen sollte (dies ist im Allgemeinen auch falsch, wie wir später sehen werden). Dennoch bekommen wir eine exponentielle Konzentration auf "kleine" Mengen:

Satz 3.1 *Sei $(\mathbb{P}_n)_n$ eine Folge von Maßen, die einem LDP mit Geschwindigkeit n und Ratenfunktion I genügt. Dann gibt es für jedes $\ell < \infty$ eine kompakte Menge D^ℓ, so dass*

$$\mathbb{P}(D^\ell) \geq 1 - e^{-n\ell} \quad \forall\, n.$$

Beweis: Nach Annahme sind die Niveaumengen

$$A_\ell = \{x : I(x) \leq \ell + 2\}$$

kompakt, sie können also für jedes k durch eine endliche Menge kompakter Kugeln vom Radius $\frac{1}{k}$ überdeckt werden. Ihre Vereinigung sei U_k. Es ist

$$I(x) \geq \ell + 2$$

auf der abgeschlossenen Menge U_k^c. Daher folgt aus der oberen Abschätzung des LDP

$$\overline{\lim}\frac{1}{n}\log \mathbb{P}_n(A_k^c) \leq -(\ell+2)$$

und somit für n hinreichend groß ($n \geq n_0$)

$$\mathbb{P}_n(A_k^c) \leq e^{-n(\ell+1)}.$$

O.B.d.A. ist $n_0 = n_0(k) \geq k$ und daher für $n \geq n_0$

$$\mathbb{P}_n(A_k^c) \leq e^{-k}e^{-n\ell}.$$

Für die ersten Indizes $j = 1, \ldots, n_0(k)$ können wir ohnehin kompakte Mengen $B_{k,1}, \ldots, B_{k,n_0}$ finden, so dass

$$\mathbb{P}_j(B_{k,j}^c) \leq e^{-k}e^{-j\ell} \quad \forall\, j$$

gilt, denn einzelne Maße sind natürlich straff. Wir setzen

$$D^\ell = E = \bigcap_k [A_k \cup (\bigcup_j B_{k,j})].$$

E ist total beschränkt (Übung) und daher kompakt, da polnische Räume vollständig sind. Weiter gilt

$$\mathbb{P}_n(E^c) \leq [\sum_{k\geq 1} e^{-k}]e^{-n\ell} \leq e^{-n\ell}.$$

□

Bemerkung 3.2 *Die Eigenschaft, die wir hier nachgewiesen haben, erinnert an das Straffheitskriterium von Prohorov. Allerdings konzentrieren sich die Maße schneller als dort. Wir nennen die Eigenschaft daher (super-) exponentielle Straffheit.*

Wir wenden uns nun Situationen zu, bei denen sich ein LDP von einer Situation auf eine verwandte Situation übertragen lässt.

Satz 3.3 *Seien $(\mathbb{P}_n)_n$ und $(\mathbb{Q}_n)_n$ zwei Folgen von Maßen, die einem LDP mit Geschwindigkeit a_n und Rate $I(\cdot)$ bzw. $J(\cdot)$ genügen. Dann genügt auch die Folge der Produktmaße $\mathbb{R}_n = \mathbb{P}_n \otimes \mathbb{Q}_n$ auf $X \times Y$ einem LDP mit Geschwindigkeit a_n. Die Rate ist gegeben durch*

$$K(x,y) = I(x) + J(y).$$

Da der Beweis typisch für einige Techniken in den großen Abweichungen ist, geben wir ihn schrittweise.

Beweis: 1. Schritt: Sei

$$(x,y) = z \in Z = X \times Y.$$

Sei $\varepsilon > 0$. Wir wollen zeigen, dass es eine offene Umgebung $N = N_{x,\varepsilon}$ von z gibt, so dass

$$\overline{\lim}\frac{1}{a_n}\log \mathbb{R}_n(N) \leq -K(z) + \varepsilon \tag{6}$$

gilt. Da I von unten halbstetig ist, gibt es eine offene Menge $U_1 \subseteq X$, so dass $I(x') \geq I(x) - \frac{\varepsilon}{2}$ für alle $x' \in U_1$ gilt. Wir wählen (da U_1 offen ist) eine offene Menge $U_2 \subseteq X$ mit

$$x \in U_2 \subseteq \overline{U_2} \subseteq U_1.$$

Wegen des LDP für $(\mathbb{P}_n)$ gilt

$$\overline{\lim}\frac{1}{a_n}\log \mathbb{P}_n(U_2) \leq \overline{\lim}\frac{1}{a_n}\log \mathbb{P}_n(\overline{U_2}) \leq -\inf_{x'\in\overline{U_2}} I(x') \leq -I(x) + \frac{\varepsilon}{2}.$$

Analog findet man offene Mengen $V_1, V_2 \subseteq Y$ mit $y \in V_2 \subseteq \overline{V_2} \subseteq V_1$ und

$$\overline{\lim}\frac{1}{a_n}\log \mathbb{Q}_n(V_2) \leq \overline{\lim}\frac{1}{n}\log Q_n(\overline{V_2}) \leq -\inf_{y'\in\overline{V_2}} J(y) \leq -J(y) + \frac{\varepsilon}{2}.$$

Wenn wir $N = U_2 \times V_2$ als offene Umgebung von z wählen, sind wir fertig mit (6).

2. Schritt: Sei $D \subseteq Z$ kompakt und $\varepsilon > 0$. Wir wollen nun zeigen, dass für eine ε-Umgebung $D_\varepsilon \supseteq D$ von D gilt

$$\overline{\lim}\frac{1}{n}\log \mathbb{R}_n(D_\varepsilon) \leq -\inf_{z\in D} K(z) + \varepsilon. \tag{7}$$

Aus Schritt 1 wissen wir, dass für alle $z \in D$ eine Umgebung N_z existiert, so dass

$$\overline{\lim}\frac{1}{n}\log \mathbb{R}_n(N_z) \leq -K(z) + \varepsilon \leq -\inf_{z\in D} K(z) + \varepsilon$$

gilt. Aufgrund der Kompaktheit von D enthält die offene Überdeckung $\{N_z\}_{z\in D}$ von D eine endliche Teilüberdeckung $\{N_j : 1 \leq j \leq k\}$. Für $D_\varepsilon = \bigcup_{j=1}^k D_j$ erhalten wird

$$\mathbb{R}_n(D_\varepsilon) \leq k \cdot \sup_{j=1,\ldots,k} \mathbb{R}_n(D_j)$$

und daher

$$\overline{\lim}\frac{1}{a_n}\log \mathbb{R}_n(D_\varepsilon) \leq -\inf_{z\in D} K(z) + \varepsilon.$$

Da ε beliebig war, folgt (7).

3. Schritt: Aus der superexponentiellen Straffheit (Satz 3.1) wissen wir, dass es für jedes ℓ kompakte Teilmengen A_ℓ und B_ℓ von X bzw. Y gibt, $\mathbb{P}_n(A_\ell^c) \leq e^{-a_n\ell}$ und $\mathbb{Q}_n(B_\ell^c) \leq e^{-a_n\ell}$. Setzen wir $C_\ell = A_\ell \times B_\ell$, so ist C_ℓ nach dem Satz von Tychonov kompakt und es gilt

$$\mathbb{R}_n(C_\ell^c) \leq 2e^{-a_n\ell}.$$

Nun schreiben wir eine beliebige abgeschlossene Menge $C \subseteq Z$ als

$$C = (C \cap C_\ell) \,\dot\cup\, (C \cap C_\ell^c).$$

Damit bekommen wird

$$\begin{aligned}\overline{\lim}\frac{1}{a_n}\log\mathbb{R}_n(C) &\le \overline{\lim}\frac{1}{a_n}\log 2\cdot\max(\mathbb{R}_n(C\cap C_\ell),\mathbb{R}_n(C\cap C_\ell^c))\\ &= \overline{\lim}\frac{1}{a_n}\log\max(R_n(C\cap C_\ell),\ e^{-a_n\ell})\\ &\le \max(-\inf_{Z\in C_\ell} K(z),-\ell)\\ &\le \max(-\inf_{Z\in C} K(z),-\ell).\end{aligned}$$

Schickt man $\ell\to\infty$, erhält man die obere Schranke.

4. Schritt: Die untere Schranke benutzt ein lokales Argument (das ist typisch) und ist viel einfacher (das ist weniger typisch). Wir müssen zeigen, dass für jedes $z\in Z$ und eine Umgebung N von Z gilt:

$$\underline{\lim}\frac{1}{a_n}\log\mathbb{R}_n(N)\ge -I(z). \tag{8}$$

Ist N nun eine Umgebung von z. Dann enthält N eine Menge $U\times V$, wobei U und V Umgebungen von x bzw. y sind. Da $\mathbb{P}_n$ und $\mathbb{Q}_n$ ein LDP und somit die untere Schranke für U bzw. V erfüllt, folgt (8). □

In der Folge wollen wir uns mit Möglichkeiten befassen, ein LDP zu zeigen. Wir beginnen mit einer Technik, die allein auf topologischen Eigenschaften beruht.

Satz 3.4 *Falls $(\mathbb{P}_n)$ einem LDP mit Geschwindigkeit (a_n) und Rate I auf einem polnischen Raum X genügt underline*

$$F: X\to Y$$

stetig ist (dabei ist Y ein weiterer polinischer Raum), so genügt auch die Folge $\mathbb{Q}=\mathbb{P}_nF^{-1}=\mathbb{P}_n^F$ einem LDP mit Geschwindigkeit (a_n) und Rate

$$J(y)=\inf_{x:F(x)=y} I(x).$$

Beweis: Sei $X\subseteq Y$ abgeschlossen. Dann ist auch $F^{-1}[C]\subseteq X$ abgeschlossen und es gilt

$$\begin{aligned}\overline{\lim}\frac{1}{a_n}\log Q_n(C) &= \overline{\lim}\frac{1}{a_n}\log\mathbb{P}_n(F_n^{-1}[C])\\ &\le -\inf_{x\in F^{-1}[C]} I(x)\\ &= -\inf_{y\in C}\inf_{X:F(x)=y} I(x)\\ &= -\inf_{y\in C} J(y).\end{aligned}$$

Analog ist für $G \subseteq Y$ offen auch $F^{-1}[G] \subseteq X$ offen und es gilt

$$\underline{\lim} \frac{1}{a_n} \log \mathbb{Q}_n(G) \geq - \inf_{x \in F^{-1}[G]} I(x) = - \inf_{y \in G} J(y).$$

□

Bemerkung 3.5 *Obwohl das obige Prinzip sehr einfach ist, ist es dennoch sehr nützlich. Die Hauptschwierigkeit besteht darin, dass die Ratenfunktion oft durch die Form in Satz 3.4 sehr implizit gegeben ist.*

Der folgende Satz stellt eine etwas komplexere Technik dar, ein LDP herzuleiten. Die Technik ist dem Beweis des Satzes von Cramér entlehnt.

Wir werden das folgende Gesetz nur für $\mathbb{R}^d$-wertige Zufallsvariablen formulieren, es gilt aber unter entsprechenden Voraussetzungen auch allgemeiner. Hierzu sei $(Z_n)_n$ eine Folge von $\mathbb{R}^d$-wertigen Zufallsvariablen. Weiterhin sei

$$\varphi_n(t) := \mathbb{E} e^{<t, Z_n>}, \quad t \in \mathbb{R}^d, \quad n \in \mathbb{N}.$$

Es gelte

$$\lim_{n \to \infty} \frac{1}{a_n} \log \varphi_n(a_n t) =: \Lambda(t) \in [-\infty, \infty] \quad \text{existiert} \tag{9}$$

und

$$0 \in \inf(D_\Lambda) \quad \text{mit} \quad D_\Lambda := \{t \in \mathbb{R}^d : \Lambda(t) < \infty\}. \tag{10}$$

Wir schauen uns zunächst Λ genauer an.

Lemma 3.6 *Λ hat die Eigenschaften:*

1. *Λ ist konvex und $\Lambda > -\infty$,*
2. *$\Lambda(x)^* = \sup_{t \in \mathbb{R}^d}(< x, t > - \Lambda(t))$ ist eine konvexe Ratenfunktion.*

Beweis:

1. Sei $0 < \lambda < 1$. Aus der Hölder-Ungleichung folgt

$$\begin{aligned} &\lim \frac{1}{a_n} \log F(e^{\lambda t_1 a_n + (1-\lambda) t_2 a_n) Z_n}) \\ \leq &\lim \frac{1}{a_n} \log (\mathbb{E} e^{t_1 a_n Z_n})^\lambda (\mathbb{E} e^{t_2 a_n Z_n})^{(1-\lambda)} \\ = &\lambda \Lambda(t_1) + (1-\lambda) \Lambda(t_2). \end{aligned}$$

   Dass $\Lambda > -\infty$ ist, folgt aus der Konvexität.

2. Da Λ konvex ist, ist auch Λ^* konvex und unterhalb stetig. Die Nicht-Negativität von Λ^* folgt wie im Satz von Cramér. Schließlich besitzt Λ^* kompakte Niveaumengen. Der Beweis zeigt die Abgeschlossenheit (die folgt aus der unteren Halbstetigkeit) und die Beschränktheit und ist eine Übung.

□

Definition 3.7 *Für Ratenfunktionen I schreibe*

$$I(S) = \inf_{s \in S} I(s). \tag{11}$$

Definition 3.8 *$x \in \mathbb{R}^d$ heißt exponierter Punkt bzgl. Λ^* genau dann wenn $t \in \mathbb{R}^d$ existiert, so dass*

$$\Lambda^*(y) - \Lambda^*(x) \geq \ < y - x, t > \quad \forall\, y \neq x$$

gilt. t (aufgefasst als Element aus dem Dualraum des $\mathbb{R}^d$) heißt exponierende Hyperebene in x. E sei die Menge der exponierten Punkte aus $\mathbb{R}^d$ mit exponierenden Hyperebenen $t \in \inf(D_\Lambda)$.

Definition 3.9 *Sei $\emptyset \neq A \subseteq \mathbb{R}^d$ konvex. Das relative Innere von A ist definiert als*

$$rint(A) := \{x \in A : \ \forall\, y \in A \exists \varepsilon > 0 : x - \varepsilon(y - x) \in A\}.$$

Es gilt stets: $rint(A) \supset$ int (A) und $rint(\{x\}) = \{x\}$.

Wir sind nun in der Lage, die angesprochene Technik, ein LDP zu beweisen, herzuleiten.

Satz 3.10 ***(Gärtner-Ellis-Theorem)*** *Sind* (9) *und* (10) *erfüllt, so gilt für die Verteilung $\mathbb{P}_n$ von Z_n*

1.

$$\limsup_{n \to \infty} \frac{1}{a_n} \log \mathbb{P}_n(A) \leq -\Lambda^*(A) \tag{12}$$

für alle $A \subseteq \mathbb{R}$ abgeschlossen.

2.

$$\liminf_{n \to \infty} \frac{1}{a_n} \log \mathbb{P}(G) \geq -\Lambda^*(G \cap E) \tag{13}$$

für alle $G \subseteq \mathbb{R}^d$ offen.

3. Gilt zudem
a) Λ ist auf $\mathbb{R}^d$ unterhalbstetig,
b) Λ ist im Inneren von D_Λ differenzierbar,
c) $D_\Lambda = \mathbb{R}^d$ oder $\lim_{\substack{t \to D_\lambda \\ t \in D_\Lambda}} |\nabla \Lambda(t)| = \infty$,
so kann man die rechte Seite in (13) *durch $\Lambda^*(G)$ ersetzen.*

Beweis: Wenig überraschend gliedert sich der Beweis in drei Teile.

1. Schritt: Hier zeigen wir (12). Wir beginnen mit einer kompakten Menge $C \subseteq \mathbb{R}^d$. Für eine Vorüberlegung sei $\delta > 0$ und

$$\Lambda_\delta^* := \min\{\Lambda^* - \delta, \frac{1}{\delta}\}.$$

Da $\Lambda_\delta^* \leq \Lambda^* = \sup(< x, t > -\Lambda(t))$, gibt es $\forall\, x$ ein $t_x \in \mathbb{R}^q$, so dass

$$< x, t_x > -\Lambda(t_x) \geq \Lambda_\delta^*(x).$$

Sei A_x eine beliebige Umgebung von x, für die gilt:

$$\begin{aligned} \inf_{J \in A_x} < y, t_x > \quad &\geq \quad < x, t_x > -\delta \\ \text{d. h.} \quad \inf_{J \in A_x} < y - x, t_x > \quad &\geq -\delta. \end{aligned}$$

Mittels der exponentiellen Chebyshev-Ungleichung erhält man wie immer für eine Zufallsvariable X

$$\mathbb{P}(X \geq \varepsilon) \leq e^{-a_n \varepsilon} \mathbb{E} e^{a_n X}$$

und daher

$$\begin{aligned} \mathbb{P}(Z_n \in A_x) &\leq \mathbb{P}(< Z_n - x, t_x > \ \geq -\delta) \\ &\leq e^{\delta a_n} \mathbb{E} e^{a_n < Z_n - x, t_x >} \\ &= e^{\delta a_n} e^{-a_n < t, x >} \mathbb{E} e^{a_n < t_x, Z_n >} \\ &= e^{\delta a_n} e^{-a_n < t, x >} \varphi_n(a_n t_x). \end{aligned}$$

Da C kompakt ist, besitzt die offene Überdeckung durch $(A_x)_{x \in C}$ eine endliche Teilüberdeckung $(A_{x_i})_{i=1}^N$. Also:

$$\begin{aligned} \frac{1}{a_n} \log \mathbb{P}_n(C) &\leq \frac{1}{a_n} \log[N \max_{1,\ldots,N} \mathbb{P}_n(A_{x_i})] \\ &= \frac{1}{a_n} \log N + \delta + \max_{i=1,\ldots,N} (- < x, t_{x_i} > + \frac{1}{a_n} \log \varphi_n(a_n t_x)) \end{aligned}$$

und daher

$$\begin{aligned} \lim_{n \to \infty} \frac{1}{a_n} \log \mathbb{P}_n(C) &\leq \delta - \min_{i=1,\ldots,N} (< x, t_{x_i} > -\Lambda(t_{x_i})) \\ &\leq \delta - \min_{i=1,\ldots,N} \Lambda_\delta^*(x_i) \\ &\leq \delta - \Lambda_\delta^*(x_i). \end{aligned}$$

Für $\delta \to 0$ folgt die Abschätzung für kompakte Teilmengen. Ist nun $A \subseteq \mathbb{R}^d$ abgeschlossen, so ist $A \cap [-N, N]^d =: A \cap \tilde{N}$ für jedes N kompakt. Verwenden wir die Vorüberlegung für kompakte Mengen, so ergibt sich

$$\begin{aligned} &\overline{\lim} \frac{1}{a_n} \log \mathbb{P}_n C(A) \leq \overline{\lim} \frac{1}{a_n} \log(\mathbb{P}_n(A \cap \tilde{N}) + \mathbb{P}(A \cap \tilde{N})) \\ &= \overline{\lim} \frac{1}{a_n} \log \max(\mathbb{P}_n(A \cap \tilde{N}), \mathbb{P}_n(A \cap \tilde{N}^c)) \\ &\leq \overline{\lim} \frac{1}{a_n} \log \max(\Lambda^*(A \cap \tilde{N}), -M_N) \end{aligned}$$

mit

$$-M_N = \overline{\lim}\frac{1}{a_n}\log \mathbb{P}_n[\tilde{N}^c].$$

Wegen (11) gibt es nun ein $\delta > 0$, sodass für alle Einheitsvektoren $e^{(i)}$ des $\mathbb{R}^d$ gilt: $\Lambda(\pm\delta e^{(i)}) < \infty$. Mithilfe des exponentiellen Chebyshev folgt dann für alle i und die i-te Koordinate $Z_n^{(i)}$ von Z_n:

$$\mathbb{P}(Z_n^{(i)} \leq -N) \leq e^{-a_n\delta_N}\varphi_n(-a_n\delta e^{(i)})$$

und

$$\mathbb{P}(Z_n^{(i)} \geq N) \leq e^{-a_n\delta N}\varphi_n(a_n\delta e^{(i)}).$$

Dadurch lässt sich M_N folgendermaßen abschätzen:

$$\begin{aligned}
-M_N &\leq \overline{\lim}\frac{1}{a_n}\log \mathbb{P}(\exists i = 1,\dots,d : Z_N^{(i)} \notin [-N,N]) \\
&\leq \overline{\lim}\frac{\log d}{a_n} + \max_{i=1,\dots,d}\frac{1}{a_n}\log\max(e^{-a_n\delta N}\varphi_n(-a_n\delta e^{(i)})) \\
&= -\delta N + \max_i \max\{\Lambda(-\delta e^{(i)}), \Lambda(\delta e^{(i)})\}
\end{aligned}$$

Der zweite Summand auf der rechten Seite ist endlich und hängt nicht von N ab. Der erste Summand geht für $N \to \infty$ gegen $-\infty$. Also ist $\lim_{N\to\infty} -M_N = -\infty$ und daraus folgt die obere Abschätzung.

