Technische Universität Dortmund
Fakultät für Mathematik
Sommersemester 2021

Vorlesungsskript:

Konzentrationsungleichungen

Dozent: Prof. Dr. Ivan Veselić

Dieses Skript ist aus einer vierstündigen Vorlesung entstanden, dass ich 2017/2018 an der TU Dortmund hielt. Es orientierte sich im Wesentlichen an dem Buch: *Concentration Inequalities: A Nonasymptotic Theory of Independence* von Stéphane Boucheron, Gábor Lugosi und Pascal Massart.

Herr Malcherczyk war einer der Hörer und hat im Anschluss ein ausgearbeitetes Skript eines großen Teiles der Vorlesung erstellt. Der vorliegende Text stellt einer Bearbeitung und Erweiterung meinerseits dar und verwendet einige von Christoph Schumacher erzeugte Graphiken. Hörer der Vorlesungen haben durch Hinweise auf Tippfehler zur Verbesserung des Skripts beigetragen.

Allen diesen Personen möchte ich an dieser Stelle danken!

Dortmund, März-Juli 2021 Ivan Veselić

Inhaltsverzeichnis

1. Motivation

Zunächst sollen klassische Situationen als motivierendes Fundament vorgestellt werden, in denen man Konzentrationsungleichungen begegnen kann.

(A) Gesetz der großen Zahlen:
Für unabhängig, identisch verteilte Zufallsvariablen $X_1, \dots, X_n$ in $\mathcal{L}^1(\Omega, P)$, äquivalent durch $E|X_1| < \infty$ ausgedrückt, gilt das *Gesetz der großen Zahlen*:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{} E(X_1). \tag{1.1}$$

Eine andere nützliche Formulierung ist folgende

$$\frac{1}{n}\sum_{i=1}^{n} \big(X_i - E(X_i)\big) \xrightarrow[n\to\infty]{} 0.$$

Letztere Schreibweise kann vor allem in Fällen nützlich sein, in denen wir keine identisch verteilten Zufallsvariablen vorliegen haben. Man beachte aber, dass in solchen Fällen nicht allgemein die Konvergenz (1.1) gelten muss.
Bisher haben wir offen gelassen, welche Art von Konvergenz hier vorliegt. Typischerweise formuliert man das Gesetz der großen Zahlen z.B. in

- fast sicherer Konvergenz (starkes Gesetz der großen Zahlen),
- stochastischer Konvergenz (schwaches Gesetz der großen Zahlen),
- der $\mathcal{L}^2$-Norm.

Ein wichtiger Aspekt bei der Anwendung von Grenzwertsätzen in der Stochastik ist die Konvergenzgeschwindigkeit. Typischerweise fragt man sich, wie groß der Approximationsfehler für endliche n ist. Solche *nicht asymptotische* Fragestellungen sind in der Anwendung wichtig, um die Güte einer Approximation für den Erwartungswert von echten gegebenen Daten $x_1, \dots, x_n$ abzuschätzen.
Dabei werden Abschätzungen der folgenden Bauart angestrebt:

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \big(X_i - E(X_i)\big) \right\| \leq f(n, P_{X_1}),$$

wobei $f(n, P_{X_1})$ eine obere Schranke ist, die von dem Stichprobenumfang n und von der Verteilung P_{X_1} selbst abhängt. Wünschenswert wäre es, wenn $f(n, P_{X_1})$ wenige Informationen über die Verteilung benötigte, um möglichst allgemeine Aussagen zu erhalten. In der $\mathcal{L}^2$-Norm für unabhängige, aber nicht notwendigerweise identisch verteilte, quadrat-integrierbare Zufallsvariablen

können wir folgende Abschätzung angeben

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \big(X_i - E(X_i)\big) \right\|_{\mathcal{L}^2(P)} \leq \frac{\max\limits_{i=1,\ldots,n} \{\sigma_i\}}{\sqrt{n}}.$$

Dabei sind σ_i die Standardabweichungen der ZVen X_i für $i = 1, \ldots, n$.
Die hier vorliegende Abschätzung bildet im Fall von identisch verteilten Zufallsvariablen sogar Gleichheit.

In Abschnitt 2 wird u.a. die Frage untersucht, welche Schranken sich durch Verwendung von höheren Momenten finden lassen.

(B) Rekonstruktion von Verteilungen in statistischer Lerntheorie
Wir betrachten den Datensatz $x_1, \ldots, x_n \in \mathbb{R}$ als Realisationen von unabhängig, identisch verteilten ZVen $X_1, \ldots, X_n$ mit unbekannter Verteilung $P_X = \mu$.
Frage: Wie lässt sich aus geg. Daten die wahre Verteilung rekonstruieren?
Wir verwenden das sogenannte *empirische Maß*

$$\mu_n = \mu_n^{x_1,\ldots,x_n} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

mit δ_{x_i} als das Punktmaß in x_i. Das liefert eine intuitive Möglichkeit für eine Schätzung von μ. Die empirische Verteilungsfunktion aus den Datensatz $x_1, \ldots, x_n$ ist dabei zum empirischen Maß assoziiert.
Auch hier stellt man sich die Frage nach der Güte der Approximation von μ_n zu μ. In welchem Sinne können wir hier überhaupt eine Konvergenz formulieren? Ein elementarer Ansatz ist durch den *Fundamentalsatz der Statistik von Glivenko-Cantelli* gegeben, durch den die Konvergenz der empirischen Verteilungsfunktion gegen die wahren Verteilungsfunktion in der Supremumsnorm geliefert wird.
Auf Ebene der Maßtheorie liegt eine schwache Konvergenz des empirischen Maßes gegen das wahre Wahrscheinlichkeitsmaß für Mengen der Bauart

$$A = (-\infty, x] \text{ für } x \in \mathbb{R}$$

vor. (Hier evtl. noch Arten der schwachen Konvergenz in der Sprache Funktionalanalysis diskutieren.)

Lässt sich die schwache Konvergenz für andere Klassen von Mengen formulieren? Diese Frage wird in einem Exkurs zur statistischen Lerntheorie in

Abschnitt 3 untersucht.

(C) Irrfahrten auf $\mathbb{Z}^2$

Wir betrachten unabhängig, identisch verteilte ZVen der Form

$$X_1, \dots, X_n \colon \Omega \to \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ -1 \end{pmatrix} \right\}.$$

Die Werte der ZVen beschreiben dabei Bewegungsrichtungen im $\mathbb{Z}^2$-Gitter. Die Wahrscheinlichkeit für alle Richtungen soll $\frac{1}{4}$ betragen. Nun summieren wir die ersten n Bewegungsschritte auf und erhalten eine neue ZVe

$$Z_n := \sum_{i=1}^{n} X_n.$$

Sie beschreibt für verschiedene Zeiten $i = 1, \dots, n$ einen Pfad auf $\mathbb{Z}^2$. (Hier evtl. noch eine Abbildung einfügen.) Man nennt solche Prozesse *Irrfahrten*.

Wir fragen uns in diesem Kontext beispielsweise, wie sich der euklidische Abstand $\|Z_n\|$ im Verlauf eines Pfades typischeweise verhält. Es kann z.B. nach einer oberen Schranke für $\max\limits_{i=1,\dots,n} \|Z_i\|$ gefragt werden. Eine sehr einfache Antwort wäre

$$\max_{i=1,\dots,n} \|Z_i\| \leq n,$$

da jeder Schritt maximal Länge 1 hat. Das ist allerdings eine grobe Abschätzung.
Mit dem *Zentralen Grenzwertsatz* können wir oft Aussagen der folgenden Bauart

$$\max_{i=1,\dots,n} \|Z_i\| \leq C \cdot \sqrt{n}$$

gewinnen. Die Frage, wann solche Abschätzungen gelten und wovon die Konstante C abhängt, bleibt noch offen. Unklar bleibt auch, wie die Verteilungen P_{X_i} in die Abschätzung eingehen.

Optional (D) Brownsche Bewegung

Hier interssieren wir uns für Suprema von stochastischen Prozessen. Ein prominentes Beispiel dafür ist die Brownsche Bewegung mit Zeithorizont T

$$B \colon \Omega \times [0, T] \to \mathbb{R}.$$

Mit einer funktionalen Version des zentralen Grenzwertsatzes gilt:

$$\text{Irrfahrt} \xrightarrow[\text{Donsker}]{} \text{Brownsche Bewegung}$$

$t \mapsto B_\omega(t)$ ist f.s. nicht differenzierbare Trajektorie.
Ferner: Für jede Zeit $t > 0$ gilt $\sup_{\omega\in\Omega} \|B_t(\omega)\| = \infty$. Aber es gilt wiederum:

$$\sup_{t\in[0,T]} \|B_t\| \leq C\sqrt{T},\ C > 0.$$

Ziel: Präzisiere die Konstante C und die Wahrscheinlichkeit!

2. Grundlegende Ungleichungen

In diesem Kapitel wollen wir erste Beispiele von Konzentrationsungleichungen kennenlernen.

2.1. **Markov-Ungleichung und Co.**

Sei X $\mathbb{R}$-wertige ZVe auf Wahrscheinlichkeitsraum (Ω, P) mit endlichem Erwartungswert $E(X)$.
Frage: Um wie viel weicht X von seinem Erwartungswert $E(X)$ ab?
Wir wollen obere Schranken für $t > 0$ der folgenden Form finden:

$$P(X - E(X) \geq t) \leq \dots$$
$$P(X - E(X) \leq -t) \leq \dots$$

einfachste Antwort: *Markov-Ungleichung*
Sei $Y \geq 0$ eine ZVe mit $E(Y) < \infty$. Dann gilt für alle $t \geq 0$:

$$Y \cdot \mathbb{1}_{\{Y\geq t\}} \geq t \cdot \mathbb{1}_{\{Y\geq t\}} \text{ auf } \Omega$$

Integration liefert:

$$E\left(Y \cdot \mathbb{1}_{\{Y\geq t\}}\right) \geq t \cdot E\left(\mathbb{1}_{\{Y\geq t\}}\right) = t \cdot P(Y \geq t).$$

Weitere Abschätzung der linken Seite:

$$E\left(Y \cdot \mathbb{1}_{\{Y\geq t\}}\right) \leq E(Y).$$

Falls die Dichtefunktion f von Y gegeben ist, ist es natürlich, den Wertebereich von Y auf die horizontale Achse aufzutragen.
Zusammenfassend: Für jede ZVe $Y\colon \Omega \to [0, \infty)$ in $\mathcal{L}^1(\Omega, P)$ gilt nach obigen Ausführungen für $t > 0$ die Markov-Ungleichung:

$$P(Y \geq t) \leq \frac{1}{t} E\left(Y \cdot \mathbb{1}_{\{Y\geq t\}}\right) \leq \frac{1}{t} E(Y). \tag{2.1}$$