2. Schritt: Wir zeigen (13). Sei $G \subseteq \mathbb{R}^d$ offen und $x \in G$. Dann gibt es ein $\varepsilon_x > 0$, sodass für alle $\varepsilon \leq \varepsilon_x$ gilt $B_\varepsilon(X) \leq G$. Wir zeigen

$$\lim_{\varepsilon\downarrow 0}\underline{\lim}_{n\to\infty}\frac{1}{a_n}\log \mathbb{P}_n(B_\varepsilon(x)) \geq -\Lambda^*(x) \quad \forall\, x \in E. \tag{14}$$

Dies gilt weil $\mathbb{P}_n(G) \geq P_n(B_\varepsilon(x))$ und wir auf der rechten Seite von (14) dann das Infimum über $x \in G \cap E$ bilden können. Die Kernidee ist es, wie beim Satz von Cramér das Wahrscheinlichkeitsmaß so zu definieren, dass G unter dem neuen Wahrscheinlichkeitsmaß typisch ist. Dazu sei $\tau \in \text{int}\,(D_\Lambda)$ eine exponierende Hyperebene. Beachte, dass $\lim\varphi_n(a_n\tau) < \infty$, also auch $\varphi_n(a_n\tau) < \infty$ für hinreichend großes n. Definiere $\mathbb{Q}_n$ vermöge

$$\frac{d\mathbb{Q}_n}{d\mathbb{P}_n}(y) = \frac{e^{a_n<y,\tau>}}{\varphi_n(a_n\tau)}.$$

Somit ist

$$\begin{aligned}
\frac{1}{a_n}\log \mathbb{P}_n(B_\varepsilon(x)) &= \frac{1}{a_n}\log\int_{B_\varepsilon(x)}\varphi_n(a_n\tau)e^{-a_n<y,\tau>}dQ_n(y) \\
&= \frac{1}{a_n}\log\varphi_n(a_n\tau) + \frac{1}{a_n}\log\int_{B_\varepsilon(x)} e^{-a_n<y,\tau>}dQ_n(y) \\
&\geq \frac{1}{a_n}\log\varphi_n(a_n\tau) + \frac{1}{a_n}\log(e^{-a_n(<x,\tau>+\varepsilon|\tau|)}\cdot\int_{B_\varepsilon(x)}\mathbb{Q}_n(dy)) \\
&= \frac{1}{a_n}\log\varphi_n(a_n\tau) - <x,\tau> - \varepsilon|\tau| + \frac{1}{a_n}\log Q_n(B_\varepsilon(x)).
\end{aligned}$$

Hierbei folgt die Ungleichung aus

$$< y, \tau >=< x, \tau > + < y - x, \tau > \ \leq \ < x, \tau > +\varepsilon|\tau|,$$

da $y \in B_\varepsilon(x)$ ist. Also zusammen:

$$\begin{aligned}\lim_{\varepsilon \downarrow 0} \varliminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}_n(B_\varepsilon(x)) &\geq [\Lambda(\tau - < x, \tau >] + \lim_{\varepsilon \downarrow 0} \varliminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{Q}_n(B_\varepsilon(x)) \\ &\geq -\Lambda^*(x) + \lim_{\varepsilon \downarrow 0} \varliminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{Q}_n(B_\varepsilon(x)),\end{aligned}$$

weil

$$\Lambda^* \geq [< x, \tau > -\Lambda(\tau)].$$

Um zu zeigen, dass der 2. Summand gleich 0 ist, bilden wir die kumulantenerzeugende Funktion $\tilde{\varphi}_n$ von Q_n und berechnen diese wie im Satz von Cramér als

$$\tilde{\varphi}_n(nt) = \frac{\varphi_n(n(t+\tau))}{\varphi_n(n\tau)}.$$

Also $\tilde{\Lambda}(t) = \Lambda(t+\tau) - \Lambda(\tau)$. Damit erhält man

$$\tilde{\Lambda}^*(x) = \Lambda^*(x) - < x, \tau > +\Lambda(\tau). \tag{15}$$

Nach Lemma 3.6 ist $\tilde{\Lambda}^*(t)$ eine Ratenfunktion und wir können Schritt 1 anwenden, um auf

$$\varlimsup \frac{1}{a_n} \log \mathbb{Q}_n(B_\varepsilon^c(x)) \leq -\tilde{\Lambda}^*(B_\varepsilon^c(x))$$

zu schließen.

Da $\tilde{\Lambda}^*$ als Ratenfunktion kompakte Niveaumengen hat und auf diesen aufgrund der Halbstetigkeit ihr Minimum annimmt, gibt es ein $x_0 \neq x_0$ mit

$$\tilde{\Lambda}^*(B_\varepsilon^c(x)) = \tilde{\Lambda}^*(x_0).$$

Wir wissen zudem, dass x ein exponierter Punkt und $\tau \in \text{int}\,(D_\Lambda)$ eine exponierende Hyperebene für x ist. Wenden wir die Definition des exponierten Punktes mit x_0 auf (15) an, ergibt sich

$$\begin{aligned}\tilde{\Lambda}^*(x_0) &= \Lambda^*(x_0) - < x_0, \tau > +\Lambda(\tau) \\ &\geq [\Lambda^*(x_0) - < x_0, \tau >] + < x, \tau > -\Lambda^*(x) \\ &> 0.\end{aligned}$$

Hierbei ist das erste Ungleichheitszeichen die Definition von Λ^*, das zweite die Definition des exponierten Punktes, aus welcher $\Lambda^*(x_0) - \Lambda^*(x) > \ < x_0 - x, \tau >$ für alle $x_0 \neq x_0$ folgt. Damit ist

$$\varlimsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}_n(B_\varepsilon^c(x)) < 0 \quad \forall\, \varepsilon > 0$$

und daraus folgt

$$\lim_{\varepsilon \downarrow 0} \varliminf_{n\to\infty} \frac{1}{a_n} \log \hat{\mathbb{P}}_n(B_\varepsilon(x)) = 0.$$

3. Schritt: Hier wollen wir sehen, dass sich die untere Abschätzung unter den stärkeren Bedingungen von $G \cap E$ auf G übertragen lässt. Wir benötigen zunächst ein Lemma aus der konvexen Analysis, das wir hier nicht beweisen können und wollen. Es findet sich in dem Buch von Rockefellar [Ro], Seite 257, Corollary 4.1.

Lemma 3.11 *Für Λ gilt unter der zusätzlichen Annahme in Satz 3.10 c)*

$$E > \text{int}(D_{\Lambda^*}).$$

Für die Anwendung dieses Lemmas benötigen wir die folgenden zusätzlichen Erkenntnisse über Λ oder Λ^*:

- Λ^* ist konvex, also ist D_{Λ^*} eine konvexe Menge.
- Da Λ^* eine Ratenfunktion und somit $\neq \infty$ ist, ist $D_{\Lambda^*} \neq \emptyset$ und somit auch rint $(D_{\Lambda^*}) \neq \emptyset$.

Wir wollen nun zeigen, dass unter den Bedingungen in Satz 3.10 c) $\Lambda^*(G \cap E) = \Lambda^*(G)$ gilt. Aus der Definition von Λ^* folgt

$$\Lambda^*(G \cap r\text{int}(D_{\Lambda^*})) \geq \Lambda^*(G)$$

und wegen 3.11 genügt es daher zu zeigen, dass

$$\Lambda^*(G \cap r\text{int}(D_{\Lambda^*})) \leq \Lambda^*(G)$$

gilt.

Ist $G \cap D_{\Lambda^*} = \emptyset$, so ist $\Lambda^*(G) = \infty$ (da D_{Λ^*} gerade die Menge der endlichen Λ^*-Werte ist). Somit können wir $G \cap D_{\Lambda^*} \neq \emptyset$ annehmen. Sei also $y \in G \cap D_{\Lambda^*}$ und $Z \in$ rint (D_{Λ^*}) (wir haben schon bemerkt, dass so ein z existiert).

Aufgrund der Konvexität besteht D_{Λ^*} entweder aus einem einzigen Punkt $(y = z)$ oder aus unendlich vielen. In beiden Fällen liegt die Strecke $\overline{yz}$ in D_{Λ^*}. Somit gilt für alle $\delta > 0$ klein genug, so dass $\delta z + (1 - \delta)y \in G \cap$ rint (D_{Λ^*}); dies ist klar, falls $y = z$ gilt, andernfalls nutzt man aus, dass G eine offene Menge ist. Also folgt:

$$\Lambda^*(G \cap r\text{int}(D_{\Lambda^*})) \leq \Lambda^*(\delta z + (1 - \delta)y)$$

für alle δ klein genug. Nun ist aber

$$\lim_{\delta \downarrow 0} \Lambda^*(\delta z + (1 - \delta)y) \leq \Lambda^*(y)$$

aufgrund der Konvexität von Λ^*. Da $y \in G \cap D_{\Lambda^*}$ beliebig war, folgt

$$\Lambda^*(G \cap r\text{int}(D_{\Lambda^*})) \leq \Lambda^*(G \cap D_{\Lambda^*}).$$

Da $\Lambda^*(G) = \Lambda^*(G \cap D_{\Lambda^*})$, da für alle $x \in G | D_{\Lambda^*}$, $\Lambda^*(x) = \infty$ ist, folgt die fehlende Ungleichung und damit die Behauptung. $\square$

Das Gärtner-Ellis-Theorem stellt eine wesentliche Technik dar, um LDPs mit konvexer Ratenfunktion herzuleiten. Wir werden im folgenden Kapitel eine Anwendung sehen. Vorher wollen wir noch einen Satz kennenlernen, der einen der wesentlichen Gründe darstellt, warum das Studium der Theorie der großen Abweichungen vertieft wurde.

Satz 3.12 ***(Varadhan):*** *Es sei $(X_n)_n$ eine Folge von Zufallsvariablen in $\mathbb{R}^d$, die einem LDP mit Geschwindigkeit n und Rate I genügt und*

$$f : \mathbb{R}^d \to \mathbb{R}$$

sei stetig und beschränkt. Dann gilt

$$\lim \frac{1}{n} \log \mathbb{E} e^{nf(X_n)} = \sup_x [f(x) - I(x)]$$

(wobei die Existenz des Limes gleich mitbehauptet wird).

Der Beweis dieses Satzes ist recht instruktiv:

Beweis: Da f stetig und beschränkt ist, lässt sich eine Überdeckung des $\mathbb{R}^d$ durch endlich viele abgeschlossene Mengen $A_1, \ldots, A_M$ finden, so dass f auf jede dieser Mengen höchstens um ein vorgegebenes $\delta > 0$ variiert, also

$$\sup_{x \in A_i} f(x) - \inf_{x \in A_i} f(x) \leq \delta \quad \forall\, i = 1, \ldots, M.$$

Dann folgt

$$\begin{aligned}
\int_{\mathbb{R}^d} e^{nf(X_n)} d\mathbb{P} &\leq \sum_{j=1}^{M} \int_{A_j} e^{nf(X_n)} d\mathbb{P} \\
&\leq \sum_{j=1}^{M} \int_{A_j} e^{n \sup_{y \in A_j} f(y)} d\mathbb{P}_{X_n}(y) \\
&\leq \sum_{j=1}^{M} \int_{A_j} e^{n(\inf_{y \in A_j} f(y) + \delta)} d\mathbb{P}_{X_n}(y) \\
&= \sum_{j=1}^{M} e^{n(\inf_{y \in A_j} f(y) + \delta)} \mathbb{P}(X_n \in A_j).
\end{aligned}$$

Also

$$\begin{aligned}
\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{E} e^{nf(X_n)} &\leq \sup_{1 \leq j \leq M} [\inf_{y \in A_j} f(y) + \delta - \inf_{y \in A_j} I(y)] \\
&\leq \sup_{1 \leq j \leq M} [\sup_{y \in A_j} (f(y) - I(y)) + \delta] \\
&= \sup_{y \in \mathbb{R}^d} (f(y) - I(y)) + \delta.
\end{aligned}$$

Lässt man δ gegen 0 gehen, ergibt sich die obere Schranke. Für die untere Schranke argumentiert man, wie oft in der Theorie der großen Abweichungen, lokal. Da f stetig ist, gibt es zu jedem $y_0 \in \mathbb{R}^d$ und $\varepsilon > 0$ eine offene Umgebung $U_\varepsilon(y_0)$, so dass für alle $y \in U_\varepsilon(y_0)$ gilt

$$f(y) \geq f(y_0) - \varepsilon.$$

Somit folgt

$$\mathbb{E}e^{nf(X_n)} \geq \int_{U_\varepsilon(y_0)} e^{nf(X_n)} d\mathbb{P} \geq e^{n(f(y_0)-\varepsilon)} \mathbb{P}(X_n \in U_\varepsilon(y_0)).$$

Mit Hilfe des LDP ergibt sich

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}e^{nf(X_n)} \geq f(y_0) - \varepsilon - \inf_{y \in U_\varepsilon(y_0)} I(y) \geq f(y_0) - I(y_0) - \varepsilon.$$

Da dies für alle y_0 stimmt, folgt

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}e^{nf(X_n)} \geq \sup_y [f(y) - I(y)].$$

□

Ganz analog zu Varadhans Lemma lässt sich der folgende Satz über große Abweichungen beweisen:

Satz 3.13 *Die Folge von Zufallsvariablen (X_n) im $\mathbb{R}^d$ genüge einem LDP mit Geschwindigkeit n und Ratenfunktion I. Für eine stetige beschränkte Funktion $F : \mathbb{R}^d \to \mathbb{R}$ setze*

$$J_n(S) = \int_S e^{nF(x)} d\mathbb{P}_{X_n}(x), \quad S \subseteq \mathbb{R}^d \text{ Borelsch.}$$

Definiere weiter

$$\mathbb{P}_n^F(S) = \frac{J_n(S)}{J_n(\mathbb{R}^d)};\ S \in \mathcal{B}^d$$

die $\mathbb{P}_n^F$ sind Wahrscheinlichkeitsmaße. Dann genügt die Folge der $(\mathbb{P}_n^F)_n$ einem LDP mit Geschwindigkeit n und Ratenfunktion

$$I^F(x) = -[F(x) - I(x)] + \sup_{y \in \mathbb{R}^d} [F(y) - I(y)].$$

Beweis: Es ist

$$\limsup \frac{1}{n} \log \mathbb{P}_n^F(S) = \limsup \frac{1}{n} \log J_n(S) - \limsup \frac{1}{n} \log J_n(\mathbb{R}^d)$$

und Analoges für liminf. Nun wissen wir aus dem vorhergehenden Satz schon, dass

$$\lim \frac{1}{n} \log J_n(\mathbb{R}^d) = \sup_y [F(y) - I(y)]$$

gilt. Somit bleibt zu zeigen, dass für alle offenen Mengen G und alle abgeschlossenen Mengen A gilt

$$\liminf \frac{1}{n} \log J_n(G) \geq - \inf_{x \in G} [F(x) - I(x)]$$

und

$$\limsup \frac{1}{n} \log J_n(A) \leq - \inf_{x \in A} [F(x) - I(x)].$$

Dies aber geht genau wie im Beweis des Varadhanschen Lemmas. □

Zuletzt bemerken wir, dass auch Satz 3.12 und Satz 3.13 ihre Gültigkeit behalten, wenn wir die Geschwindigkeit n durch die Geschwindigkeit a_n ersetzen.