Setze $Y = |E(X) - X|$. Das liefert eine erste Antwort:

$$P(X - E(X) \geq t) + P(X - E(X) \leq -t) = P(Y \geq t) \leq \frac{E(Y)}{t} = \frac{E(|X - E(X)|)}{t}.$$

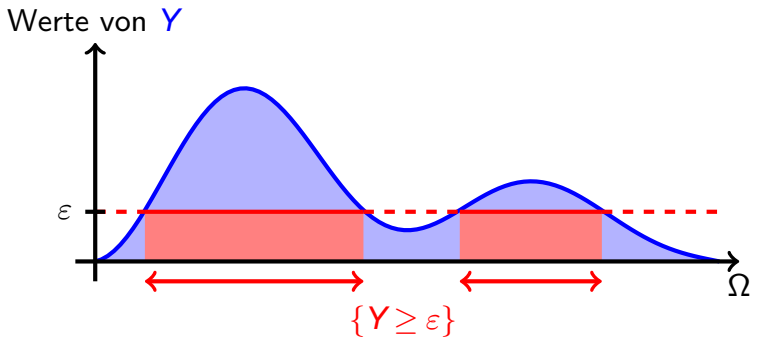


ABBILDUNG 1. Ω auf der horizontalen Achse.

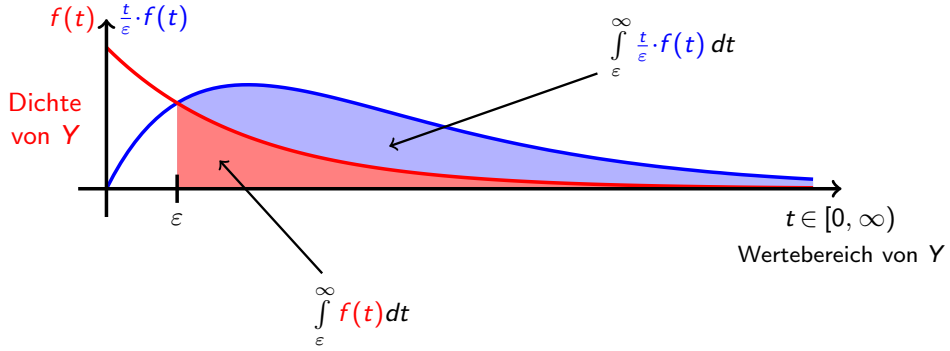


ABBILDUNG 2. Wertebereich auf der horizontalen Achse.

Frage: Gibt es eine bessere Wahl von Y in (2.1)?

Hat X z.B. eine endliche Varianz $\mathrm{Var}(X)$, so gilt für $\Phi(Y)$ mit $\Phi(y) = y^2$

$$E(\Phi(Y)) = E(Y^2) = E(|X - E(X)|^2) = \mathrm{Var}(X) < \infty \quad \Rightarrow \quad \Phi(Y) \in \mathcal{L}^1(\Omega, P).$$

Wende nun für $t \geq 0$ die Markov-Ungleichung an:

$$P(|X - E(X)| \geq t) = P(\Phi(Y) \geq \Phi(t)) \leq \frac{E(\Phi(Y))}{\Phi(t)} = \frac{Var(X)}{t^2} \tag{2.2}$$

und erhalte damit die *Markov-Čebyšev-Ungleichung*. Die Ungleichung (2.2) gilt für sämtliche isotone (monoton wachsende) Funktionen $\Phi\colon I \to [0, \infty)$ auf einem Intervall $I \subset \mathbb{R}$, so dass $\Phi(t) > 0$ ist. [1] Für die Anwendbarkeit der

[1] In dieser Vorlesung benutzen wir der Kürze halber den Begriff *isoton* für monoton wachsend und *antiton* für monoton fallend.

Methode braucht man

$$\Phi(Y) = \Phi(|X - E(X)|) \in \mathcal{L}^1(\Omega, P).$$

Gilt für eine ZVe X: $E|X^q| < \infty \; \forall\, q \in \mathbb{N}$, erhalten wir $\forall\, q, t \geq 0$:

$$P(|X - E(X)|) \geq t) \leq \frac{E(|X - E(X)|^q)}{t^q}.$$

$\rightarrow$ schöne Form, da linke Seite unabhängig von q ist (Optimierungsaspekt)

$$\Rightarrow \quad P(|X - E(X)|) \geq t) \leq \inf_{q>0} \frac{E(|X - E(X)|^q)}{t^q}. \tag{2.3}$$

Erstes Fazit: Je mehr Informationen über eine ZVe vorliegen, desto breiter das Spektrum an potentiellen Abschätzungen und Methoden.

Warum spielt der Fall $\mathbf{q = 2}$ eine zentrale Rolle?
Seien $X_1, \ldots, X_n$ unabhängige ZVen mit endlichen Varianzen. Für $Z = \sum_{i=1}^n X_i$ gilt nach dem Additionssatz (Bienaymé):

$$\mathrm{Var}(Z) = \sum_{i=1}^n \mathrm{Var}(X_i).$$

$\rightarrow$ Formulierung einer Konzentrationsungleichung für gemittelte ZVen mit Hilfe von Ungleichung (2.2):

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - E(X_i))\right| > t\right) &= P\left(\left|\sum_{i=1}^n (X_i - E(X_i))\right| > t \cdot n\right) \\ &\leq \frac{\mathrm{Var}(Z)}{t^2 \cdot n^2} = \frac{\sigma^2}{t^2 \cdot n}, \end{aligned}$$

wobei $\sigma^2 := \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(X_i)$ die gemittelte Varianz ist.

Bemerkung: In dieser Vorlesung spielt die Eigenschaft der *Identischen Verteilung* eine weniger zentrale Rolle als die Unabhängigkeit, da uns nicht der explizite Wert des Limes interessiert, sondern gute Abschätzungen für *endliches* n (bzw. die Konvergenzordnung). Natürlich vereinfachen sich viele Aussagen, falls die ZVen identisch verteilt sind.

2.2. Cramér-Chernoff-Methode.

Die Methoden dieses Kapitels werden in den nachfolgenden Kapiteln 2.3 - 2.9 benötigt. Zusammenhänge zur Entropie werden in Kapitel 4.9 dargestellt. In Kapitel 5.2, 5.4, 5.5 taucht sie ebenso auf.

Idee: Wähle statt $\Phi(t) = t^2$ nun $\Phi(t) = e^{\lambda t}$ für $\lambda > 0$.
Analog zu (2.2) mit $\Phi(y) = e^{\lambda y}$ nach Anwendung der Markov-Ungleichung:

$$P(X \geq t) = P\left(e^{\lambda X} \geq e^{\lambda t}\right) \leq \frac{E\left(e^{\lambda X}\right)}{e^{\lambda t}}. \tag{2.4}$$

$\rightarrow$ Schranke mit exponentiellem Abfall gewonnen.
Studiere nun die Schranken genauer. Die folgende Abbildung

$$M\colon \mathbb{R} \to [0, \infty],\ M(\lambda) := E\left(e^{\lambda X}\right)$$

nennen wir die *momentenerzeugenden Funktion* von X (kurz: MEF von X). Sie ist gegebenenfalls unendlich. Betrachte $Z = \sum_{i=1}^n X_i$ mit unabhängigen ZVen $X_1, \ldots, X_n$. Für die MEF von $Z - E(Z)$ gilt dann

$$E\left(e^{\lambda \sum_{i=1}^n (X_i - E(X_i))}\right) = \prod_{i=1}^n E\left(e^{\lambda (X_i - E(X_i))}\right),$$

wegen der Unabhängigkeit der $X_1, \ldots, X_n$. Sind $(X_i - E(X_i))$ zusätzlich identisch verteilt mit MEF $M(\lambda) := M_{X_1 - E(X_1)}(\lambda)$, so gilt mit (2.4):

$$P\left(\frac{1}{n}(Z - E(Z)) \geq t\right) \leq \frac{\prod_{i=1}^n M(\lambda)}{e^{\lambda t n}} = \frac{(M(\lambda))^n}{e^{\lambda t n}}.$$

Vorgehensweise bisher:

- Gewinne Klassen von Ungleichungen für verschiedene λ
- Schranke über den Parameter λ optimieren (Minimierungsaufgabe)
- Lösen der Optimierungsaufgabe liefert eine (hoffentlich) gute Abschätzung

Bemerkungen:

(a) Im Allg. liefern polynomielle Transformationen $\Phi(t) = t^q$ aus (2.3) bessere Schranken als exponentielle Transformationen $\Phi(t) = e^{\lambda t}$ aus (2.4). D.h.: für beliebige $t > 0$ und ZVe $X \geq 0$:

$$\inf_{q>0} \frac{E(X^q)}{t^q} \leq \inf_{\lambda > 0} \frac{E(e^{\lambda X})}{e^{\lambda t}}.$$

Beweisidee: Taylorentwicklung der Exponentialfunktion (Übung).

(b) Wir interessieren uns für die Wahrscheinlichkeit der Abweichungen vom Mittelwert mit folgender Bauart:

$$P\left(|Z - E(Z)| \geq t\right) = P\left(Z - E(Z) \geq t\right) + P\left(Z - E(Z) \leq -t\right) \text{ für } t > 0.$$

Wegen (b) genügt es o.B.d.A. die zentrierte Version der ZVe $\tilde{Z} := (Z - E(Z))$ zu betrachten und Abschätzungen für $P\left(|\tilde{Z}| \geq t\right)$ herzuleiten.

2.2.1. *Die kumulantenerzeugenden Funktion und die Cramér-Transformierte.* Sei Z ZVe mit MEF $M_Z(\lambda)$, d.h. (äquivalent zu (2.4)):

$$P(Z \geq t) \leq e^{-\lambda t} M(\lambda) = e^{-\lambda t + \ln(M(\lambda))}. \tag{2.5}$$

Fasse die obere Schranke als Klasse von Funktionen auf, um sie dann zu minimieren. Dazu definieren wir die *kumulantenerzeugende Funktion* $\psi_Z(\lambda)$ (kurz: KEF):

$$\psi_Z(\lambda) := \ln(M(\lambda)) = \ln(E(e^{\lambda Z})).$$

Betrachte dazu die sogenannte *Cramér-Transformierte*:

$$\psi_Z^*(t) := \sup_{\lambda \geq 0} (\lambda t - \psi_Z(\lambda)) \text{ für } \lambda \geq 0.$$

Die Tansformierte ψ^* einer konvexen Funktion ψ wird abstrakt in Lemma 2.14 untersucht. Bedeutung der Cramér-Transformierte:
Minimiere die Schranke von (2.5) über λ und betrachte nur noch den Exponenten. Beachte: Vorzeichenwechsel liefert Maximierungsproblem:

$$P(Z \geq t) \leq \inf_{\lambda \geq 0} e^{-\left(\lambda t - \ln(M(\lambda))\right)} = e^{-\psi_Z^*(t)}.$$

Untersuche nun den Definitions- und Wertebereich der Cramér-Trafo ψ_Z^*. Zunächst bemerkt man eine Eigenschaft der KEF für beliebige ZVen Z:

$$\psi_Z(0) = \ln(1) = 0.$$

Dieser Zusammenhang ist nützlich für Randwertuntersuchungen. Auf die Cramér-Trafo überträgt sich dies, wenn man für $\lambda = 0$ setzt, wie folgt:

$$\psi_Z^*(t) = \sup_{\lambda \geq 0} (\lambda t - \psi_Z(\lambda)) \geq 0 - 0 = 0.$$

Insbesondere wissen wir nun, dass der Wertebereich von ψ_Z^* nichtnegativ ist.