4 Zwei Anwendungsbeispiele

Der Sinn dieses Kapitels besteht darin zu zeigen, dass die Resultate des vorhergehenden Abschnitts sinnvoll sind, um interessante Resultate zu erzielen. Wir werden uns mit Namen mit zwei Situationen beschäftigen. Die erste geht wieder von einer Folge von i.i.d. Zufallsvariablen $(X_i)_{i \in \mathbb{N}}$ aus. Der Satz von Cramér besagt, dass $(\frac{\sum X_i}{n} = S_n)_n$ einem LDP genügt mit Geschwindigkeit n. Allerdings zeigt eine einfache Anwendung der Chebyshevschen Ungleichung, dass für $\frac{1}{2} < \alpha < 1$, dass

$$\mathbb{P}\left(\left| \frac{\sum_{i=1}^n (X_i - \mathbb{E}X)}{n^\alpha} \right| > \varepsilon \right) \leq \frac{1}{n^{2\alpha} \varepsilon^2} \mathbb{V}(\sum X_i) \leq \frac{n \cdot \mathbb{V}X}{\varepsilon^2 n^{2\alpha}} = \text{const} \cdot n^{1-2\alpha} \to 0.$$

Also konergiert $\frac{\sum (X_i - \mathbb{E}X_i)}{n^\alpha}$ gegen 0. Kann man auch hierfür ein LDP zeigen? Unter exponentiellen Momentebedingungen ist dies in der Tat der Fall. Wir zeigen das folgende

Satz 4.1 ***(Moderate Abweichungen für i.i.d. Folgen).*** *Sei $X_1, \ldots, X_n$ eine Folge von i.i.d. $\mathbb{R}^d$-wertigen Zufallsvariablen, sodass*

$$\Lambda_X(\lambda) := \log \mathbb{E} e^{<\lambda, X_1>} < \infty$$

für alle $\lambda \in B_\varepsilon(0)$ für ein $\varepsilon > 0$. Ohne Einschränkung sei $\mathbb{E}X_1 = 0$ und dass die Kovarianzmatrix C von X_1 invertierbar ist. Für $\frac{1}{2} < \alpha < 1$ sei

$$Z_n = \frac{1}{n^\alpha} \sum_{i=1}^n X_i.$$

Dann genügt Z_n einem LDP mit Geschwindigkeit $n^{2\alpha - 1}$ und Rate

$$I(x) = \frac{1}{2} < x, C^{-1} x > .$$

Bemerkung 4.2 *a) Satz* **??** *heißt moderates Abweichungsprinzip. Der Unterschied zu einem großen Abweichungsprinzip ist die unterschiedliche Skala – im Prinzip gibt es keinen formalen Unterschied zu einem LDP. Ein moderates Abweichungsprinzip liegt dem Sprachgebrauch nach auf der Ebene zwischen einem Zentralen Grenzwertsatz und einem Gesetz der großen Zahlen. Ein MDP erbt von einem Zentralen Grenzwertsatz die* Universalität *des Grenzwerts: Sind die (X_i) so, dass $C = Id$, so ergibt sich stets als Ratenfunktion $I(x) = \frac{1}{2}\sum_{i=1}^{d} x_i^2$. Auf der anderen Seite erbt das MDP vom LDP das exponentielle Verhalten.*

b) Die Bedingungen in Satz **??** *sind nicht optimal. Die bisher besten sind in [EL2003] gegeben.*

Beweis: Der Beweis ist eine Anwendung des Gärtner-Ellis-Theorems. Zu diesem Zweck sei

$$\begin{aligned}\Lambda(\lambda) :&= \mathbb{E}[< \lambda, X_1 >^2]/2 \\ &= < \lambda, C\lambda > /2.\end{aligned}$$

Dann ist die Legendre-Fenchel-Transformierte von Λ gegeben durch

$$\begin{aligned}\Lambda^*(x) &= \sup_{\lambda\in\mathbb{R}^d}\{< \lambda, x > -\Lambda(\lambda)\} \\ &= \sup_{\lambda\in\mathbb{R}^d}\{< \lambda, x > -\frac{1}{2} < \lambda, C\lambda >\} \\ &= \frac{1}{2} < \lambda, C^{-1}\lambda > .\end{aligned}$$

Nun sei

$$\Lambda_n(\lambda) = \log \mathbb{E}e^{\lambda Z_n>},\ \lambda \in \mathbb{R}^d.$$

Wenn wir zeigen können, dass für alle $\lambda \in \mathbb{R}^d$ gilt

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n^{2\alpha-1}}\Lambda(n^{2\alpha-1}Z_n), \tag{16}$$

so folgt Satz 4.1 aus dem Gärtner-Ellis-Theorem, da $\Lambda(\cdot)$ überall endlich und differenzierbar ist. Bleibt (16) zu zeigen. Wir bemerken, dass

$$\begin{aligned}\Lambda_n(n^{2\alpha-1}Z_n) &= \log \mathbb{E}e^{n^{2\alpha-1}<\lambda,Z_n>} \\ &= \sum_{i=1}^{n} \log \mathbb{E}e^{n^{\alpha-1}<\lambda,X_i>} \\ &= n \cdot \log \mathbb{E}e^{n^{\alpha-1}<\lambda,X_1>}.\end{aligned}$$

Da $\Lambda_X(\lambda) < \infty$ ist für alle $\lambda \in B_\varepsilon(0)$, folgt $\mathbb{E}e^{n^{\alpha-1}<\lambda,X_1>} < \infty$ für n hinreichend groß, denn $n^{\alpha-1} \to 0$. Hithilfe einer Taylor-Entwicklung und dem Satz über dominierte Konvergenz erhalten wir

$$\mathbb{E}e^{n^{\alpha-1}<\lambda,X_1>} = 1 + \mathbb{E}n^{\alpha-1} < \lambda, X_1 > +n^{2\alpha-2}\frac{1}{2}E < \lambda, X_1 >^2 +\mathcal{O}(n^{3\alpha-3}).$$

Der zweite Summand verschwindet wegen

$$\mathbb{E} < \lambda, X_1 >= 0.$$

Also

$$\frac{1}{n^{2\alpha-1}}\Lambda_n(n^{2\alpha-1}\lambda) = n \cdot \frac{1}{n^{2\alpha-1}}\Lambda_n(n^{2\alpha-1}\lambda)$$
$$= \frac{1}{n^{2\alpha-2}} \log(1 + \frac{1}{2}n^{2\alpha-2}\mathbb{E} < \lambda, X_1 >^2 + 0(n^{3\alpha-3}))$$

und daher

$$\lim_{n\to\infty} \frac{1}{n^{2\alpha-1}}\Lambda_n(n^{2\alpha-1}\lambda) = \frac{1}{2}\mathbb{E} < \lambda, X_1 >^2 = \Lambda(\lambda).$$

□

Bemerkung 4.3 *Der Satz besagt im Wesentlichen, dass man nicht ganz falsch liegt, wenn man die Wahrscheinlichkeit für* $\mathbb{P}(\frac{S_n}{n^\alpha} \geq \alpha)$ *entweder mithilfe des Zentralen Grenswertsatzes annähert (obwohl das aus dem Zentralen Grenzwertsatz heraus nicht zu rechtfertigen ist) oder indem man aus dem Satz von Cramér die folgende Näherung probiert*

$$\mathbb{P}(S_n \geq n^\alpha a) \approx e^{-n} I(n^{\alpha-1}a)$$

und dann I (die Rate im Satz von Cramér) Taylor-entwickelt. Obwohl beide Rechnungen nicht mathematisch zu rechtfertigen sind, ist es schwierig, Gegenbeispiele zu finden (siehe z. B. [EL2004] oder [LM2012].

Eine zweite Anwendung befasst sich mit einem Modell für Ferromagnetismus aus der statistischen Mechanik. Naiv lässt sich eine magnetische Substanz so modellieren, dass sie aus vielen Atomen aufgebaut ist, die alle einen magnetischen Dipol besitzen, der entweder nach oben zeigt $(+1)$ oder nach unten (-1). Wenn alle Atome rein zufällig ihren Dipol wählen und nicht interagieren, so wird die durchschnittliche Magnetisierung

$$m_N(\sigma) := \frac{\sum_{i=1}^N \sigma_i}{N}$$

nach dem Gesetz der großen Zahlen nahezu 0 sein, wir sehen kein magnetisches Verhalten. Hierbei bezeichne σ_i den "Spin" des i-ten Atoms. Die Interaktion, die für das magnetische Verhalten notwendig ist, kodiert man in einer Energie- oder Hamiltonfunktion $H_N(\sigma)$ und wählt die Spinkonfiguraten $\sigma = (\sigma_i)_{i=1}^N$ gemäß dem sogenannten Gibbs-Maß aus:

$$\mu_N(\sigma) = \exp(-\beta H_N(\sigma))/Z_N.$$

Dabei ist $\beta > 0$ ein zusätzlicher Parameter, die inverse Temperatur $\beta = \frac{1}{T}$, Z_N die Zustandssume, macht μ_N zu einem Wahrscheinlichkeitsmaß

$$Z_N = \sum_{\sigma'} e^{-\beta H_N(\sigma')}.$$

Das Minuszeichen entspricht der physikalischen Konvention, dass Systeme versuchen, ihre Energie zu minimieren. Eine Möglichkeit, die zu einer möglichst gleichen Ausrichtung von Spins sorgt, ist es, interagierende Spins σ_i und σ_j miteinander zu multiplizieren. Im einfachsten Fall, dem so genannten Curie-Weiss-Modell, interagieren einfach alle Spins miteinander mit gleicher Stärke.

Dies führt zu der Energie- bzw. Hamiltonfunktion

$$H_N(\sigma) = -\frac{1}{2N} \sum_{1 \le i,j \le N} \sigma_i \sigma_j - h \sum_{i=1}^{N} \sigma_i.$$

Hierbei modelliert $h > 0$ ein externes Magnetfeld, dem die Spins ausgesetzt sind. Der große Vorteil des Curie-Weiss-Modells ist, dass die Energiefunktion eine Funktion des Ordnungsparameters m_N ist

$$H_N(\sigma) = -\frac{N}{2}[m_N(\sigma)]^2 - hN\, m_N(\sigma).$$

Diese Hamiltonfunktion hängt also von der mittleren Magnetisierung ab, man spricht auch von einem meanfield-Modell, einem Mittelwertmodell. Das zugehörige Gibbs-Mass ist von der Form

$$\mu_{N,\beta,h}(\sigma) = \frac{e^{\frac{\beta N}{2} m_N^2(\sigma) + N\beta h m_N(\sigma)}}{Z_{N,\beta,h}}$$

mit

$$Z_{N,\beta,h} = \sum_{\sigma} e^{\frac{\beta N}{2} m_N^2(\sigma) + N\beta h m_N(\sigma)}.$$

Um dies zu behandeln, erinnern wir an das LDP für i.i.d. Zufallsvariablen (X_i), die $Ber(p)$-verteilt sind auf $\{0,1\}$. Dann genügte $\frac{S_n}{n}$ einem LDP mit Geschwindigkeit n und Rate

$$H(x)\rho = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}.$$

Betrachtet man anstelle der (X_i) eine Folge (Y_i) mit

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = -1) = \frac{1}{2},$$

so entspricht dies einer Transformation

$$Y_i = 2X_i - 1.$$

Somit erhalten wir

$$\mathbb{P}(\frac{\sum_{i=1}^n Y_i}{n} = y) = \mathbb{P}(2S_n - 1 = y) = \mathbb{P}(S_n = \frac{y+1}{2}).$$

Wir sehen also, dass mit S_n auch $\frac{\sum_{i=1}^n Y_i}{n}$ einem LDP genügt und zwar mit Rate

$$\begin{aligned} I(y) = H(\frac{y+1}{2}|\frac{1}{2}) &= \frac{y+1}{2} \log \frac{y+1}{2} + \frac{1-y}{2} \log \frac{1-y}{2} + \log 2 \\ &= \frac{(1+y)}{2} \log \frac{(1+y)}{2} + \frac{(1-y)}{2} \log(1-y). \end{aligned}$$

Nun können wir Satz 3.12 und Satz 3.13 anwenden, um das Curie-Weiss-Modell zu studieren.

Satz 4.4 *Für jede inverse Temperatur $\beta > 0$ und jedes magnetische Feld h gilt*

$$\begin{aligned} f_{\beta,h} := \lim_{N\to\infty} -\frac{1}{\beta N}\log Z_{N,\beta,h} &= \inf_{m\in[-1,+1]}[-\frac{m^2}{2} - hm - \frac{1}{\beta}(\log 2 - I(m))] \\ &= -\sup_{m\in[-1,+1]}[\frac{m^2}{2} + hm + \frac{1}{\beta}(\log 2 - I(m))]. \end{aligned}$$

Hierbei ist

$$I(m) = \frac{1+m}{2}\log(1+m) + \frac{1-m}{2}\log(1-m).$$

Beweis: Es ist

$$Z_{N,\beta,h} = \sum_{\substack{\sigma_i\in\{\pm 1\}\\ \forall\, i=1,\dots,N}} e^{\frac{\beta N}{2}(\frac{\sum_{i=1}^N \sigma_i}{N})^2 + N\beta h(\frac{\sum_{i=1}^N \sigma_i}{N})} = 2^N \frac{1}{2^N} \sum_{\substack{\sigma_i\in\{\pm 1\}\\ \forall\, i=1,\dots,N}} e^{\frac{\beta N}{2}(\frac{\sum_{i=1}^N \sigma_i}{N})^2 + N\beta h(\frac{\sum_{i=1}^N \sigma_i}{N})}.$$

(Dies ist der Grund für das Auftreten des log 2-Terms im Ausdruck für $f_{\beta,h}$). Nun haben wir schon in einem Beispiel geklärt, dass $\frac{\sum_{i=1}^N \sigma_i}{N}$ unter dem Produktmaß einem LDP mit Geschwindigkeit N und Ratenfunktion $I(\cdot)$ genügt. Nun ist

$$F(x) = \frac{\beta x^2}{2} + \beta h x$$

natürlich im allgemeinen zwar stetig, aber nicht beschränkt. Nun lebt $\frac{\sum_{i=1}^N \sigma_i}{N}$ aber auf $[-1,+1]$. Dort ist $F(\cdot)$ sehr wohl beschränkt. Eine Anwendung des Varadhanschen Lemmas ergibt somit

$$\lim \frac{1}{N}\log(2^{-N} Z_{N,\beta,h}) = \sup_{m\in[-1,+1]}[\frac{\beta m^2}{2} + \beta h m - I(m)].$$

Dies ist äquivalent zu unserer Behauptung. □

Ebenso lässt sich zeigen, dass die Magnetisierung

$$m_N := \frac{1}{N}\sum_{i=1}^N \sigma_i$$

unter dem Gibbs-Maß

$$\mu_{N,\beta,h}(\sigma) = \frac{e^{-\beta H_N(\sigma)}}{Z_{N,\beta,h}}$$

einem LDP genügt.

Satz 4.5 *$m_N(\cdot)$ genügt unter dem Gibbs-Maß $\mu_{N,\beta,h}$ einem LDP mit Geschwindigkeit N und Ratenfunktion*

$$J(x) = -\frac{\beta x^2}{2} - \beta h x + I(x) + \sup_{y\in[-1,1]}[\frac{\beta y^2}{2} + \beta h y - I(y)],$$

wobei wieder

$$I(x) = \frac{1+x}{2}\log(1+x) + \frac{1-x}{2}\log(1-x)$$

ist.

Beweis: Dies folgt direkt aus den Sätzen über große Abweichungen, die wir zuvor bewiesen haben, da $\mu_{N,\beta,h}$ genau die dortige exponentielle Struktur besitzt. □

Das Schöne an Prinzipien der großen Abweichungen ist, dass sie auch Gesetze der großen Zahlen implizieren. Dieses Faktum haben wir schon in Wahrscheinlichkeitstheorie II kennengelernt. Es soll hier kurz wiederholt werden.

Satz 4.6 *Eine Folge von Zufallsvariablen $(X_n)_n$ in $\mathbb{R}^d$ genüge einem LDP mit Rate $I(\cdot)$ und Geschwindigkeit n. Dann gilt für die Menge der Nullstellen*

$$\mathcal{N} = \{x \in \mathbb{R}^d : I(x) = 0\}$$

der Ratenfunktion und jedes $\varepsilon > 0$

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \in \mathcal{N}_\varepsilon^c) < +\infty.$$

Hierbei ist

$$\mathcal{N}_\varepsilon = \{x : \|x - \mathcal{N}\| < \varepsilon\}.$$

Ist insbesondere $\mathcal{N}$ einelementig, so folgt für $\nu \in \mathcal{N}$

$$\mathbb{P}(X_n \to \nu) = 1.$$

Beweis: Da $\mathcal{N}_\varepsilon^c$ abgeschlossen ist, I von unten halbstetig und $\mathcal{N}$ die Menge der globalen Minima von I ist, folgt

$$a := \inf_{x \in \mathcal{N}_\varepsilon^c} I(x) > 0.$$

Aus der oberen Abschätzung der großen Abweichungen folgt, dass für hinreichend große n

$$\mathbb{P}(X_n \in \mathcal{N}_\varepsilon^c) \leq e^{-na/2}$$

gilt. Somit ist

$$\sum_{n} \mathbb{P}(X_n \in \mathcal{N}_\varepsilon^c) < +\infty.$$

Die fast sichere Konvergenz ist eine unmittelbare Konsequenz dieser Summierbarkeit und des Borel-Cantelli-Lemmas. □

Wir müssen uns somit, um die Minima von

$$J(x) = -\frac{\beta x^2}{2} - \beta h x + I(x) + \sup_{y \in [-1,+1]} [\frac{\beta y^2}{2} + \beta h y - I(y)]$$

mit

$$I(x) = \frac{1+x}{2} \log(1+x) + \frac{1-x}{2} \log(1-x)$$

kümmern. Diese Minima erfüllen also

$$I'(x) = \frac{1}{2} \log \frac{1+x}{1-x} = \beta(x+h),$$

also

$$e^{2\beta(x+h)} = \frac{1+x}{1-x}.$$

Durch langes Hinschauen erkennt man, dass dies äquivalent ist zu

$$x = \frac{e^{2\beta(x+h)} - 1}{e^{2\beta(x+h)} + 1} = \tanh(\beta(x+h)).$$

Man unterscheidet nun verschiedene Fälle:

- Ist $h > 0$, hat diese Gleichung zwei Lösungen, von denen aber nur die positive ein Minimum von $J(\cdot)$ liefert (die andere ein Maximum).
- Für $h < 0$ gibt es ebenfalls zwei Lösungen, von denen aber nur die negative ein Minimum, die positive aber ein Maximum ist.
- Für $h = 0$ ist die Situation symmetrisch zum Ursprung. Für $\beta \leq 1$ ist $x = 0$ die einzige Lösung dieser Gleichung und liefert somit ein Minimum von $J(\cdot)$. Für $\beta > 1$ hat die Gleichung allerdings drei Lösungen, von denen die Lösung $x = 0$ diesmal ein Maximum von $J(\cdot)$ liefert, die Lösungen, die verschieden sind von 0 aber Minima.

Zusammen erhält man:

Satz 4.7 *Im Curie-Weiss-Modell gelten für die Magnetisierung m_N die folgenden Grenzwertsätze unter $\mu_{N,\beta,h}$.*

1. *Ist $h > 0$, so konvergiert m_N exponentiell schnell gegen die positive Lösung von*
$$x = \tanh(\beta(x+h)).$$
2. *Ist $h < 0$, so konvergiert m_N exponentiell schnell gegen die negative Lösung von*
$$x = \tanh(\beta(x+h)).$$
3. *Ist $h = 0$ und $\beta \leq 1$, so konvergiert m_N exponentiell schnell gegen 0.*
4. *Ist $h = 0$ und $\beta > 1$, so konvergiert m_N exponentiell schnell gegen die beiden von Null verschiedenen Lösungen von*
$$x = \tanh(\beta x).$$

5. *Insbesondere erhält man*

$$\lim_{h\downarrow 0}\lim_{N\to\infty}\mu_{N,\beta,h}\circ m_N^{-1} \Rightarrow \delta_{m^*(\beta)},$$

wobei $m^*(\beta)$ die größte Lösung von

$$x = \tanh(\beta x)$$

ist.

5. ist in der physikalischen Literatur auch als "spontane Magnetisierung" bekannt: Ein Material, das in ein Magnetfeld gehalten wird, merkt sich dies bei genügend tiefen Temperaturen und wird selbst magnetisch. Bei hohen Temperaturen bleibt dieses Phänomen aus.

5 Der Satz von Sanov

In diesem Kapitel wollen wir ein LDP für eine maßwertige Zufallsgröße betrachten. Da sich aus diesem Satz der Satz von Cramér mittels des Kontraktionsprinzips herleiten lässt, ist das Prinzip, das wir zeigen wollen, gewissermaßen eine Ebene höher angesiedelt als der Satz von Cramér; man spricht auch von einem LDP auf Level II.

Die Ausgangssituation ist wieder die, dass wir eine Folge von i.i.d. Zufallsvariablen $(X_i)_{i\in\mathbb{N}}$ gegeben haben. Die X_i mögen auf einem polnischen Raum X mit Topologie $\mathcal{X}$ leben. Das empirische Maß

$$L_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

(wobei δ das Dirac-Maß ist) schickt X^n auf die Menge der Wahrscheinlichkeitsmaße $\mathcal{M}^1(X)$ auf X. L_n ist also unter dem Produktmaß $\mathbb{P}^n$, das die Verteilung der $(X_i)_{i=1}^n$ steuert, ein zufälliges Maß. Eine natürliche Topologie auf $\mathcal{M}^1(X)$ ist die schwache Topologie, die gröbste Topologie, so dass alle Abbildungen

$$\mu \mapsto \int f d\mu,\ f \in C^b(X)$$

stetig sind. Eine Basis dieser Topologie bilden die "ε-Kugeln"

$$B_\varepsilon^{f_1,\dots,f_N}(\mu) = \{\nu : |\int f_i d\nu - \int f_i d\mu| < \varepsilon \quad \forall\, i = 1,\dots,N\}.$$

Hierbei sind die $f_i \in C^b(X)$. Das Gesetz der großen Zahlen zeigt nun, dass

$$\int f_i dL_n = \frac{1}{n}\sum_{i=1}^{n} f_i(X_i) \to \mathbb{E}f_i(X_1) = \int f_j d\mathbb{P}$$

für alle $j = 1,\dots,N$ und da N endlich ist somit, dass $L_n \in B_\varepsilon^{f_1,\dots,f_N}(\mathbb{P})$ für alle $f_1,\dots,f_N$ und alle N. Somit konvergiert L_N schwach gegen $\mathbb{P}$. Wir wollen ein LDP für dieses Konvergenzresultat beweisen.

Satz 5.1 *Die Folge (L_n) genügt einem LDP in der schwachen Topologie auf $\mathcal{M}^1(X)$ mit Geschwindigkeit n und Ratenfunktion*

$$H(Q|\mathbb{P}) = \begin{cases} \int \log \frac{dQ}{d\mathbb{P}} dQ, & \text{falls } Q \ll \mathbb{P} \quad \text{und} \quad \log \frac{dQ}{d\mathbb{P}} \in \mathcal{L}^1|Q \\ +\infty & \text{sonst} \end{cases} .$$

Bemerkung 5.2 *Die Ratenfunktion $H(Q|\mathbb{P})$ haben wir im Diskreten schon für den Münzwurf kennengelernt. Dass sie hier wieder auftaucht, sollte nicht so sehr erstaunen, denn sowohl beim Münzwurf als auch bei δ_{X_i} handelt es sich um Zufallsvariablen mit den Werten 0 und 1. Manchmal werden wir die alternative Darstellung*

$$H(\mathbb{Q}|\mathbb{P}) = \int \frac{dQ}{d\mathbb{P}} \log \frac{dQ}{d\mathbb{P}} d\mathbb{P}$$

benutzen (dort wo $H(Q|\mathbb{P}) < \infty$).

Zunächst benötigen wir noch eine alternative Darstellung der Entropie. Dazu sei $\mathcal{B}$ die Borelsche σ-Algebra auf X und $\mathcal{B}(X)$ die Menge der beschränkten, messbaren Funktionen auf X. Dann gilt

Lemma 5.3 *Für $\alpha, \beta \in \mathcal{M}^1 X$ gilt*

$$H(\beta|\alpha) = \sup_{f \in \mathcal{B}(X)} [\int f d\beta - \log \int e^f d\alpha].$$

Beweis: Der Schlüssel zu dem Beweis ist eine Dualität von $x \log x - x + 1$ und $e^x - 1$ in dem Sinne, dass

$$\begin{aligned} x \log x - x + 1 &= \sup_y [xy - (e^y - 1)] \\ \text{und} \quad e^y - 1 &= \sup_x [xy - (x \log x - x + 1)] \end{aligned}$$

(das nachzurechnen, ist eine Übung). Nutzt man das mit $b(x) = \frac{d\beta}{d\alpha}(x)$, dann erhält man

$$\begin{aligned} \int f(x) d\beta &= \int f(x) b(x) d\alpha \\ &\leq \int [b(x) \log b(x) - b(x) + 1] + [e^{f(x)} - 1] d\alpha \\ &= H(\beta|\alpha) + \int e^{f(x)} - 1 d\alpha \\ &= H(\beta|\alpha) + \int e^{f(x)} d\alpha - 1. \end{aligned}$$

Nun ist mit $f \in \mathcal{B}(X)$ auch $f \pm c \in \mathcal{B}(X)$ für jede Konstante $c \in \mathbb{R}$. Wir schreiben $f = f + c - c$ und erhalten

$$\int f(x) + c d\beta = H(\beta|\alpha) + c + \int e^{f(x)} d\alpha - 1$$

und mit $g(x) = f(x) + c$

$$\int g(x)d\beta = H(\beta|\alpha) + c + \int e^{g(x)-c}d\alpha - 1.$$

Wählen wir $c = \log \int e^{g(x)}d\alpha$, so ergibt sich

$$\int g(x)d\beta \leq H(\beta|\alpha) + \log \int e^{g(x)}d\alpha$$

für alle $g \in \mathcal{B}(X)$, also

$$H(\beta|\alpha) \geq \sup_{f\in\mathcal{B}(X)} [\int f(x)d\beta - \log \int e^{f(x)}d\alpha].$$

Andererseits sei

$$C = \sup_{f\in\mathcal{B}(X)} [\int f(x)d\beta - \log \int e^{f(x)}d\alpha].$$

Dann ist

$$\int f(x)d\beta \leq C + \log \int e^{f(x)}d\beta$$

für alle $f \in \mathcal{B}(X)$. Setzt man $f(x) = \lambda \mathbb{1}_A(x)$ für ein $A \in \mathcal{B}$, so erhält man

$$\beta(A) \leq \frac{1}{\lambda}[C + \log[e^\lambda \lambda(A) + (1 - \alpha(A))]]$$

und mit $\lambda = -\log \alpha(A)$

$$\beta(A) \leq \frac{C+2}{\log \frac{1}{\alpha(A)}}.$$

Somit ist für $\alpha(A) = 0$ auch $\beta(A) = 0$, also $\beta \ll \alpha$. Setze nun $b(x) = \frac{d\beta}{d\alpha}(x)$, dann erhält man

$$\int f(x)b(x)d\beta \leq C + \log \int e^{f(x)}d\beta.$$

Wählt man $f(x) = \log b(x)$ und führt ein Abschneideargument durch, falls $b(x)$ von oben unbeschränkt ist, so erhält man

$$H(\beta|\alpha) \leq C,$$

also die fehlende Richtung. □

Das Supremum im vorhergendenden Satz lässt sich auch allein über die Menge der stetigen und beschränkten Funktionen $C^b(X)$ bilden. Dazu benötigt man

Satz 5.4 ***(Lusin):*** *Ist* $(X, \mathcal{X}, \mathbb{P})$ *ein Maßraum und*

$$f : X \to \mathbb{R}$$

eine beschränkte, messbare Abbildung. Dann gibt es zu jedem $\varepsilon > 0$ *eine messbare Menge* $A_\varepsilon \in \mathcal{X}$*, sodass*

$$f_\varepsilon := f|_{A_\varepsilon}$$

eine stetige Funktion ist. A_ε *lässt sich als abgeschlossene Menge wählen.*

Um die Funktion f_ε wieder als stetige Funktion auf ganz X ansehen zuz können, benötigt man noch

Satz 5.5 ***(Tietze-Urysohn).*** *Sei X ein normaler topologischer Raum $A \subset X$ abgeschlossen und*

$$f : A \to \mathbb{R}$$

stetig. Dann gibt es eine stetige Fortsetzung

$$F : X \to \mathbb{R}$$

mit $F|_A = f$ und $\|F\|_\infty = \|f\|_\infty$.

Damit erhält man

$$H(\beta|\alpha) = \sup_{f \in C^b(X)} [\int f(x) d\beta(x) - \log \int e^{f(x)} d\alpha].$$

Wir kommen nun zum

Beweis von Satz 5.1: Wir beginnen damit zu zeigen, dass $H(\cdot|\mathbb{P})$ eine Ratenfunktion ist. Da wieder

$$H(Q|\mathbb{P}) = \int \psi \left(\frac{dQ}{d\mathbb{P}} \right) d\mathbb{P}$$

mit $\psi(x) = x \log x - x + 1$ und ψ strikt konvex ist, folgt, dass $H(\cdot|\mathbb{P})$ strikt konvex und unterhalbstetig ist mit $H(Q|\mathbb{P}) = 0 \Leftrightarrow Q = \mathbb{P}$. Sei nun

$$D_L := \{Q : H(Q|\mathbb{P}) \leq L\}.$$

Da H halbstetig ist, ist D_L abgeschlossen. Um zu zeigen, dass es auch kompakt ist, genügt es nach dem Satz von Prokorov für jedes $\varepsilon > 0$ eine kompakte Menge K_ε vorzuzeigen, sodass $Q(K_\varepsilon^c) < \varepsilon$ für alle $A \in D_L$. Nun gibt es eine kompakte Menge K_L, sodass $\mathbb{P}(K_L^c) \leq e^{-(L+2)/\varepsilon}$ ist. Aus der im vorigen Beweis gewonnenen Abschätzung erhalten wir dann, dass

$$Q(K_L^c) \leq \frac{L+2}{\log \frac{1}{\mathbb{P}(K_L^c)}} \leq \varepsilon$$

gilt.

Für die obere Abschätzung seien $Q \in \mathcal{M}^1(X)$ und $\varepsilon > 0$ beliebig. Wir zeigen, dass es eine kleine offene Umgebung U_Q von Q gibt, sodass

$$\overline{\lim} \frac{1}{n} \log \mathbb{P}(L_n \in U_Q) \leq -H(Q|\mathbb{P}) + 2\varepsilon.$$

Zunächst wählen wir ein $f \in C(X)$, sodass

$$\int f(x) dQ - \log \int e^{f(x)} d\mathbb{P} \geq H(Q|\mathbb{P}) - \varepsilon.$$

Nun ist

$$\mathbb{E}[n\int f(x)dL_n] = [\mathbb{E}\int e^f d\mathbb{P}]^n.$$

Wählen wir

$$U_Q = \{R : |\int f dQ - \int f dR| < \varepsilon\},$$

so folgt mit dem exponentiellen Chebyshev:

$$\begin{aligned}\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{P}(L_n \in U_Q) &\leq -\int f dQ + \varepsilon + \log\int e^{f(x)}d\mathbb{P}\\ &\leq -H(Q|\mathbb{P}) + 2\varepsilon.\end{aligned}$$

Ist nun $D \subset \mathcal{M}^1(X)$ kompakt, dann können wir es mit einer endlichen Anzahl von solchen $U_Q,\ Q \in D$ überdecken. Aus dem Prinzip des maximalen Terms erhalten wir dann

$$\overline{\lim}\,\frac{1}{n}\log\mathbb{P}(L_n \in D) \leq -\inf_{Q\in D} H(Q|\mathbb{P}) + 2\varepsilon$$

und da $\varepsilon > 0$ beliebig auch

$$\overline{\lim}\,\frac{1}{n}\log\mathbb{P}(L_n \in D) \leq -\inf_{Q\in D} H(Q|\mathbb{P}).$$

Wenn wir auch die exponentielle Straffheit zeigen können, also dass es zu jedem $L < \infty$ ein kompaktes $D_l \subset \mathcal{M}^1 X$ gibt, sodass

$$\overline{\lim}\frac{1}{n}\log\mathbb{P}(L_n \in D_L^c) \leq -L,$$

dann bekommen wir die obere Abschätzung auch für beliebige abgeschlossene Mengen A, indem wir

$$\mathbb{P}(L_n \in A) = \mathbb{P}(L_n \in A, D_L) + \mathbb{P}(L_n \in A \cap D_L^c)$$

schreiben und $L \to \infty$ gehen lassen. Wählen wir nun Folgen $(\varepsilon_j),\ (\delta_j)$ und $(\vartheta_j)_j$, so dass $\delta_j = \frac{1}{j},\ \vartheta_j = j(L + \log 2 + j)$ und $\varepsilon_j = e^{-\vartheta_j}$ (bemerke, dass $\varepsilon_j \downarrow 0,\ \delta_j \downarrow 0$ und $\vartheta_j \to \infty$) und Mengen K_j, sodass

$$\mathbb{P}(K_j^c) \leq \varepsilon_j.$$

Sei

$$D = \{Q : Q(K_j^c) \leq \delta_j \quad \forall\, j\}.$$

Wieder ist D nach dem Satz von Prohorov kompakt. Es gilt

$$\begin{aligned}\mathbb{P}(L_n \in D^c) &\leq \sum_{j=1}^{\infty}\mathbb{P}(L_n(K_j^c) \geq \delta_j)\\ &= \sum_{j=1}^{\infty}\mathbb{P}(\sum_{k=1}^{n}\delta_{X_k}(K_j^c) \geq n\delta_j)\\ &\leq \sum_{j=1}^{\infty}[\varepsilon_j e^{\vartheta_j} + (1-\varepsilon_j)]^n e^{-n\vartheta_j\delta_j}\end{aligned}$$

wieder nach dem exponentiellen Chebyshev. Nach Wahl von $\varepsilon_j, \delta_j, \vartheta_j$ ist daher

$$\mathbb{P}(L_n \in D^c) \leq e^{-Ln}.$$

Dies zeigt die obere Schranke.

Die untere Schranke benutzt den schon bekannten exponentiellen Maßwechsel. Es sei G eine offene Menge in $\mathcal{M}^1(X)$ und $Q \in G$. OBdA existiere die Dichte $\frac{dQ}{d\mathbb{P}}$, denn sonst ist die Behauptung trivial. Somit ist dann für eine offene Umgebung $U_Q \subseteq G$ von Q

$$\begin{aligned}\mathbb{P}(L_n \in G) &\geq \mathbb{P}(L_n \in U_Q) \\ &= \int_{\{L_n \in U_Q\}} \left(\frac{dQ}{d\mathbb{P}}\right)^{-1}(x_1), \ldots, \left(\frac{dQ}{d\mathbb{P}}\right)^{-1}(x_n) dQ(x_1), \ldots, dQ(x_n).\end{aligned}$$

Falls $\frac{dQ}{d\mathbb{P}} = 0$ auf einer Menge positiven Maßes, so fällt diese nicht ins Gewicht, denn

$$Q\left(\frac{dQ}{d\mathbb{P}} = 0\right) = \int_{\frac{dQ}{d\mathbb{P}}=0} dQ = \int_{\frac{dQ}{d\mathbb{P}}=0} \frac{dQ}{d\mathbb{P}} d\mathbb{P} = 0.$$

Somit ergibt sich

$$\begin{aligned}\mathbb{P}(L_n \in G) &\geq \int_{L_n \in U_Q | \int \log \frac{dQ}{d\mathbb{P}} dL_n - H(Q|\mathbb{P}| < \varepsilon} \left(\frac{dQ}{d\mathbb{P}}\right)^{-1}(x_1), \ldots, \left(\frac{dQ}{d\mathbb{P}}\right)^{-1}(x_n) dQ(x_1), \ldots, dQ(x_n) \\ &\geq e^{-n(H(Q|\mathbb{P})+\varepsilon)} \int_{L_n \in U_Q | \int \log \frac{dQ}{d\mathbb{P}} dL_n - H(Q|\mathbb{P})| \varepsilon} dQ(x_1, \ldots, dQ(x_n) \\ &\to e^{-n(H(Q|\mathbb{P})+\varepsilon)}\end{aligned}$$

wenn $n \to \infty$, nach dem Gesetz der großen Zahlen. Logarithmieren und durch n teilen ergibt die Behauptung. □

6 Anwendungen des Satzes von Sanov

Hier wollen wir Konsequenzen aus dem Satz von Sanov vorstellen. Wir beginnen mit der schon versprochenen Begründung, warum der Satz von Sanov ein LDP auf Level II genannt wird. Wir leiten den Satz von Cramér her, allerdings gleich für Banachraum-wertige Zufallsvariablen.

Satz 6.1 *Seien $X_1, X_2, \ldots$ i.i.d. Zufallsvariablen mit Werten in einem separablen Banachraum X und gemeinsamer Verteilung $\mathbb{P}$. Sei*

$$\mathbb{E}e^{\vartheta \|X_1\|} < \infty \quad \textit{für alle} \quad \vartheta > 0.$$

Dann genügt $\frac{1}{n}\sum_{i=1}^n X_i$ einem LDP mit Geschwindigkeit n und Ratenfunktion

$$H(x) = \sup_{y \in X^*} [< y, x > - \log \int e^{<y,x>} d\mathbb{P}].$$

Hierbei ist X^* der Dualraum von X und $<,>$ die duale Paarung.

Die Kernidee wird es natürlich sein, den Satz von Sanov für die $(X_i)_i$ zu verwenden. Tatsächlich ist ja mit der Notation des vorangegangenen Kapitels

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \int z dL_n(z).$$

Allerdings kann die Identität nur dann als stetige und beschränkte Funktion angesehen werden, wenn die X_i beschränkt sind. Dies ist eine der beiden Schwierigkeiten. Die andere ist, dass, selbst wenn man das Kontraktionsprinzip anwenden kann, aus diesem eine Rate der Form

$$I(x) = \inf_{Q:\int z dQ = x} H(Q|\mathbb{P})$$

bekommt. Das ist ja nicht ganz $H(x)$. Im ersten Schritt skizzieren wir, warum $H(x)$ und $I(x)$ doch gleich sind.