Warum betrachten wir nicht $\lambda \in \mathbb{R}$, sondern nur $\lambda \geq 0$ im Supremum?
Falls für die ZV $Z \in \mathcal{L}^1$ gilt, so gilt nach der Jensen-Ungleichung

$$e^{\lambda E(Z)} \leq E(e^{\lambda Z}) = M(\lambda),$$

wobei der letzte Ausdruck auch unendlich sein könnte. Logarithmieren dieser Ungleichung ergibt

$$\lambda \cdot E(Z) \leq \ln(M(\lambda)) = \psi_Z(\lambda).$$

Betrachte $\lambda < 0$ und $t \geq E(Z)$ und schätze die linke Seite ab

$$\lambda \cdot E(Z) \geq \lambda \cdot t$$

Insgesamt folgt aus beiden Ungleichungen für $\lambda < 0$ und $t \geq E(Z)$:

$$\lambda \cdot t - \psi_Z(\lambda) \leq 0$$

Die Annahme $t \geq E(Z)$ lässt sich allgemein dadurch motivieren, dass wir zentrierte ZVe Z mit $E(Z) = 0$ betrachten wollen.

Fazit: Obige Randbetrachtung für $\lambda = 0$ das liefert, dass das Supremum über λ nicht im negativen Bereich angenommen wird, d.h.

$$(2.6)\quad \tilde{\psi}_Z(t) := \sup_{\lambda \in \mathbb{R}} (\lambda t - \psi_Z(\lambda)) = \sup_{\lambda \geq 0} (\lambda t - \psi_Z(\lambda)) = \psi_Z^*(t) \text{ für } t \geq E(Z).$$

Wir nennen die Funktion $\tilde{\psi}_Z$ in (2.6) die *Fenchel-Legendre-Transformierte* oder auch *Fenchel-Legendre-Duale* von ψ_Z.

Nicht jedes $t \geq 0$ liefert brauchbare Chernoff-Schranken. Falls $\psi_Z^*(t) = 0$ ist, ergibt sich eine triviale Schranke $e^{-\psi_Z^*(t)} = 1$. In welchen Fällen tritt dies noch ein?
Einerseits ist der Fall $\psi_Z(\lambda) \equiv \infty$ für $\lambda > 0$ problematisch. Dann folgt

$$\psi_Z^*(t) = \sup_{\lambda \geq 0} (\lambda t - \psi_Z(\lambda)) = 0.$$

Andererseits ist der Fall $t \leq E(Z)$ problematisch wegen

$$\lambda t \leq \lambda E(Z) \leq \psi_Z(\lambda) \quad \text{für } \lambda \geq 0$$
$$\Rightarrow \quad \lambda t - \psi_Z(\lambda) \leq 0 \text{ und } = 0 \text{ für } \lambda = 0.$$

Um diese Situation zu vermeiden, nehmen wir in diesem Kapitel an, dass ein $\lambda_0 > 0$ existiert, so dass $E\left(e^{\lambda_0 Z}\right) < \infty$ gilt. Mit der Hölder-Ungleichung lässt sich zeigen, dass dann auch das exponentielle Moment für $\lambda \leq \lambda_0$ existiert (Übung). Es gilt dann also:

$$\text{für alle } \lambda \in [0, \lambda_0] : E(e^{\lambda Z}) < \infty.$$

Setze dazu die Zahl $b := \sup\{\lambda \geq 0 \mid E(e^{\lambda Z}) < \infty\} \in [0, \infty]$. Da die Voraussetzung $E(e^{\lambda Z}) < \infty$ für $\lambda = 0$ immer erfüllt ist, genügt es auch hier statt $\lambda \in \mathbb{R}$ nur den Bereich $\lambda \geq 0$ zu untersuchen. In den meisten Fällen ist $b \in \{0, \infty\}$. Dagegen ist für exponentialverteilte ZVen der Wert b gerade der Parameter der Exponentialverteilung.

Bemerkung 2.1 (Eigenschaften der kumulantenerzeugenden Funktion). Folgende Eigenschaften werden sich für Optimierungsaufgaben als nützlich erweisen:

(a) $\psi = \psi_Z$ ist konvex auf $I := (0, b)$ (auch gültig, falls $b = \infty$)
(b) ψ ist strikt konvex auf I, falls Z nicht fast sicher konstant
(c) $\psi \colon I \to \mathbb{R}$ ist $\mathcal{C}^\infty$

Für zentrierte Zufallsvariablen Z gilt darüber hinaus:

(d) $\psi_Z : [0, b) \to \mathbb{R}$ ist $\mathcal{C}^1$. Beachte: 0 ist *zusätzlich* im Definitionsbereich.
(e) $\psi'_Z(0) = 0$ (zusätzlich zu $\psi_Z(0) = 0$)
(f) Es genügt die Cramér-Trafo auf dem Intervall I zu bestimmen:

$$\psi_Z^*(t) = \sup_{\lambda \geq 0} (\lambda t - \psi_Z(\lambda)) = \sup_{\lambda \in I} (\lambda t - \psi_Z(\lambda))$$

Die Beweise von (a) - (f) werden als Übungsaufgaben gestellt.