Lemma 6.2 *Sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf X mit*

$$\int e^{\vartheta \|z\|} d\mathbb{P} < \infty \quad \forall\, \vartheta > 0.$$

Sei

$$M(y) = \int e^{<y,z>} d\mathbb{P}(z)$$

und

$$H(x) = \sup_{y \in X}[< y, x > - \log M(y)].$$

Dann ist

$$\begin{aligned} H(x) &= \inf_{Q:\int z dQ = x} H(Q|\mathbb{P}) \\ &= \inf \int f(z) \log f(z(d\mathbb{P}(z) \end{aligned}$$

$$f = \frac{dQ}{d\mathbb{P}} : \int z f(z) d\mathbb{P}(z) = x.$$

Der Beweis des Lemmas wiederum beruht auf einem Minimax-Prinzip aus der konvexen Analysis, das wir hier ohne Beweis angeben. Dazu sei $F(x, y)$ eine konvexe, unterhalb stetige Funktion

$$F : X \times Y \to \mathbb{R}$$

(X, Y metrische Räume). Es seien $C_1 \leq X$ und $X_2 \leq Y$ abgeschlossen. Sind endweder C_1 oder C_2 kompakt oder alle Miveaumengen

$$\begin{aligned} D_X^L &= \{y : F(x, y) \leq L\} \quad \text{oder} \\ D_y^L &= \{x : F(x, y) \geq L\} \end{aligned}$$

für alle $L \in \mathbb{R}$ kompakt, so gilt:

$$\inf_{y \in C_2} \sup_{x \in C_1} F(x,y) = \sup_{x \in C_1} \inf_{y \in C_2} F(x,y).$$

Beweis von Lemma 6.2 (Skizze): Betrachte die Funktion

$$F(y, f(\cdot)) = < y, x > - \int < y, z > f(z) d\alpha(z) + \int f(z) \log f(z) d\alpha(z)$$

auf $X^* \times \mathcal{N}$, wobei $\mathcal{N}$ die Menge aller Dichten bezüglich α mit

$$\int f(z) \log f(z) d\alpha < \infty$$

ist. Man rechnet aus, dass

$$\sup_y \inf_f F(y, f(\cdot)) = \sup_y [< y, x > - \log M(y)] = H(x)$$

ist, indem man zeigt, dass $H(Q|\mathbb{P})$ eine entsprechende Variationsformel erfüllt. Andererseits ist

$$\sup_y [< y, x > - \int < y, z > d\alpha(z)] = \infty,$$

falls nicht $x = \int z d\alpha(z)$ (und indem Fall ist es 0). Also gilt

$$\inf_f \sup_y F(y, f(\cdot)) = \inf_{f : \int z f(z) d\alpha = x} \int f(z) \log f(z) d\alpha.$$

Man überprüft nun, dass sich der oben zitierte Satz aus der konvexen Analysis anwenden lässt und ist fertig. □

Beweis von Satz 6.1:
1. Schritt: Wir beginnen mit beschränkten Zufallsvariablen. Sei also $\|X_1\| \leq L$. Angenommen, β_n ist eine Folge von Wahrscheinlichkeitsmaßen mit Träger in $B_L(0)$ und $\beta_n \to \beta$ schwach. Setzt man

$$x_n := \int z d\beta_n,$$

so gilt $< y, x_n > \to < y, x > = \int z d\beta$ für alle $y \in X$. Gemäß dem Satz von Prohorov gibt es eine kompakte Menge K_2, so dass $\beta_n(K_\varepsilon) \geq 1 - \varepsilon$ für alle n (und alle $\varepsilon > 0$) und $\beta(K_\varepsilon) \geq 1 - \varepsilon$. Setzen wir

$$x_n^\varepsilon := \int_{K_\varepsilon} z d\beta_n,$$

dann ist x_n in der konvexen Hülle von $\{0\}$ und K_ε und $\|x_n^\varepsilon - x_n\| \leq \varepsilon \ell$. Also folgt

$$\|x_n - x\| \to 0.$$

2. Schritt: Wir zerlegen X_i in einen beschränkten und einen unbeschränkten Anteil. Also sei für $L > 0$

$$X_i = Y_i^L + Z_i^L$$

mit $Y_i^L := X_i \mathbb{1}_{\{\|X_i\| \leq L\}}$. Beachte, dass $Z_i^L \to 0$ für $L \to \infty$ und alle i und daher

$$\mathbb{E}e^{\vartheta \|Z_i^L\|} \to 1 \quad \text{für alle } i,\ \vartheta > 0,$$

wenn $L \to \infty$. Wir können daher die exponentielle Chebyshev-Ungleichung benutzen, um

$$\begin{aligned} &\limsup_{L\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\frac{1}{n}\sum_{j=1}^n \|Z_j^L\| \geq \varepsilon\right] \\ \leq\ &\limsup_{L\to\infty} \inf_{\vartheta>0} [-\vartheta\varepsilon + \log \mathbb{E}e^{\vartheta\|Z_i^L\|}] \\ \leq\ &\inf_{\vartheta>0} [\vartheta\varepsilon + \frac{\varepsilon}{2}] \end{aligned}$$

für L hinreichend groß abzuschätzen. Dies zeigt aber, dass

$$\overline{\lim_{L\to\infty}}\,\overline{\lim_{n\to\infty}}\ \frac{1}{n} \log \mathbb{P}[\frac{1}{n}\sum_{j=1}^n \|Z_j^L\| \geq \varepsilon] = -\infty$$

gilt.

3. Schritt: Hier fügen wir die vorbereitenden Schritte zusammen. Sei $C \subseteq X$ abgeschlossen und

$$C^\varepsilon = \{y \in X : d(y, C) \leq \varepsilon\}.$$

Dann gilt:

$$\mathbb{P}\left[\frac{1}{n}\sum_{j=1}^n X_j \in C\right] \leq \mathbb{P}\left[\frac{1}{n}\sum_{j=1}^n Y_j^L \in C^\varepsilon\right] + \mathbb{P}\left[\frac{1}{n}\sum_{j=1}^n \|Z_j^L\| \geq \varepsilon\right].$$

Setzen wir

$$H^L(x) := \sup_{y\in X} [< y, x > - \log \mathbb{E}e^{<y, Y^L>}],$$

so ergibt sich nach dem Kontraktionsprinzip, dem Satz von Sanov und Lemma 6.2:

$$\frac{1}{n} \log \mathbb{P}\left[\frac{1}{n}\sum_{j=1}^n Y_j^L \in C^\varepsilon\right] = \frac{1}{n} \log \mathbb{P}[\int x dL_N(Y^L) \in C^\varepsilon] \to \inf_{y\in C^\varepsilon} \inf_{Q:\int x dQ = y} H(Q|\mathbb{P}^L).$$

Hierbei ist $L_N(Y^L)$ das empirische Maß der $Y_1^L, \ldots, Y_n^L$ und $\mathbb{P}^L$ ihre Verteilung. Nach Lemma 6.2 ist die rechte Seite gleich

$$= - \inf_{y\in C^\varepsilon} H^L(y).$$

Somit

$$\overline{\lim_{n\to\infty}}\ \frac{1}{n} \log \mathbb{P}\left[\frac{1}{n}\sum_{j=1}^n X_j \in C\right] \leq \max\{- \inf_{x\in C^\varepsilon} H^L(x), \overline{\lim_{n\to\infty}} \log \mathbb{P}\left[\frac{1}{n}\sum_{j=1}^n \|z_j^L\| > \varepsilon\right].$$

Dies gilt für alle $L > 0$ und alle $\varepsilon > 0$. Lässt man L groß werden, sieht man, dass das Maximum im ersten Term angenommen wird und wir bekommen

$$\begin{aligned}
&\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}\left[\frac{1}{n}\sum_{j=1}^{n} X_j \in C\right] \\
&\leq -\varliminf_{\varepsilon\to 0} \overline{\lim_{L\to\infty}} \inf_{x\in C^\varepsilon} H^L(x) \\
&= -\inf_{x\in C} H(x).
\end{aligned}$$

Dieser letzte Schritt bedarf noch einer Begründung. Können wir diese liefern, so sind wir fertig, denn dann geht die untere Schranke analog. Die Begründung folgt im anschließenden Lemma. □

Die Idee dieses Lemmas ist es, dass wir uns für L die $X_L^\varepsilon \in C^\varepsilon$ anschauen und zeigen, dass aus der Beschränktheit von $H^L(x_L)$ folgt, dass $(x_L^\varepsilon)_L$ eine kompakte Folge ist, derart dass jeder Häufungspunkt x_ε in C_ε liegt und

$$H(x) \leq \liminf_{\varepsilon\to 0} \liminf_{L\to\infty} H_L(x_L^\varepsilon)$$

erfüllt. Dazu benötigen wir

Lemma 6.3 *Für eine Indexmenge I sei $(\mu_i)_{i\in I}$ eine straffe Folge von Wahrscheinlichkeitsmaßen auf X, dergestalt dass*

$$\sup_{i\in I} \int e^{\vartheta\|z\|} d\mu_i(z) = m(\vartheta) < \infty$$

für alle $\vartheta > 0$. Es sei

$$H_i(x) = \sup_{y\in X^x} [< y, x > - \log \int e^{<y,z>} d\mu_i(z)].$$

Dann hat für jedes L die Menge

$$\bigcup_{i\in I} \{x : H_i(x) \leq L\}$$

einen kompakten Abschluss in X.

Beweis: Aufgrund der Darstellung

$$H_i(x) = \inf_{\substack{Q:\frac{dQ}{d\mu_i}=f \\ \int zf(z)d\mu_i(z)=x}} H(Q||\mu_i)$$

gibt es für alle x_i mit $H_i(x_i) \leq L$ eine Dichte f_i, so dass

$$x_i = \int z f_i(z) d\mu_i(z) =: \int z d\nu_i(z)$$

und

$$\int f_i(z) \log f_i(z) d\mu_i(z) \leq L.$$

Nun ist die Folge der μ_i straff, d. h. für alle $\varepsilon > 0$ gibt es ein K_ε mit $\mu_i(K_\varepsilon) \geq 1 - \varepsilon$ für alle $i \in I$. Da aber $H(\nu_i|\mu_i) \leq L$ ist, bedeutet dies, dass auch die Folge der $(\nu_i)_i$ straff ist. In der Tat: Wäre ν_i nicht straff, so gäbe es für alle $\varepsilon > 0$ eine Konstante c_ε mit $c_\varepsilon \to 0$, sodass $\nu_i(K_\varepsilon) \leq c_\varepsilon \mu_i(K_\varepsilon)$ für mindestens ein i gilt. Dann ist aber

$$\begin{aligned} H(\nu_i|\mu_i) &= \int_{K_\varepsilon} \frac{d\nu_i}{d\mu_i} \log \frac{d\nu_i}{d\mu_i} d\mu_i + \int_{K_\varepsilon^c} \frac{d\nu_i}{d\mu_i} \log \frac{d\nu_i}{d\mu_i} d\mu_i \\ &\geq -\frac{1}{e} + \int_{K_\varepsilon^c} \frac{1 - c_\varepsilon(1-\varepsilon)}{\varepsilon} \log \frac{1 - c_\varepsilon(1-\varepsilon)}{\varepsilon} \end{aligned}$$

wegen $x \log x \geq -\frac{1}{e}$ und aufgrund der Konvexität der Entropie. Der 2. Summand divergiert aber für $\varepsilon \to 0$.

Somit ist $(\nu_i)_i$ relativ kompakt nach dem Satz von Prohorov. Ebenso überlegt man sich, dass $\int \|z\| d\nu_i(z)$ gleichgradig integrierbar ist, was die Kompaktheit der Folge (x_i) impliziert (das ist eine Übung). □

Eine weitere wichtige Anwendung des Satzes von Sanov finden wir beim sogenannten Gibbsschen Konditionieren. Die Fragestellung hierbei stammt (wie der Name Gibbs vermuten lässt) aus der statistischen Mechanik, hat aber auch als statische Frage ihre Berechtigung. In der statistischen Mechanik unterscheidet man, je nachdem, ob man die Energie und die Teilchenzahl eines Systems festhält oder veränderlich lässt, zwischen dem mikrokanaischen, kanaischen und makrokanaischen Ensemble. Eine wichtige Frage ist: Was weiß man über die Energie eines einzelnen Teilchens, wenn man die Energie des Gesamtsystems kennt? Dies lässt sich so mathematisieren: Es sei $Y_1, \ldots, Y_n$ eine Folge von i.i.d. Zufallsvariablen mit Verteilung μ auf einer endlichen Menge X (und o.B.d.A. $\mu(i) > 0 \quad \forall\, i \in X$). Weiter sei

$$f : X \to \mathbb{R}$$

eine Abbildung und $A \subseteq \mathbb{R}$. Wir fragen nun nach der bedingten Verteilung

$$\mu_n^{(A)}(x) := \mathbb{P}(Y_1 = x | \int f dL_n \in A)$$

(beachte, dass $\int f dL_n = \frac{1}{n} \sum_{i=1}^n f(Y_i)$ ist). Die Schwierigkeit dieser Fragestellung stammt daher, dass die Bedingung $\int f dL_n \in A$ zwar die identische Verteilung der (Y_i) erhält, aber nicht ihre Unabhängigkeit (das ist ähnlich wie beim Ziehen ohne Zurücklegen). Immerhin lässt die identische Verteilung der $(Y_i)_i$ noch die folgende Rechnung für jede Testfunktion $g : X \to \mathbb{R}$ zu:

$$\begin{aligned} \int g \, d\mu_n(A) &= \mathbb{E}[g(Y_1) | \int f \, dL_n \in A] \\ &= \mathbb{E}[\frac{1}{n} \sum_{i=1}^n g(Y_i) | \int f \, dL_n \in A] \\ &= \mathbb{E}[\int g \, dL_n | \int f \, dL_n \in A]. \end{aligned}$$

Also ist $\mu_n^{(A)} = \mathbb{E}[L_n | L_n \in \Sigma_A]$, wobei

$$\Sigma_A = \{\nu \in \mathcal{M}^1(X) : \int f \, d\nu \in A\}$$

ist. Es ist naheliegend, dass sich Fragen über $\mu_n^{(A)}$ mithilfe des Satzes von Sanov beantworten lassen. Der folgende Satz beschreibt die Menge aller Häufungspunkte der Folge $(\mu_n^{(A)})$ für Mengen A, sodass Σ_A Entropie-stetig ist, d. h. sodass

$$\inf_{\nu \in \Sigma_A} H(\nu|\mu) = I_{\Sigma_A} = \inf_{\nu \in \Sigma_A^0} H(\nu|\mu) = \inf_{\nu \in \Sigma_A} H(\nu|\mu) \tag{17}$$

gilt.

Satz 6.4 ***(Gibbssches Prinzip).*** *Es sei A so, dass* (17) *erfüllt ist und sei*

$$\mathcal{M} = \{\nu \in \Sigma_A : H(\nu|\mu) = I_{\sum_A}\}$$

die Menge der "Lösungen" von (17)*. Dann gilt:*

a) Die Menge der Häufungspunkte von $(\mu_n^{(A)})_n$ ist enthalten in $\overline{co(\mathcal{M})}$, dem Abschluss der konvexen Hülle von $\mathcal{M}$.

b) Ist $\sum_A$ konvex und $\sum_A^0 \neq \emptyset$, dann besteht $\mathcal{M} = \{\mu^*\}$ aus einem einzigen Punkt und

$$\mu_n \to \mu^*, \quad n \to \infty.$$

Bemerkung 6.5 *a) Da $\mathcal{M}$ kompakt ist, lässt sich auch $co(\mathcal{M})$ statt $\overline{co(\mathcal{M})}$ schreiben.*

b) $co(\mathcal{M})$ ist aber notwendig. Nimmt man die $(Y_i)_i$ als i.i.d. Ber($\frac{1}{2}$)-verteilte Zufallsvariablen, f als die Identität und

$$A = [0, \frac{1}{4}] \cup [\frac{3}{4}, 1],$$

so besteht $\mathcal{M}$ (aufgrund der Konvexität von $H(\cdot|Ber(\frac{1}{2}))$ und der Symmetrie aus den beiden Verteilungen Ber($\frac{1}{4}$) und Ber($\frac{3}{4}$). Nun ist aber sowohl μ als auch A symmetrisch bzgl. $\frac{1}{2}$. Der einzige mögliche Häufungspunkt von $\mathbb{E}[L_n|L_n \in A]$ ist daher die Ber($\frac{1}{2}$)-Verteilung. Diese liegt nicht in $\mathcal{M}$, wohl aber in $co(\mathcal{M})$.

Beweis von Satz 6.4: Bemerke, dass X, also auch $\mathcal{M}^1(X)$ kompakt sind. Da $H(\cdot|\mu)$ unterhalb stetig ist, ist $\mathcal{M}$ nicht-leer. Eine Übung ist es zu sehen, dass aus der Tatsache, dass $\sum_A$ konvex ist und $\sum \neq \emptyset$ folgt, dass $\mathcal{M}$ einelementig ist (dies benutzt die strikte Konvexität von $H(\cdot|\mu)$). Dies zeigt Teil b).

Für Teil a) sei d der Abstand der totalen Variation auf $\mathcal{M}_1(X)$, also

$$d(\nu, \nu') = \frac{1}{2} \sum_{x \in X} |\nu(x) - \nu'(x)|.$$

Wir wählen ein $\delta > 0$ und setzen U_δ als die δ-Umgebung von $\mathcal{M}$, also

$$U_\delta := \{\nu : d(\nu, \mathcal{M}) < \delta\}.$$

Wir wollen zeigen, dass

$$\mathbb{P}(L_n \in U_\delta | L_n \in \Sigma_A) \to 1 \tag{18}$$

gilt. Der Satz von Sanov für endliche Zustandsräume liefert nun

$$\begin{aligned} &\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in U_\delta^c \cap \Sigma_A) \\ \leq &\limsup_{n\to\infty} \frac{1}{2} \log \mathbb{P}(L_n \in U_\delta^c \cap \bar{\Sigma}_A) \\ \leq &- \inf_{\nu \in U_\delta^c \cap \bar{\Sigma}_A} H(\nu|\mu). \end{aligned} \tag{19}$$

Andererseits folgt ebenfalls aus dem Satz von Sanov und der Entropie-Stetigkeit von Σ_A (17), dass

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in \Sigma_A) = -I_{\Sigma_A}.$$

Wegen der Kompaktheit von $\mathcal{M}^1(X)$ und der Abgeschlossenheit von $\bigcup_\delta^c$ und Σ_A wird das Infimum in (19) auch angenommen und zwar in einem Punkt $\nu \notin \mathcal{M}$. Somit ist das Infimum in (19) strikt kleiner als I_{Σ_A}. Dies aber bedeutet, dass

$$\begin{aligned} \mathbb{P}(L_n \notin U_\delta | L_n \in \Sigma_A) &= \mathbb{P}(L_n \in U_\delta^c \cap \Sigma_A) | \mathbb{P}(L_n \in \Sigma_A) \\ &\approx \exp(-n \inf_{\nu \in U_\delta^c \cap \Sigma_A} H(\nu|\mu) - I_{\Sigma_A}) \\ &\to 0 \end{aligned}$$

exponentiell schnell. Also folgt (18). Andererseits impliziert (18) aber

$$\lim_{n\to\infty} d(\mu_n^{(A)}, co(U_\delta)) = 0. \tag{20}$$

Dies sieht man wie folgt: Es gilt

$$\begin{aligned} &\mu_n^{(A)} - \mathbb{E}[L_n | L_n \in U_\delta \cap \Sigma_A] = \mathbb{E}[L_n | L_n \in \Sigma_A] - \mathbb{E}[L_n | L_n \in U_\delta \Sigma_A] \\ = &\mathbb{P}(L_n \in U_\delta^c | L_n \in \Sigma_A)(\mathbb{E}(L_n | L_n \in U_\delta^c \cap \Sigma_A) - \mathbb{E}(L_n | L_n \in U_\delta \cap \Sigma_A)) \end{aligned} \tag{21}$$

(die letzte Gleichung ist eine kleine Übung zum Rechnen mit bedingten Wahrscheinlichkeiten). Da nun $\mathbb{E}[L_n | L_n \in U_\delta \cap \Sigma_A] \in co(U_\delta)$, folgt, dass der Abstand von μ_n^* zu $co(U_\delta)$ durch die rechte Seite von (21) beschränkt ist. Nun sind $\mathbb{E}[L_n | L_n \in U_\delta^c \cap \Sigma_A)$ und $\mathbb{E}(L_n | L_n \in U_\delta \cap \Sigma_A)$ Wahrscheinlichkeitsmaße auf X, also ist ihr Abstand höchstens 1. Daher ist

$$d(\mu_n^{(A)}, co(U_\delta)) \leq d(\mu_n^*, \mathbb{E}[L_n | L_n \in U_\delta \cap \Sigma]) \leq \mathbb{P}(L_n \in U_\delta^c | L_n \in U) \to 0.$$

Das zeigt (20). Da $\delta > 0$ beliebig war, liefert das die gesuchte Behauptung. □

7 Der Satz von Schilder

Wir haben im ersten Abschnitt sehr allgemein definiert, was ein LDP ist, und im zweiten abstrakte Sätze über solche Prinzipien hergeleitet. Ein Vorteil des Abstraktonsgrades liegt darin, dass man die Prinzipien auch für komplexe Situationen, beispielsweise unendlich-dimensionale Räume, herleiten kann. Eine solch unendlich dimensionale Situation ist die Brownsche Bewegung (sie besteht aus unendlich vielen Zeitpunkten, die man gleichzeitig kontrollieren muss). Sei $(\eta(t))_{t\in[0,1]}$ diese Brownsche Bewegung. Es ist intuitiv klar, dass man, wenn man diesen Prozess aus großer Distanz betrachtet, also

$$x_\varepsilon(t) = \sqrt{\varepsilon}\beta(t) = \beta(\varepsilon t)$$

studiert, "im Wesentlichen" eine Null-Linie sieht. Tatsächlich ist ja für ein festes t

$$\begin{aligned}\mathbb{P}(|x_\varepsilon(t)| \geq x) &= 2\mathbb{P}(x_\varepsilon(t) \geq x)\\ &= 2\cdot\int_x^\infty \frac{1}{\sqrt{2\pi\varepsilon}} e^{-y^2/2\varepsilon} dy\\ &\approx \sqrt{\frac{2}{\pi\varepsilon}} e^{-x^2/2\varepsilon},\end{aligned}$$

was für $\varepsilon \to 0$ sehr schnell gegen 0 geht. Eines der historisch ältesten Resultate über große Abweichungen beschäftigt sich mit der Wahrscheinlichkeit, dass der Prozess $(x_\varepsilon(t))_{t\in[0,1]}$ für $\varepsilon \to 0$ wesentlich von 0 verschieden ist. Dies ist der Satz von Schilder. Dazu sei Q_ε die Verteilung von $(x_\varepsilon(\cdot))$.

Satz 7.1 ***(Schilder):*** *Die Folge Q_ε genügt einem Prinzip der großen Abweichungen für $\varepsilon \to 0$ mit Geschwindigkeit ε^{-1} und Ratenfunktion*

$$I(f) = \begin{cases} \frac{1}{2}\int_\infty^1 (f'(t))^2 dt & f \in \mathcal{A} \\ +\infty & sonst \end{cases}$$

Dabei ist $\mathcal{A}$ die Menge der absolut stetigen Funktionen auf $[0,1]$ mit quadratisch-integrierbarer Ableitung und $f(0) = 0$.

Erinnerung: $f : [0,1] \to \mathbb{R}$ heißt absolut stetig genau dann, wenn für alle $\varepsilon > 0$ ein $\delta > 0$ existiert, sodass für jedes disjunkte System offener Intervalle $\{(a_k, b_k)\}$ in $[0,1]$ gilt:

$$\sum_k (b_k - a_k) < \delta \Rightarrow \sum_k |f(b_k) - f(a_k)| < \varepsilon.$$

Beweis: Wir verwenden zunächst I, die Ratenfunktion, um eine Hölder-Stetigkeit von f nachzuweisen (mit Hölder-Exponent $\frac{1}{2}$). In der Tat gilt für $I(f) \leq L$

$$|f(t) - f(s)| = |\int_s^t f'(x)dx| \leq |t-s|^{1/2}(\int_s^t (f'(x))^2 dx)^{1/2} \leq \sqrt{2L}|t-s|^{1/2}.$$

Dies hilft, um zu zeigen, dass I wirklich eine Ratenfunktion ist. I ist halbstetig. Das zeigen wir im Anschluss. Die vorangegangene Rechnung beweist, dass die $f \in$

$\{I \leq L\}$ gleichgradig stetig sind. Außerdem gilt für jedes f mit $I(f) \leq L$, dass $f(0) = 0$ ist und somit folgt aus der Rechnung auch die punktweise Beschränktheit von $\{I \leq L\}$. Dabei ist $\{I \leq L\}$ nach dem Satz von Arzela und Ascoli kompakt.

Nun widmen wir uns der Herleitung der oberen Schranke für die großen Abweichungen. Sei also $A \subseteq C([0,1])$ abgeschlossen. Für $f \in C([0,1])$ und $N \in \mathbb{N}$ zerlegen wir $[0,1]$ in N disjunkte Teilintervalle der Lage $\frac{1}{N}$ und approximieren f durch f_N. Dabei ist $f(x) = f_N(x)$ für alle $x = 0, \frac{1}{N}, \frac{2}{N}, \ldots, 1$ und ansonsten sei f_N die lineare Interpolation zwischen diesen Stützstellen. Die Rate von f_N lässt sich einfach berechnen:

$$I(f_N) = \frac{1}{2} \sum_{j=0}^{N-1} \left[f\left(\frac{j+1}{N}\right) - f\left(\frac{j}{N}\right) \right]^2,$$

da der Ausdruck in Klammern bis auf einen Faktor N gerade $f_N'(x)$ für alle $x \in [\frac{j-1}{N}, \frac{j}{N}]$ ist. Schreiben wir wieder A^δ für die Menge

$$A^\delta = \{g \in C([0,1]) : d(g, A) \leq \delta\}$$

für $\delta > 0$, so können wir $Q_\varepsilon(A)$ wie folgt abschätzen:

$$Q_\varepsilon(A) = Q_\varepsilon[f \in A] \leq Q_\varepsilon[f_N \in A^\delta] + Q_\varepsilon[\|f_N - f\| \geq \delta],$$

wobei $\|\cdot\|$ die Supremumsnorm auf $C([0,1])$ ist. Kürzen wir mit

$$L_\delta = \inf_{f \in A^\delta} I(f)$$

und

$$L = \inf_{f \in A} I(f)$$

ab, so folgt aus der Tatsache, dass $I(\cdot)$ halb-stetig, und aus der Kompaktheit der Niveaumengen, dass $L_\delta \to L$ für $\delta \to 0$ konvergiert. In der Tat ist $L_\delta \leq L$ für alle δ. Andererseits sei $f_\delta \in A^\delta$ so, dass $I(f_\delta) \leq L_\delta + \frac{\delta}{2}$. Für $\delta \to 0$ nehmen die Funkionen auch an der Infimumsbildung für L teil, d. h. die Häufungspunkte der Folge $(f_\delta)_\delta$ liegen in A, andererseits folgt aus der Halbstetigkeit von I für jeden Häufungspunkt f der $(f_\delta)_\delta$

$$\lim_{\delta \downarrow 0} I(f_\delta) \geq I(f) \geq L.$$

Also gilt

$$\lim_{\delta \downarrow 0} L_\delta = L.$$

Definitionsgemäß impliziert $f_N \in A^\delta$, dass $I(f_N) \geq L_\delta$ ist, also

$$Q^\varepsilon(f_N \in A^\delta) \leq Q^\varepsilon(I(f_N) \geq L_\delta).$$

$I_N(f_N)$ lässt sich aaber unter Q^ε gut ausrechnen. Die Pfade haben ja unter Q^ε unabhängige normalverteilte Zuwächse, also ist $I(f_N)$ die Summe von Quadraten zentrierter Gaußscher Zufallsvariablen mit Varianz $\frac{1}{N}$. Also ist $I(f_N)$ bis auf Skalierung

χ^2_N-verteilt. Genauer ist

$$Q_\varepsilon[I(f_N) \geq L_\delta] = Q_\varepsilon\left[\sum_{j=0}^{N-1}\left(f\left(\frac{j+1}{N}\right) - f\left(\frac{j}{N}\right)\right)^2 \geq 2L_\delta\right]$$
$$= \mathbb{P}\left[\sum_{i=1}^{N} Y_i^2 \geq 2L_\delta \frac{N}{\varepsilon}\right],$$

wobei die (Y_i) unter $\mathbb{P}$ i.i.d. $\mathcal{N}(0,1)$-verteilt sind. Die rechte Seite ist gegeben durch

$$\frac{1}{\Gamma(\frac{N}{2})}\int_{2L_\delta\frac{N}{\varepsilon}}^{\infty} \frac{1}{2^{N/2}} x^{N/2-1}e^{-x/2}dx.$$

Auf einer logarithmischen Skala ist dieser Integral so groß wie der maximale Term (das ist die ursprüngliche Laplace-Methode) und dies ist

$$e^{-\frac{N}{\varepsilon}L_\delta}\frac{1}{\Gamma(\frac{N}{2})2^{N/2}}2^{N/2-1}\frac{L_\delta^{N/2-1}N^{N/2-1}}{\varepsilon^{N/2-1}}.$$

Also gilt für festes N und $\varepsilon \to 0$

$$\limsup_{\varepsilon\to 0} \varepsilon \cdot \log Q_\varepsilon[I(f_N) \geq L_\delta] \leq -L_\delta.$$

Es bleibt der $\|f_N - f\|$-Term zu verarzten. Nun ist

$$\|f_N - f\| \leq 2 \sup_{0\leq j\leq N} \sup_{\frac{j}{N}\leq t\leq\frac{j+1}{N}} |f(t) - f(j/N)|.$$

Beachte, dass alle Ereignisse $\{\sup_{\frac{j}{N}\leq t\leq\frac{j+1}{N}} |f(t) - f(j/N)| \geq \delta/2\}$ die gleiche Wahrscheinlichkeit besitzen:

$$\begin{aligned} Q_\varepsilon[\sup_{\frac{j}{N}\leq t\leq\frac{j+1}{N}} |f(t) - f(j/N)| \geq \delta/2] &= Q_\varepsilon[\sup_{0\leq t\leq 1/N} |f(t)| \geq \delta/2] \\ &\leq 2Q_\varepsilon[\sup_{0\leq t\leq 1/N} f(t) \geq \delta/2] \\ &= 4Q_\varepsilon[f(\frac{1}{N}) \geq \delta/2]. \end{aligned}$$

Letzte Gleichheit folgt aus dem Spiegelungsprinzip für die 1-dimensionale Brownsche Bewegung: Ist $M_t := \max_{s\in[0,t]} \beta_s$, so gilt

$$\mathbb{P}[M_t > a] = 2\mathbb{P}[\beta_t > a]$$

(der Beweis hierfür ist eine Übung). Nun ist $f(\frac{1}{N})$ unter Q_ε normalverteilt mit Erwartungswert 0 und Varianz $\frac{N}{\varepsilon}$. Somit ist:

$$Q_\varepsilon[f(\frac{1}{N}) \geq \delta/2] \leq e^{-\frac{N}{\varepsilon}\frac{\delta^2}{8}}$$

und daher

$$Q_\varepsilon[\|f_N - f\| \geq \delta/2] \leq 4Ne^{-\frac{N}{\varepsilon}\delta^2/8},$$

also

$$\overline{\lim_{\varepsilon\to 0}}\, \varepsilon \log Q_\varepsilon[\|f_N - f\| \geq \delta/2] \leq -\frac{N\delta^2}{8}.$$

Führt man das mit der vorhergehenden Rechnung zusammen, ergibt sich

$$\limsup_{\varepsilon\to 0} \varepsilon \log Q_\varepsilon[A] \leq -\inf\{L_\delta, \frac{N\delta^2}{8}\}.$$

Da N und δ beliebig waren, können wir erst $N \to \infty$ und dann $\delta \to 0$ gehen lassen und erhalten

$$\limsup_{\varepsilon\to 0} \varepsilon \log Q_\varepsilon[A] \leq -L.$$

Dies zeigt die obere Schranke.

Für die untere Schranke sei $G \subseteq C([0,1])$ offen. Wir wollen zeigen, dass

$$\liminf_{\varepsilon\to 0} \varepsilon \log Q_\varepsilon[G] \geq -\inf_{g\in G} I(g).$$

Sei $g \in G$ und N eine offene Umgebung von g in G. Da $Q_\varepsilon[G] \geq Q_\varepsilon[N]$ und g beliebig ist, genügt es zu zeigen:

$$\liminf_{\varepsilon\to 0} \varepsilon \log Q_\varepsilon[N] \geq -I(g).$$

O.B.d.A. sei $I(g) < \infty$. Wir werden N als hinreichend kleinen δ-Ball $B_\delta(g)$ um g wählen. Wir benötigen für diesen Schritt den Satz von Cameron und Martin, einen Spezialfall des Satzes von Girsanov. Dieser Satz besagt, dass, falls $\mathbb{P}^\beta$ die Verteilung der Brownschen Bewegung ist, $\psi \in C([0,1])$ eine stetige Funktion und $\mathbb{P}^{\beta+\psi}$ die Verteilung der Brownschen Bewegung mit Drift ψ, das Maß $\mathbb{P}^{\beta+\psi}$ absolut stetig ist bezüglich $\mathbb{P}^\beta$ und die Radon-Nikodym-Ableitung gegeben ist durch

$$\frac{d\mathbb{P}^{\beta+\psi}}{d\mathbb{P}^\beta}(w) = \exp(Z(\psi)(w) - I(\psi)).$$

Hierbei ist

$$Z(\psi)(w) = \int_0^1 \varphi'(t)dw(t)$$

das Ito-Integral unter $\mathbb{P}^\beta$ und $Z(\psi)(w) < \infty$ fast sicher. Damit gilt

$$\begin{aligned} Q_\varepsilon[B_\delta(g)] &= \mathbb{P}[\beta - \frac{1}{\sqrt{\varepsilon}}g \in B_{\delta/\sqrt{\varepsilon}}(0)] \\ &= \mathbb{P}^{\beta-\frac{1}{\sqrt{\varepsilon}}g}[B_{\delta/\sqrt{\varepsilon}}(0)] \\ &= \mathbb{E}_\beta\left[\exp\left(-\frac{1}{\sqrt{\varepsilon}}Z(g) - \frac{1}{\varepsilon}I(g)\right)\mathbb{1}_{B_{\delta/\sqrt{\varepsilon}}(0)}\right], \end{aligned}$$

wobei wir die quadratische Gestalt von I ausnutzen. Die letzte Zeile ist für beliebiges $h > 0$

$$\begin{aligned} &\geq e^{-\frac{1}{\varepsilon}I(g)}\mathbb{E}_\beta\left[e^{-\frac{1}{\sqrt{\varepsilon}}Z(g)}\mathbb{1}_{B_{\frac{\delta}{\sqrt{\varepsilon}}}(0)}\mathbb{1}_{\{\frac{1}{\sqrt{\varepsilon}}Z(g)\leq\frac{h}{\varepsilon}\}}\right] \\ &\geq e^{-\frac{1}{\varepsilon}[I(g)+h]}\mathbb{P}^\beta(\{Z(g) \leq \frac{h}{\sqrt{\varepsilon}}\} \cap B_{\frac{\delta}{\sqrt{\varepsilon}}}(0)). \end{aligned}$$

Nach der Vorbemerkung ist $Z(g)$ fast sicher endlich, selbiges gilt für die Brownsche Bewegung β. Also konvergiert die Wahrscheinlichkeit auf der rechten Seite gegen 1. Also

$$\liminf_{\varepsilon\to 0} \varepsilon \log Q_\varepsilon[B_\delta(g)] \geq -I(g) - h.$$

Da $h > 0$ beliebig war folgt die Behauptung. □

Es bleibt die Halbstetigkeit von I zu zeigen.

Lemma 7.2 *I ist halbstetig.*

Beweis: Sei $(f_n)_n \subseteq C([0,1])$ eine gegen $f \in C([0,1])$ konvergierte Folge. Zu zeigen ist, dass

$$I(f) \leq \liminf_{n\to\infty} I(f_n)$$

gilt. O.B.d.A. existiere $s = \lim I(f_n)$, sonst gehen wir zu der entsprechenden Teilfolge für den Limes-Inferior über.

Untersuchen wir also f. Wegen $f_n(0)$ für alle n ist auch $f(0) = 0$. Weiter ist f absolutstetig. In der Tat gilt für alle $m \in \mathbb{N}$ und alle $[a_1, b_1], \ldots, [a_m, b_m] \subseteq [0,1]$, die disjunkt sind, mit $A = \bigcup_{i=1}^m [a_i, b_i]$:

$$\begin{aligned}
\sum_{i=1}^m |f_n(b_i) - f_n(a_i)| &\leq \int_A |f_n'(x)| dx \\
&\leq \sqrt{|A|} \sqrt{\int_0^1 |f_n'(x)^2| dx} \\
&\leq \left(\sum_{i=1}^m (b_i - a_i) \right)^{1/2} (2I(f_n))^{1/2} \\
&\underset{n\to\infty}{\longrightarrow} \left(\sum_{i=1}^m (b_i - a_i) \right)^{1/2} \sqrt{2s}.
\end{aligned}$$

Somit ist auch f absolut stetig, wenn man auch links den Limes $n \to \infty$ bildet. Somit können wir das Lemma von Lebesgue verwenden. Dies besagt, dass die absolut stetige Funktion eine Ableitung f' besitzt und dass für (eine messbare Version) diese(r) und fast alle t gilt:

$$\lim_{h\downarrow 0} \int_0^{t+h} f'(x) dx = f'(t).$$

Wir approximieren f und f_n mit stückweisen linearen Funktionen. Dazu sei $t_i^{(r)} = \frac{i}{r},\ i = 0, \ldots, r$ und φ_r sei die lineare Interpolation von $f(t_i^{(r)})$. Analog definieren wir die lineare Interpolation der Werte $f_n(t_i^{(r)})$ und nennen sie $\varphi_{r,n}$. Aus dem eben zitierten Lemma von Lebesgue folgt, dass für fast alle $t \in [0,1]$ gilt:

$$\lim_{\substack{v\uparrow t \\ u\downarrow t}} \frac{f(v) - f(u)}{v - u} = \lim_{\substack{v\uparrow t \\ u\downarrow t}} \frac{1}{u - v} \int_v^u f'(x) dx = f'(t).$$

Also folgt für fast alle $t \in [0,1]$

$$\lim_{r\to\infty} \varphi_r(t) = \lim_{r\to\infty} \frac{f(t_i^{(r)}) - f(t_{i-1}^{(r)})}{t_i^{(r)} - t_{i-1}^{(r)}} = f'(t).$$

Daher erhält man mit dem Lemma von Fatou:

$$I(f) = \frac{1}{2}\int_0^1 |f'(t)|^2 dt \leq \underline{\lim}_{r\to\infty} \frac{1}{2}\int_0^1 |\varphi_r'(t)|^2 dt = \liminf_{r\to\infty} I(\varphi_r).$$

Außerdem überlegt man sich, dass

$$\begin{aligned} I(\varphi_{r,n}) &= \frac{1}{2}\sum_{i=1}^{r} \frac{|f_n(t_i^{(r)}) - f_n(t_{i-1}^{(r)})|^2}{t_i^{(r)} - t_{i-1}^{(r)}} \\ &\leq \frac{1}{2}\sum_{i=1}^{r} \int_{t_{i-1}^{(r)}}^{t_i^{(r)}} |f_n'(x)|^2 dx \\ &= I(f_n) \end{aligned}$$

aufgrund der Jensenschen Ungleichung gilt ($|\cdot|^2$ ist konvex). Schicken wir nun $n \to \infty$, so erhalten wir mit der vorhergehenden Ungleichung, dass einerseits

$$I(\varphi_r) \leq \liminf_{n\to\infty} I(f_n)$$

gilt und andererseits

$$I(f) \leq \liminf_{r\to\infty} I(\varphi_r),$$

somit auch

$$I(f) \leq \liminf_{n\to\infty} I(f_n).$$

Dies beweist die Unterhalbstetigkeit von $I(\cdot)$, somit auch die Abgeschlossenheit der Niveaumengen von $I(\cdot)$ und beendet den Beweis des Satzes von Schilder. □

8 Quellenkodierung

Dieser Abschnitt befasst sich mit einer Anwendung der Theore großer Abweichungen in der Informationstheorie. Dieses Gebiet der angewandten Mathematik wurde Mitte des letzten Jahrhunderts im Wesentlichen von dem britischen Mathematiker C. Shannon gegründet. Bedeutende Beiträge stammen u. a. von I. Cszisar und R. Ahlswede.