2.2.2. *Bestimmung des Supremums mittels der Ableitung.* Ansatz:

- ψ_Z ist $\mathcal{C}^1 \ \rightarrow$ mittels Ableitung stationäre Punkte berechnen
- strikte Konvexität liefert Eindeutigkeit des Optimums auf I

$$\begin{aligned} 0 &= \frac{d}{d\lambda}(\lambda t - \psi(\lambda)) = t - \psi'(\lambda) \\ \Leftrightarrow \ t &= \psi'(\lambda) \end{aligned}$$

Sei λ_t eine Lösung dieser Gleichung. Falls man den trivialen Fall einer fast sicher konstanten Zufallsvariable ausschließt, ist ψ strikt konvex.
$\Rightarrow \quad \lambda_t$ ist eindeutig.

Definition: Sei $B := \psi'_Z(b)$. Dann ist $\psi'_Z : I \to (0, B)$ wegen strikter Monotonie bijektiv mit strikt monotoner Inversen $(\psi'_Z)^{-1}$.

$$\text{Daher gilt für alle } t \in (0, B) : \ \lambda_t = (\psi'_Z)^{-1}(t).$$

Diese Formel können wir nun nutzen, um für konkrete Verteilungen die Cramér-Trafo auszurechnen.

2.2.3. *Cramér-Transformierte verschiedener Verteilungsklassen.*

(a) Cramér-Transformierte für zentrierte Normalverteilungen:

Die Kenntnis der Cramér-Transformierten der zentrierten Normalverteilung ist beim Verständnis von sub-gaußschen Zufallsvariablen in Kapitel 2.2 relevant. Sei $Z \sim \mathcal{N}(0, \sigma^2)$ mit Varianz σ^2. Beachte, dass wir Zentriertheit brauchen.

- zunächst berechne MEF: $M(\lambda) = e^{\frac{\lambda^2\sigma^2}{2}}$ (Übung)
- MEF ergibt sofort die KEF $\psi_Z(\lambda) = \frac{\lambda^2\sigma^2}{2}$.
- erste Ableitung ist $\psi_Z'(\lambda) = \lambda\sigma^2$
- obige Formel ergibt $\lambda_t = (\psi')^{-1}(t) = \frac{t}{\sigma^2}$ als Lösung der Optimierungsaufgabe

$\forall\, t > 0$ besitzt die Cramér-Trafo also die folgende Gestalt

$$\begin{aligned}\psi_Z^*(t) &= \lambda_t t - \psi_Z(\lambda_t) \\ &= \frac{t^2}{\sigma^2} - \frac{\sigma^2 t^2}{2\sigma^4} \\ &= \frac{t^2}{2\sigma^2}.\end{aligned}$$

Das liefert die Chernoff-Schranke für $\forall\, t > 0 : P(Z \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$.

Wie gut ist diese Schranke? Kann man sie noch verbessern? Zur Beantwortung dieser Fragen formuliere die Chernoff-Abschätzung um

$$P(Z \geq t) \cdot e^{\frac{t^2}{2\sigma^2}} \leq 1.$$

Diese Abschätzung lässt sich aber verbessern.

$$\text{Für alle } t > 0 \text{ gilt } P(Z \geq t) \cdot e^{\frac{t^2}{2\sigma^2}} \leq \frac{1}{2} \text{ (Übung).}$$

$\rightarrow$ globale Vorfaktor $\frac{1}{2}$ wird verschenkt, Größenordnung passt zumindest. Zudem kann man zeigen, dass letztere Abschätzung scharf ist:

$$\sup_{t>0} \left(P(Z \geq t) \cdot e^{\frac{t^2}{2\sigma^2}} \right) = \frac{1}{2} \text{ (ebenfalls Übung).}$$

Unter Normalität sind Gewinnungen von Abschätzungen mittels anderer Techniken noch einfach, wodurch wir einen Vergleich zwischen ihnen und der Chernoff-Methode erhalten. In anderen Fällen ist dies nicht mehr so einfach möglich.

(b) Cramér-Transformierte für zentrierte Poisson-Verteilungen:

Die Kenntnis der KEF der zentrierten Poisson-Verteilung wird im Beweis

der Bennett-Ungleichung 2.20 in Kapitel 2.7 benutzt. Sei $Y \sim Poi(\nu)$ für ein $\nu > 0$. Nach Definition der Poisson-Verteilung gilt

$$\forall\, k \in \mathbb{N}_0 :\ P(Y = k) = \frac{\nu^k}{k!} \cdot e^{-\nu}.$$

$\Rightarrow \quad E(Y) = \nu$. Wir arbeiten mit zentrierten ZVen und definieren daher $Z := Y - E(Y)$ mit $E(Z) = 0$. Berechne im ersten Schritt die MEF:

$$\begin{aligned} M_Z(\lambda) = E(e^{\lambda Z}) &= e^{-\lambda\nu} \sum_{k\in\mathbb{N}_0} \left(e^{\lambda k} \cdot \frac{\nu^k}{k!}\right) \cdot e^{-\nu} \\ &= e^{-\lambda\nu-\nu} \sum_{k\in\mathbb{N}_0} \frac{\left(\nu e^{\lambda}\right)^k}{k!} = e^{-(\lambda+1)\nu} \cdot e^{\nu e^{\lambda}}. \end{aligned}$$