Die Motivation für die Fragestellung der Quellenkodierung stammt aus der Übertragung (bzw. dem Speichern, was heute vermutlich das aktuellere Problem ist) von Nachrichten. Nehmen wir an, diese entstammen einem 0-1-Alphabet. Dann können wir natürlich versuchen, genau diese 0-1-Sequenz zu übermitteln. Eventuell ist es aber sparsamer, wenn man sie kodiert. Besteht sie zum überwiegenden Teil aus

0'en, so könnte man immer Anzahl der aufeinander folgenden 0'en speichern statt der 0'en selbst. Die Frage ist, wie stark man eine Nachricht komprimieren kann. Da dies von der Art der Nachricht nicht abhängen soll, betrachtet man zufällige Nachrichten.

Man kann sich neben der Frage, wie stark man denn Nachrichten komprimieren kann, darüber hinaus damit beschäftigen, was geschieht, wenn man diese Grenze unterschreitet. Mit derartigen Problemen wollen wir uns in diesem Abschnitt beschäftigen. Um das Problem zu formalisieren, sei $\{1, \dots, d\},\ d < \infty$, das Alphabet. Wir stellen uns vor, dass die Wörter unserer Botschaft allesamt die Länge n besitzen. Diese Wörter wollen wir nun mithilfe von Codewörtern der Länge r, d. h. aus $\{0,1\}^r$ kodieren. Also gibt es 2^r Codewörter. Da es d^n Codewörter gibt, erreicht man nur dann eine Datenkompression, wenn $2^r < d^n$ ist. Bei der Codierung mit möglichen Fehlern schreibt man $2^r - 1$ Quellwörtern ein eindeutiges Codewort zu. Das verbleibende Codewort zeigt im Wesentlichen einen Kodierungsfehler an. Wir nehmen nun an, dass die Buchstaben des Quellcodes i.i.d. sind (das ist ein wenig unrealistisch) und die Marginalverteilung

$$\mathbb{P}(X_1 = i) = p_i$$

besitzen. Benötigen wir eine zweite Wahrscheinlichkeit auf $\{1, \dots, d\}$, so heißt diese in der Regel $Q = (q_1, \dots, q_d)$. Mit U bezeichnen wir die Gleichverteilung auf der Menge $\{1, \dots, d\}$. Die Entropie von $\mathbb{P}$ ist gegeben durch

$$H(P) = -\sum_{i=1}^{d} p_i \log p_i.$$

H ist so etwas wie die "mittlere Überraschung", die P zu bieten hat. Offenbar ist $H(U) = \log d$ und $H(U)$ ist maximal. Der erste Kodierungssatz von Shannon besagt im Wesentlichen, dass Entropie das grundlegende Maß für den Informationsinhalt einer Quelle ist. Genauer: Für jedes $\varepsilon > 0$ gibt es einen Code, also eine Abbildung von $\{1, \dots, d\}^n \to \{0,1\}^r$, sodass die Wahrscheinlichkeit eines Kodierungsfehlers kleiner ist als ε, falls n groß genug und

$$R := r \log \frac{2}{n} > H$$

gilt. Mit anderen Worten konvergiert die Wahrscheinlichkeit für einen Kodierungsfehler für $n \to \infty$ gegen 0. Wir werden hier sehen, dass diese Konvergenz sogar exponentiell schnell ist und die Raten mithilfe der Theorie großer Abweichungen bestimmen. Sei also

$$H < R < \log d$$

(das lässt sich offenbar für $n \to \infty$ erreichen). Sei $x = (x_1, \dots, x_n)$ ein Quellwort. Wir betrachten das empirische Maß

$$L_{n,x}(j) = \frac{1}{n} \sum_{i=1}^{n} \delta_j(x_i),$$

das die Häufigkeit des Vorkommens einzelner Buchstaben in x misst. Seine Entropie ist offenbar

$$H(L_{n,x}) = -\sum_{j=1}^{d} L_{n,x}(j) \log L_{n,x}(j).$$

Wir fragen uns nun, wieviele der d^n Quellwörter ein $L_{n,x}$ haben mit $H(L_{n,x}) < R$. Diese Menge heiße S_n, also

$$S_n = \{x : H(L_{n,x}) < R\}.$$

Formal ist S_n offen (als Urbild von (O, R) unter dem stetigen H, aber in der Tat sind ja in allen Mengen in $\{1, \ldots, d\}^n$ zugleich offen und abgeschlossen). Offenbar gilt

$$|S_n| = d^n U(S_n).$$

Somit folgt aus dem Satz von Sanov

$$\begin{aligned}
\overline{\lim}\frac{1}{n} \log U(S_n) &\leq \overline{\lim}\frac{1}{n} \log U(\{x : H(L_{n,x} \leq R\}) \\
&\leq - \inf_{Q:H(Q)\leq R} \left(\sum_{i=1}^{d} q_i \log \frac{q_i}{d^{-1}} \right) \\
&= -\log d + \sup_{Q:H(Q)\leq R} H(Q) \\
&= R - \log d.
\end{aligned}$$

Also $\overline{\lim}\frac{1}{n} \log |S_n| \leq R$. Nun nehmen wir einen Code, bei dem jedes Wort aus S_n ein eindeutiges Codewort bekommt. Wir können dies mit m_n Buchstaben aus $\{0, 1\}$ tun, wobei

$$m_n = \frac{\log(|S_n| + 1)}{\log 2}.$$

Wir nehmen ein Extracodewort, um alle Nachrichten zu kodieren, die nicht in S_n liegen. Also ist die Wahrscheinlichkeit für einen Codierungsfehler $\mathbb{P}(H(L_{n,x}) \geq R)$. Nach dem Satz von Sanov ist aber

$$\overline{\lim}\frac{1}{n} \log \mathbb{P}(H(L_{n,x}) \geq R) \leq - \inf_{Q:H(Q)\geq R} H(Q|\mathbb{P}).$$

Somit haben wir gezeigt

Satz 8.1 ***(Jelinek-Cszisar):*** *Für jedes R mit $H < R < \log d$ gibt es eine Folge von Codes in $\{0, 1\}^{m_n}$, sodass $\overline{\lim}\frac{m_n}{n} \leq R/\log 2$ ist, mit*

$$\overline{\lim}\frac{1}{n} \log \mathbb{P}\textit{(Codierungsfehler)} \leq - \inf_{Q:H(Q)\geq R} H(Q|\mathbb{P}).$$

Dieser Satz ist auch optimal in dem Sinne, als dass die Rate optimal ist. Um das einzusehen betrachten wir irgendeinen Code mit binären Codewörtern der Länge m_n und $\overline{\lim}\frac{m_n}{n} \leq \frac{R}{\log 2}$. Nun ist für endliches d nicht nur

$$H : \mathcal{M}^1(\{1, \ldots, d\}) \to \mathbb{R}$$

stetig, sondern auch

$$H(\cdot|\mathbb{P}) : \mathcal{M}^1(\{1,\dots,d\}) \to \mathbb{R}.$$

Sei nun $Q_0 \subset \mathcal{M}^1(\{1,\dots,d\})$ beliebig mit $H(Q_0) > R$. Aufgrund der Stetigkeit von $H(\cdot)$ und $H(\cdot|\mathbb{P})$ gibt es dann eine offene Menge $G \subseteq \mathcal{M}^1(\{1,\dots,d\})$, sodass für alle $Q \in G$

$$|H(Q) - H(Q_0)| < \varepsilon \quad \text{und} \quad |H(Q|\mathbb{P}) - H(Q_0|\mathbb{P})| \leq \varepsilon$$

gilt. Wie viele Quellwörter haben nun eine empirische Verteilung $L_{n,x}$ mit Werten in G? Wir bezeichnen diese Menge mit

$$S_{G,n} := \{x : L_{n,x} \in G\}.$$

Benutzt man das gleiche Argument wie oben, erhält man:

$$\begin{aligned}
\underline{\lim}\frac{1}{n}\log|S_{G,n}| &= \underline{\lim}\frac{1}{n}\log d^n U(S_{G,n}) \\
&\geq \log d - \inf_{Q\in G} H(Q,U) \\
&= \sup_{Q\in G} H(Q) \\
&\geq H(Q_0) - \varepsilon > R,
\end{aligned}$$

wobei wir $\varepsilon > 0$ gerade so klein wählen, dass die letzte Ungleichung stimmt. Sei nun S_n die Menge der Wörter, die man eindeutig dechiffrieren kann. Nach Konstruktion ist $|S_1| \leq 2^{m_n}$, also

$$\overline{\lim}_{n\to\infty}\frac{1}{n}\log|S_n| \leq \overline{\lim}\frac{m_n}{n}\log 2 \leq R.$$

Somit ist für hinreichend großes n jeder positive Bruchteil (zum Beispiel die Hälfte) aller Quellwörter in $S_{G,n}$ nicht in S_n. Ist darüber hinaus $x \in \{1,\dots,d\}^n$ mit $L_{n,x} =: Q \in G$, so wissen wir, so sind für jedes i genau nq_i Buchstaben von x gleich. Somit gilt

$$\begin{aligned}
\frac{1}{n}\log\mathbb{P}(x) = \frac{1}{n}\log\prod_{i=1}^{d} p_i^{nq_i} &= \sum_{i=1}^{d} q_i \log p_i \\
&= \sum_{i=1}^{d} q_i \log q_i - q_i \log\frac{q_i}{p_i} \\
&= -H(Q) - H(Q|\mathbb{P}) \\
&\geq -\sup_{Q\in S_{G,n}} [H(Q) + H(Q|\mathbb{P})] \\
&\geq -\sup_{Q\in S_{Gn}} [H(Q_0) + H(Q_0|\mathbb{P}) + 2\varepsilon] \\
&= -H(Q_0) - H(Q_0|\mathbb{P}) + 2\varepsilon.
\end{aligned}$$

Daher folgt

$$\begin{aligned}\underline{\lim}_{n\to\infty}\frac{1}{n}\log\mathbb{P}(\text{Codierungsfehler}) &\geq \underline{\lim}_{n\to\infty}\frac{1}{n}\log\mathbb{P}(x\notin S_n \text{ und } x\in S_{G,n})\\ &\geq \lim_{n\to\infty}\frac{1}{n}\log\left[\frac{1}{2}|S_{G,n}|\cdot\mathbb{P}(x|x\in S_{G,n})\right]\\ &\geq H(Q_0)-\varepsilon-[H(Q_0)+H(Q_0|\mathbb{P})+2\varepsilon]\\ &= -H(Q_0|\mathbb{P})-3\varepsilon.\end{aligned}$$

Schickt man $\varepsilon\to 0$, so erhält man

Satz 8.2 *Für jede Folge von Codewörtern in* $\{0,1\}^{m_n}$ *mit* $\overline{\lim}\frac{m_n}{n}\leq\frac{R}{\log 2}$ *gilt*

$$\underline{\lim}\frac{1}{n}\log\mathbb{P}\text{ (Codierungsfehler)}\geq -\inf_{Q:H(Q)\geq R}H(Q|\mathbb{P}).$$

Wir wenden uns nun einer zweiten Fragestellung zu. Hierzu ist es wichtig zu bemerken, dass $H(\cdot)$ eine untere Schranke an dem Informationsgehalt einer zufälltigen Nachricht ist und eine untere Schranke an die erwartete Codewortlänge. Die Frage, die wir nun behandeln wollen, was geschieht, wenn man probiert, mit weniger Platz auszukommen. Hierzu sei $\rho(\cdot,\cdot)$ ein Verzerrungsmaß. $\rho(j,k)$ ist das Maß für die Verzerrung, wenn wir den Buchstaben j aus dem Quellalphabet mit dem Buchstaben k kodieren. Damit können wir eine Durchschnittsverzerrung berechnen, wen wir das Wort x mit dem Wort y (beide n-elementig) kodieren. Die Verzerrung ist dann

$$\rho_n(x,y)=\frac{1}{n}\sum_{i=1}^{n}\rho(x_i,y_i).$$

Sei $\mathbb{P}$ die Wahrscheinlichkeitsverteilung auf der Quelle. Die Tatsache, dass wir die Buchstaben i.i.d. auf $\{1,\ldots,d\}$ wählen heißt hier "gedächtnislose Quelle". Es sei $Q(k|j)$ eine bedingte Wahrscheinlichkeit dafür, dass wir den Buchstaben j des Quellalphabets mit dem Buchstaben k im Zielalphabet wiedergeben. $Q(\cdot|\cdot)$ zusammen mit $\mathbb{P}(\cdot)$ ergibt eine Verteilung

$$Q(k)=\sum_{j=1}^{d}\mathbb{P}(j)Q(k|j)$$

auf dem Zielalphabet. Ähnlich können wir die erwartete Verzerrung berechnen:

$$d(Q)=\sum_{j,k}\mathbb{P}(j)Q(k|j)\rho(j,k).$$

Eine bedingte Wahrscheinlichkeit Q heißt D-zulässig, wenn $d(Q)\leq D$. Die Menge der entsprechenden Q's bezeichnen wir mit $\mathcal{Q}_D$. Mit der folgenden Größe I bezeichnen wir für eine bedingte Wahrscheinlichkeit $Q(\cdot|\cdot)$ die durchschnittliche gegenseitige Information.

$$I(Q)=\sum_{j,k}\mathbb{P}(j)Q(k|j)\log\frac{Q(k|j)}{Q(k)}.$$

$I(\cdot)$ misst die Information, die man über die Quelle erhält, wenn man nur die Ausgabe beobachten kann. Schließlich definieren wir

$$R(D) := \min_{Q \in \mathcal{Q}_D} I(Q)$$

als die Ratenverzerrungsfunktion. Q^* sei der Minimierer.

Der Beweis des foglendes Lemmas ist eine Übung.

Lemma 8.3 ***(Gallagher):*** *Es gilt*

$$R(D) \geq \sup_{\vartheta \leq 0} [\vartheta D - f(\vartheta)],$$

wobei

$$f(\vartheta) = \sum_j \mathbb{P}(j) \log \left[\sum_k \exp(\vartheta \rho(j,k)) Q(k) \right]$$

und $Q(\cdot)$ ist aus einer bedingten Wahrscheinlichkeit $Q(\cdot|\cdot) \in \mathcal{Q}_D$ abgeleitet.

Wir kommen nun dazu, unser zentrales Resultat zu formulieren. Jedes $B := \{y_1, \ldots, y_k\}$ n-dimensionaler Vektoren mit Komponenten aus dem Zielalphabetheißt Code der Größe j und Blocklänge n. Die Elemente von B heißten Code-Wort. Jedes Quellwort x wird auf ein beliebiges y mit $\rho_n(x,y) = \min_z \rho_n(x,z)$ abgebildet. Die Verzerrung notieren wir mit

$$\rho_n(x|B) := \min_{y \in B} \rho_n(x,y).$$

Die durchschnittliche Verzerrung des Codes ist definiert als

$$\mathbb{E}[\rho_n(x|B)] := \sum_x \mathbb{P}(x) \rho_n(x|B).$$

Falls $\mathbb{E}[\rho_n(x|B)] \leq D$, heißt der Code D-zulässig. Die Codierungsrate ist

$$R := \frac{1}{n} \log k.$$

Um das Konzept zu verstehen, stelle man sich vor, der log würde zur Basis 2 genommen. Dann wäre in R die Anzahl an 0en und 1en, die man braucht, um jeden Codewort in B eine eindeutige $0-1$-Sequenz zuzuordnen. R wäre dann die Anzahl an Binärziffern, die man pro Buchstaben im Quellalphabet braucht. Nun ist bei uns der Logarithmus zur Basis e, aber qualitativ ändert das nichts.

Satz 8.4 ***(Shannon):*** *Für eine gedächtnislose Quelle und gegebenes Verzerrungsmaß $\rho(\cdot,\cdot)$ und gegebene $\varepsilon > 0$ und $D > 0$ gibt es ein $n \in \mathbb{N}$, so dass es einen $(D+\varepsilon)$-zulässigen Code der Bloklänge n mit Rate*

$$R < R(D) + \varepsilon$$

gibt.

Wir suchen k Codewörter als unabhängige Zufallsvariablen gemäß der Verteilung

$$Q(y) := \prod_{i=1}^{n} Q(y_i),$$

wobei $y = (y_1, \ldots, y_n)$ und Q eine Verteilung auf dem Zielalphabet ist, die wir später wählen werden. Sei

$$S(x) := \{y : \rho_n(x, y) \leq D + \delta\}.$$

Weiter sei

$$Q(S(x)) := \sum_{y \in S(x)} Q(y)$$

die Wahrscheinlichkeit dafür, dass ein zufällig gewähltes Codewort in $S(x)$ liegt. Somit ist die Wahrscheinlichkeit dafür, dass keines der k zufällig gewählten Codewörter in $S(x)$ liegt $(1 - Q(S(x))^k$. Sei

$$\rho_{\max} = \max_{k,l} \rho(l, h).$$

Wir wollen nun eine obere Schranke für die erwartete Verzerrung herleiten, die ein zufälliger Code erzeugt. Diese nennen wir $\bar{\rho}$. Dann gilt

$$\begin{aligned}
\bar{\rho} &\leq \sum_x \mathbb{P}(x)[(D + \delta)(1 - (1 - Q(S(x)))^k] + \rho_{\max}[1 - Q(S(x))]^k \\
&\leq D + \delta + \rho_{\max} \sum_x \mathbb{P}(x)(1 - Q(S(x)))^k \\
&\leq D + \delta + \rho_{\max} \sum_x \mathbb{P}(x) e^{-kQ(S(x))}.
\end{aligned}$$

Wir werden in der Folge

$$k := e^{nR(D)+2\delta}$$

wählen und die Frage ist dann, wie schnell $Q(S(x))$ gegen 0 konvergiert. Falls dies langsamer geschieht als k wächst, dann wird

$$\rho_{\max} \sum_x \mathbb{P}(x) e^{-kQ(S(x))}$$

kleiner als beispielsweise δ. Daraus folgt dann die Behauptung des Satzes.

Dazu wählen wir

$$Q(y) := \prod_{i=1}^{n} Q^*(y),$$

wo, wie schon erwähnt, Q^* der Minimierer in $\mathcal{Q}_D$ ist. Dann ist

$$Q(S(x)) = Q^* \left(\frac{1}{n} \sum_{i=1}^{n} \rho(x_i, y_i) \leq D + \delta \right).$$

Es geht also darum, das große Abweichungsverhalten von $\frac{1}{n}\sum_{i=1}^n \rho(x_i, y_i)$ zu verstehen. Bemerke, dass die $\rho(x_i, y_i)$ als Funktion von y zwar unabhängig aber nicht identisch verteilt sind. Wir bemühen daher das Gärtner-Ellis-Theorem. Es ist

$$\begin{aligned}\rho(\vartheta) :&= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{Q^*} e^{\vartheta \sum_{i=1}^n \rho(x_i, y_i)} \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \log \mathbb{E}_{Q^*} e^{\vartheta \rho(x_i, y_i)} \\ &= \mathbb{E}_{\mathbb{P}} \log \mathbb{E}_{Q^*} e^{\vartheta \rho(x_i, y_i)}\end{aligned}$$

nach dem Gesetz der großen Zahlen. Somit ist

$$\rho(\vartheta) = \sum_j \mathbb{P}(j) \log(\sum_k e^{\vartheta \rho(j,k)} Q^*(k)) = f(\vartheta)$$

nach Definition. Somit folgt aus dem Gärtner-Ellis-Theorem

$$\begin{aligned}&\underline{\lim} \frac{1}{n} \log Q^* \left(\frac{1}{n} \sum_{k=1}^n \rho(x_i, y_i) \le D + \delta \right) \\ &\ge - \sup_{\vartheta} [\vartheta D - f(\vartheta)] \\ &= - \sup_{\vartheta \le 0} [\vartheta D - f(\vartheta)]\end{aligned}$$

(das zeigt man wie im Satz von Cramér). Nach dem Lemma von Gallagher ist somit

$$\underline{\lim} \frac{1}{n} \log Q^* \left(\frac{1}{n} \sum_{i=1}^n \rho(x_i, y_i) \right) \ge -R(D),$$

also geht $k \cdot Q(S(x)) \to \infty$. Das beweist die Behauptung. □

9 Markov-Ketten

Wir werden nun eine nächste abhängige Situation betrachten. Sei also $X_1, X_2, \ldots$ eine Markovkette auf der endlichen Menge F. Es sei $\pi = (\pi(x, y))$ ihre Übergangsmatrix. Wir nehmen an, dass π und somit (X_n) ergodisch ist, sogar dass $\pi(x, y) > 0$ für alle x, y. Die stationäre Verteilung heiße p, also gilt

$$p\pi = p.$$

Es sei $V : F \to \mathbb{R}$ eine Abbildung mit p-Ewartungswert $m = \sum V(x)p(x)$. Da die Verteilung von X_n nach dem Ergodensatz gegen p konvergiert, folgt

$$\lim_{n\to\infty} \mathbb{P}_x \left[|\frac{1}{n} \sum_j V(X_j) - m| \ge a \right] = 0$$

für alle $a > 0$. Dabei bezeichnet $\mathbb{P}_x$ die Wahrscheinlichkeit für eine Markov-Kette mit Start in $x \in F$. Wir wollen sehen, dass – wie im Satz von Cramér – die Wahrscheinlichkeiten

$$\frac{1}{n} \log \mathbb{P} \left[\frac{1}{n} \sum_j V(X_j) \geq a \right]$$

für $a > m$ gegen einen negativen Limes konvergieren. Zunächst zeigen wir

Lemma 9.1 *Für jedes $V : F \to \mathbb{R}$ und jedes $x \in F$ existiert*

$$\log \sigma(V) := \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}_x e^{\sum_{i=1}^n V(X_i)}.$$

Dabei ist $\sigma(V)$ der betragsmäßig größte Eigenwert der Matrix

$$\pi_V := (\pi_V(x,y))_{x,y \in F} := (\pi(x,y) e^{V(x)})_{x,y}.$$

Bemerkung 9.2 *$\sigma(V)$ ist positiv, da alle Zeilen von π_V positive Summen haben. Als betragsmäßig größter Eigenwert hat $\sigma(V)$ einen rein positiven Eigenvektor nach dem Satz von Gershgorin.*

Beweis: Per Induktion zeigt man

$$F_x \left[\exp \left(\sum_{i=1}^n V(X_i \right) \right] = \sum_y [\pi_V]^n (x,y)$$

(das ist eine Übung). Nun wächst $[\pi_V]^n$ aber wie der betragsmäßig große Eigenwert, also

$$\sum_y [\pi_V]^n = c \cdot \sigma(V)^n.$$

Daraus folgt die Behauptung. □

Wendet man diese Erkenntnis auf die exponentielle Chebyshev-Ungleichung an, so ergibt sich

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}_x \left[\frac{1}{n} \sum_{i=1}^n V(X_i) \geq a \right] \leq \log \sigma(V) - a$$

bzw. wenn man V durch λV, $\lambda > 0$, ersetzt,

$$\limsup \frac{1}{n} \log \mathbb{P}_X \left[\frac{1}{n} \sum_{i=1}^n V(X_i) \geq a \right] \leq \log \sigma(\lambda V) - \lambda a.$$

Wir können diese Schranke über alle $\lambda > 0$ optimieren und erhalten:

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}_X \left[\frac{1}{n} \sum_{i=1}^n V(X_i) \geq a \right] \leq -h(a) := - \sup_{\lambda \geq 0} [\lambda a - \log \sigma(\lambda V)].$$

Nun erhält man mit der Jensenschen Ungleichung:

$$\begin{aligned}\log \sigma(\lambda V) &= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_x[\exp(\lambda \sum_{i=1}^{n} V(X_i)] \\ &\geq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\lambda X_i) \\ &= \lambda m.\end{aligned}$$

Damit ist

$$\sup_{\lambda\geq 0}[\lambda a - \log \sigma(\lambda V)] = \sup_{\lambda\in\mathbb{R}}[\lambda a - \log \sigma(\lambda V)],$$

denn würde das Supremum in $\lambda_0 < 0$ angenommen, so wäre

$$\begin{aligned}h(a) &= -\lambda_0 a - \log \sigma(\lambda_0 V) \\ &\leq -\lambda_0 a - \lambda_0 m \\ &= -\lambda_0(a-m) > 0\end{aligned}$$

und wir hätten eine Wahrscheinlichkeit durch einen exponentiell wachsenden Term beschränkt.

Für die untere Schranke wenden wir wieder eine exponentielle Maßtransformation an. Die Idee ist es dabei $\mathbb{P}_x$ wieder durch ein Maß Q_x zu ersetzen, für das das betrachtete Verhalten typisch ist. Die Radon-Nikodym-Ableitung $\frac{dQ_x}{d\mathbb{P}_x}$ wird dann den Strafterm geben. Konkret können wir das erreichen, indem wir die Übergangsmatrix π zu $\bar{\pi}$ verändern, sodass $\bar{\pi}$ das invariante Maß q hat und $\sum_{x\in F} q(x)V(x) = a$ ist. Falls Q_x also die Verteilung der Kette mit Übergangsmatrix $\bar{\pi}$ ist, dann ist

$$\mathbb{P}_x(E) = \int_E R_n dQ_x,$$

wobei

$$R_n = \exp\left[\sum_{j=1}^{n-1} \log \frac{\pi(X_j, X_{j+1})}{\bar{\pi}(X_j, X_{j+1})}\right],$$

wie man durch Ausschreiben der Wahrscheinlichkeiten feststellt. Damit erhalten wir

$$\begin{aligned}&\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x\left(\frac{1}{n}\sum_{i=1}^{n} V(X_i) \geq a\right) = \underline{\lim}_{n\to\infty} \frac{1}{n} \log \int_{\{\frac{1}{n}\sum V(X_i) > a\}} R_n dQ_x \\ &\geq \underline{\lim}_{n\to\infty} \frac{1}{n} \log \int_{\{\frac{1}{n}\sum V(X_i)\geq a\}\cap\{\log R_n \geq \mathbb{E}_{Q_x} \log R_n\}} e^{\log R_n} dQ_x \\ &\geq \underline{\lim}_{n\to\infty} \frac{1}{n} \log e^{\mathbb{E}_{Q_x} \log R_n} Q_x\left(\left\{\frac{1}{n}\sum V(X_i) \geq a\right\} \cap \{\log R_n \geq \mathbb{E}_{Q_x} \log R_n\}\right) \\ &= \lim_{n\to\infty} -\frac{1}{n} \mathbb{E}_{Q_x}[-\log R_n],\end{aligned}$$

da die Wahrscheinlichkeiten rechts nach dem Ergodensatz für Markov-Ketten gegen 0 gehen. Nun konvergiert ebenfalls nach dem Ergodensatz

$$\lim_{n\to\infty} -\frac{1}{n} \log R_n = \frac{1}{n} \sum_{j=1}^{n} \log \frac{\bar{\pi}(X_j, X_{j+1})}{\pi(X_j, X_{j+1})}$$

gegen

$$J(\bar{\pi}) = \sum_{x,y} \log \frac{\bar{\pi}(x,y)}{\pi(x,y)} \bar{\pi}(x,y) q(x)$$

(Übung: man zeige das), somit haben wir gezeigt, dass

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x \left(\frac{1}{n} \sum V(X_i) \geq a \right) \geq -J(\bar{\pi})$$

ist. Da das für jedes $\bar{\pi}$ funktioniert, bekommen wir daraus sofort

$$\lim \frac{1}{n} \log \mathbb{P}_x \left[\frac{1}{n} \sum_{i=1}^{n} V(X_i) \geq a \right] \geq - \inf_{\pi : \sum V(X_i) q(X) = a} J(\pi).$$

Wir haben also für die uns interessierenden Wahrscheinlichkeiten eine obere und eine untere Schranke bewiesen mit dem kleinen Schönheitsfehler, das sich diese bislang wenig ähnlich sehen. Dies versuchen wir nun zu ändern. Dazu nehmen wir an, dass das Supremum bei der Bildung von $h(a)$ in einem λ angenommen wird, dass also

$$\lambda a - \log(\sigma(\lambda V)) = ha$$

gilt.

Dann ist $a = \frac{\sigma'(\lambda V)}{\sigma(\lambda V)}$ und die Matrix π_V besteht aus Einträgen der Form $\pi(x,y)e^{\lambda V(y)}$. σ ist natürlich wieder ihr größter Eigenwert. Zu diesem gibt es rechte und linke Eigenvektoren, die wir mit f bzw. g bezeichnen. Der Gag ist es nun, die richtige stochastische Matrix zu wählen. Wir setzen

$$\bar{\pi}(x,y) = \frac{1}{\sigma} \pi(x,y) e^{\lambda V(y)} \frac{f(y)}{f(x)}.$$

In der Tat ist π eine stochastische Matrix, denn für jedes x gilt

$$\begin{aligned} \sum_y \bar{\pi}(x,y) &= \frac{1}{\sigma} \frac{1}{f(x)} \sum_y \pi(x,y) e^{\lambda V(y)} f(y) \\ &= \frac{1}{\sigma f(x)} \cdot \sigma f(x) = 1. \end{aligned}$$

Das invariante Maß von $\bar{\pi}$ berechnet sich als

$$q(x) = \frac{f(x)g(x)}{Z},$$

wobei

$$Z = \sum_y f(y) g(y)$$

ist. Hier ist natürlich wichtig zu bemerken, dass f und g als Eigenvektoren zum größten Eigenwert σ nicht-negativ sind. In der Tat gilt

$$\begin{aligned} \sum_x q(x) \bar{\pi}(x,y) &= \frac{1}{Z} \sum_x \frac{1}{\sigma} g(x) \pi(x,y) e^{\lambda V(y)} f(y) \\ &= \frac{f(y)}{Z} \frac{1}{\sigma} \cdot \sigma g(y) = \frac{f(y)g(y)}{Z} = q(y). \end{aligned}$$

Mit ein wenig Störungstheorie (die wir hier nicht durchführen wollen) berechnet man, dass zudem auch

$$a = \sum_x q(x)V(x)$$

gilt, $\bar{\pi}$ also zur Menge der Maße gehört, die an der Infimumsbildung für J teilnehmen. Dann ist

$$\begin{aligned} J(\bar{\pi}) &= \sum_{x,y}(-\log\sigma + \lambda V(y) + \log f(y) - \log f(x)) \\ &= \sum_{x,y} -\log\sigma\pi(x,y)\cdot q(y) + \lambda V(y)\pi(x,y)q(y) \\ &= \lambda\sum_x q(x)V(x) - \log\sigma \\ &= \lambda a - \log\sigma \\ &= n(a). \end{aligned}$$

Somit sind h und J dasselbe und wir haben gezeigt:

Satz 9.3 *Für jede Markov-Kette auf einem endlichen Zustandsraum F und Übergangsmatrix π mit strikt positiven Einträgen und jedes*

$$V : F \to \mathbb{R}$$

genügt $(\frac{1}{n}\sum_{i=1}^n V(X_i))$ einem LDP mit Geschwindigkeit n und Ratenfunktion

$$n(a) = \sup_\lambda[\lambda a - \log\sigma(\lambda V)].$$

Übung 9.4 *Man versuche diesen Satz aus dem Gärtner-Ellis-Theorem abzuleiten.*

Wir wollen abschließend versuchen, auch den Satz von Sanov auf der Ebene von Markov-Ketten zu beweisen. Sei also für $y \in F$

$$L_n^y = \frac{1}{n}\sum_{j=1}^n \delta y(X_j)$$

die Anzahl der Besuche der Markovkette in y (da F endlich ist, genügt es, L_n für jedes y zu betrachten, um das empirische Maß zu verstehen). Dieses empirische Maß heiße ν_x^n, also

$$\nu_x^n(\cdot) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(\cdot).$$

ν_x^n hängt natürlich vom Startpunkt x ab. Wir zeigen

Satz 9.5 *(ν_x^n) genügt einem LDP mit Geschwindigkeit n und Ratenfunktion*

$$I(q) = \sup_{V \in \mathcal{M}} \sum q(x)V(x).$$

Hierbei ist $\mathcal{M}$ die Menge

$$\mathcal{M} = \{V : V = \log \frac{u}{\pi u} \quad \textit{für ein} \quad u \geq 0\}.$$

Beweisskizze: Die Kernidee ist, dass man für Potentiale

$$V(x) = \log \frac{u(x)}{(\pi(u)(x)} \quad \text{mit} \quad (\pi u)(x) = \sum_y \pi(x,y)u(y)$$

schnell ausrechnet, dass $f(x) = (\pi u)(x)$ ein rechter Eigenvektor zum Eigenwert 1 ist. Da für $f(x) \geq 0$ für alle x ist, muss daher $\sigma = 1$ sein. Also

$$\log \sigma(\log \frac{u}{\pi u}) = 0.$$

Da nun

$$\log \sigma(V + c) = \log \sigma(V) + c$$

für jedes c ist, ist $\log \sigma(V)$ gerade die Größe, um die wir V verschieben müssen, damit es in $\mathcal{M}$ ist.

Wir beweisen nun die obere Schranke: Sei $q \in \mathcal{M}^1(F)$ beliebig und $V \in \mathcal{M}$. Dann ist

$$\mathbb{E}_{\mathbb{P}_x}\left[\exp\left(\sum_{i=1}^n (X_i)\right)(\pi u)(X_{n+1})\right] = (\pi u)(x),$$

da wie eben bemerkt (πu) ein Eigenvektor zu Eigenwert 1 für $(\pi(x,y)e^{v(y)})$ ist und dies dann natürlich auch für Produkte solcher Matrizen gilt. Daraus folgt, dass

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_x}\left[\exp\left(\sum_{j=1}^n V(X_j)\right)\right] \leq 0.$$

Wählt man nun eine Folge offener Umgebungen U_ε um q, die mit $\varepsilon \downarrow 0$ gegen q schrumpft, ergibt sich aus dem vorhergehenden Satz:

$$\lim_{\varepsilon \downarrow 0} \limsup_{n\to\infty} \frac{1}{n} \log \nu_x^n(U_\varepsilon) \leq -\sum_x q(x)v(x).$$

Dies genügt aber schon, um die obere Schranke für das LDP herzuleiten, da der Grundraum kompakt ist.

Die untere Schranke führt wieder zu dem Problem, eine Übergangsmatrix $\bar{\pi}$ zu finden, sodass q das invariante Maß für $\bar{\pi}$ ist und

$$J(\bar{\pi}) = I(q)$$

gilt. Wir müssen also das Variationsproblem lösen

$$\min\{J(\bar{\pi}) : \bar{\pi} \text{ hat das invariante Maß } q\}.$$

Man kann zeigen, dass das Minimum in

$$\bar{\pi}(x,y) = \pi(x,y)\frac{f(y)}{(\pi f)(x)}$$

angenommen wird. Da dieses $\bar{\pi}$ q als invariantes Maß hat, kann man schnell die folgende Rechnung durchführen:

$$\begin{aligned} J &= \sum_{x,y}[\log f(y) - \log(\pi u)(x)]\bar{\pi}(x,y)q(x) \\ &= \sum_{x}[\log f(x) - \log(\pi u)(x)]q(x) \\ &= \sum_{x} V(x)q(x) \quad \text{für alle} \quad V \in \mathcal{M} \\ &\leq I(q). \end{aligned}$$

Da wir schon eine entsprechende obere Schranke haben, sind wir fertig. □

Bemerkung 9.6 *Man kanna aus dieser Überlegung wieder die Situation von i.i.d. Zufallsvariablen zurückgewinnen. Wir wählen $\pi(x,y) = \pi(y)$ für alle x (π die Verteilung der Zufallsvariablen). Dann ist für jedes V die Matrix $\pi(y)e^{v(y)}$ vom Rang 1 und der einzige nicht-verschwindende Eigenwert ist*

$$\sigma(V) = \sum_{x} e^{v(x)}\pi(x).$$

Somit sind wir wieder in der Situation des Satzes von Cramér, denn σ ist die momenten-erzeugende Funktion von V. Insbesondere ist wieder

$$I(q) = \sum_{x} q(x)\log\frac{q(x)}{\pi(x)}$$

die relative Entropie bezüglich π.