Logarithmieren liefert für $\lambda > 0$ die KEF ψ_Z. Berechne außerdem ψ'_Z

$$\psi_Z(\lambda) = \nu\left(e^{\lambda} - \lambda - 1\right),\ \psi'_Z(\lambda) = \nu\left(e^{\lambda} - 1\right).$$

Ansatz: $t = \psi'_Z(\lambda)$, um λ_t als Lösung des Optimierungsproblems herzuleiten:

$$\begin{aligned} & t = \psi'_Z(\lambda) = \nu\left(e^{\lambda_t} - 1\right) \\ \Leftrightarrow\ & e^{\lambda_t} = \frac{t}{\nu} + 1 \\ \Leftrightarrow\ & \lambda_t = \ln\left(\frac{t}{\nu} + 1\right). \end{aligned}$$

Der optimierende Parameter λ_t liefert nun die Gestalt der Cramér-Trafo:

$$\begin{aligned} \psi^*_Z(t) &= t\lambda_t - \psi_Z(\lambda_t) \\ &= t\ln\left(\frac{t}{\nu} + 1\right) - \nu\left(\frac{t}{\nu} + 1 - \ln\left(\frac{t}{\nu} + 1\right) - 1\right) \\ &= (t+\nu) \cdot \ln\left(\frac{t}{\nu} + 1\right) - t \\ &= \nu \cdot h\left(\frac{t}{\nu}\right), \end{aligned}$$

wobei wir für $x \geq -1$ die Funktion h einführen:

$$h(x) := (1+x) \cdot \ln(1+x) - x.$$

Die ZVe $-Z$ ist ebenfalls zentriert und analog folgt:

$$\psi^*_{-Z}(t) = \nu \cdot h\left(-\frac{t}{\nu}\right),\ \text{sofern } t \leq \nu \text{ gilt.}$$

(c) Cramér-Transformierte für zentrierte Bernoulli-Verteilungen:
Eigentliches Ziel: Cramér-Trafo von binomialverteilten ZVen.
Betrachte zunächst in (c) Bernoulli-ZVen und in (d) Summen von unabhängigen ZVen, um in Teil (e) die Binomialverteilung zu untersuchen.

Sei $Y : \Omega \to \{0, 1\}$ eine ZVe mit $P(Y = 1) = p = 1 - P(Y = 0)$ für ein $p \in [0, 1]$. Der Erwartungswert ist dann $E(Y) = p$. Die ZVe $Z := Y - p$ ist dann die zentrierte Version. Die MEF von Z ergibt sich aus der Definition

$$M_Z(\lambda) = \left(p \cdot e^{\lambda} + (1 - p)\right) e^{-\lambda p},$$

mit KEF

$$\psi_Z(\lambda) = -\lambda p + \ln\left(pe^{\lambda} + 1 - p\right).$$

Schließlich folgt mit gleicher Strategie (Übung) für $t \in (0, 1 - p)$:

$$\begin{aligned}\psi_Z^*(t) &= (1 - p - t) \cdot \ln\left(\frac{1 - p - t}{1 - p}\right) + (p + t) \cdot \ln\left(\frac{p + t}{p}\right) \\ &= (1 - a) \cdot \ln\left(\frac{1 - a}{1 - p}\right) + a \cdot \ln\left(\frac{a}{p}\right) \\ &=: h_p(a).\end{aligned}$$

Hierbei haben wir $a := t + p$ gesetzt, so dass $a \in (p, 1)$.

Definition 2.2. Die Funktion $h_p \colon (0, 1) \to \mathbb{R}$, $h_p(a) = (1-a) \cdot \ln\left(\frac{1 - a}{1 - p}\right) + a \cdot \ln\left(\frac{a}{p}\right)$ nennt man die *Kulback-Leibler-Divergenz* $D(P_a \| P_p)$ zwischen zwei Bernoulli-Verteilungen mit Parameter a bzw. p.

Wir werden im Kapitel 4 den Begriff der *Entropie* einführen und auch die Funktion h_p diesem Begriff zuordnen können (vgl. insbesondere die Dualitätsformel 4.38 aus Kapitel 4.9.). Die Schranken einiger Konzentrationsungleichungen entsprechen solchen Entropien.

(d) Cramér-Transformierte für Summen unabhängiger Zufallsvariablen:
Die Cramér-Chernoff-Methode erlaubt einfachen Umgang mit Summen von unabhängig, identisch verteilten ZVen $X_1, \ldots, X_n$. Dazu betrachten wir die ZVe $Z := \sum_{i=1}^n X_i$ und die KEF ψ_{X_1}, sowie die Cramér-Trafo $\psi_{X_1}^*$ für X_1.

Ziel: Darstellung der KEF von Z in Abhängigkeit der KEF der $X_1, \ldots, X_n$.

Für λ mit $\psi_{X_1}(\lambda) < \infty$ bestimmen wir die KEF von Z:

$$\begin{aligned}
\psi_Z(\lambda) &= \ln\left(E\left(e^{\lambda Z}\right)\right) \;=\; \ln\left(E\left(e^{\lambda \sum_{i=1}^n X_i}\right)\right) \\
&= \ln\left(\prod_{i=1}^n \left(E\left(e^{\lambda X_i}\right)\right)\right) \text{ (wegen Unabhängigkeit)} \\
&= \ln\left(E\left(e^{\lambda X_1}\right)^n\right) \text{ (wegen identischer Verteilung)} \\
&= n \cdot \ln\left(E\left(e^{\lambda X_1}\right)\right) \\
&= n \cdot \psi_{X_1}(\lambda).
\end{aligned} \tag{2.7}$$

Falls die X_i nur unabhängig sind, erhalten wir mit denselben Schritten zumindest

$$\psi_Z(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda). \tag{2.8}$$

Bemerkung 2.3. Wir nehmen in diesem Abschnitt durchgängig an, dass die betrachtete ZV X integrierbar ist und ein $\lambda > 0$ existiert mit $M(\lambda) < \infty$, was sich dann auch auf die KEF überträgt. Insbesondere ist $b > 0$ in der Definition des Intervalls $I = (0, b)$. Im Spezialfall einer f.s. konstanten ZV gilt

$$\psi^*(t) = \begin{cases} 0, & \text{falls } t = EX \\ +\infty, & \text{falls } t > EX, \end{cases}$$

anderenfalls

$$\psi^*(t) = \begin{cases} 0, & \text{falls } t = EX \\ \in (0, \infty), & \text{falls } t > EX. \end{cases}$$

Auch die Cramér-Trafo hat ein einfaches verhalten bei einer i.i.d.-Summe:

Lemma 2.4. *Seien $X, X_1, \ldots, X_n$ unabhängige, identisch verteilten ZVen und $Z := \sum_{i=1}^n X_i$. Dann gilt:*

$$\psi_Z^*(t) = n \cdot \psi_{X_1}^*\left(\frac{t}{n}\right).$$

Beweis. Es sei an die Gestalt des optimierenden Parameters $\lambda_t = (\psi')^{-1}(t)$ der Cramér-Trafo erinnert. Also untersuchen wir:

$$\begin{aligned}
\psi_Z'(\lambda) &\overset{(2.7)}{=} n \cdot \psi_X'(\lambda) \\
&= \mathcal{M}_n(\psi_X'(\lambda)) = (\mathcal{M}_n \circ \psi_X')(\lambda)
\end{aligned}$$

mit $\mathcal{M}_n(x) = n \cdot x$. Außerdem ist $(\mathcal{M}_n)^{-1}(x) = \frac{x}{n}$.

$$\Rightarrow \quad \lambda_t = (\psi_Z')^{-1}(t) = (\psi_X')^{-1}((\mathcal{M}_n)^{-1}(t)) = (\psi_X')^{-1}\left(\frac{t}{n}\right).$$

Für die Cramér-Trafo von Z folgt schließlich

$$\begin{aligned}\psi_Z^*(t) &= t\lambda_t - \psi_Z(\lambda_t) \\ &= t(\psi_Z')^{-1}(t) - \psi_Z\left((\psi_Z')^{-1}(t)\right) \\ &= t(\psi_X')^{-1}\left(\frac{t}{n}\right) - n\psi_X\left((\psi_X')^{-1}\left(\frac{t}{n}\right)\right) \\ &= n\psi_X^*\left(\frac{t}{n}\right)\end{aligned}$$

□

(e) Cramér-Transformierte für zentrierte Binomialverteilungen::
Jetzt sind alle Vorbereitungen getroffen, die Cramér-Trafo der Binomialverteilung zu berechnen. Seien $X_1, \ldots, X_n \sim Ber(p)$ unabhängig.
$\Rightarrow$ $Y := \sum_{i=1}^n X_i \sim Bin(n,p)$. Betrachte zentrierte Version $Z := Y - np = \sum_{i=1}^n (X_i - p)$, da ja $E(X_i) = p$. Für $t \in (0, n(1-p))$ gilt nach (d) und (c) mit $a = p + \frac{t}{n}$:

$$\psi_Z^*(t) = n \cdot \psi_{X-p}^*\left(\frac{t}{n}\right) = n \cdot h_p\left(\frac{t}{n} + p\right),$$

wobei h_p die in (c) bereits eingeführte Kulback-Leibler-Divergenz zwischen zwei Bernoulli-Verteilungen ist.
Wir erhalten insbesondere die Chernoff-Schranke:

$$P(Z \geq t) \leq \exp\left(-n \cdot h_p\left(\frac{t}{n} + p\right)\right).$$

→ Guter Vorfaktor n für $n \to \infty$, aber wie schnell und wohin konvergiert h_p für $n \to \infty$? → Noch zu untersuchen!

2.3. **Sub-Gaußsche Zufallsvariablen.**

Gauß-ZVen $\hat{=}$ gut umgängliche Klasse von ZVen
Klasse von Verteilungen, die von Gauß-ZVen dominiert werden, sind ähnlich gut umgänglich.

Definition 2.5. Eine reellwertige, zentrierte ZVe X heißt *sub-gaußsch* mit Varianzfaktor ν, in Zeichen: $P_X \in \mathcal{G}(\nu)$, falls

$$\forall \lambda \in \mathbb{R} : \psi_X(\lambda) \leq \frac{\lambda^2 \nu}{2}. \tag{2.9}$$

Die Klasse aller solchen ZVen, bezeichnen wir (auch) mit $\mathcal{G}(\nu)$. Die Schranke entspricht der KEF einer zentrierten Normalverteilung mit Varianz ν. Wir betrachten also die Klasse von Verteilungen, die durch gaußsche ZVen im Sinne der KEF dominiert werden.
Die Konzept von sub-gaußsch findet in den nachfolgenden Kapiteln 2.5 und 2.9 Anwendung.

Bemerkung 2.6 (Eigenschaften sub-gaußscher Zufallsvariablen). (a) Varianz: Es folgt $\mathrm{Var}(X) \leq \nu$, im Allg. gilt aber $\mathrm{Var}(X) \neq \nu$. Die Ungleichung kann man durch eine Taylorentwicklung zeigen (Übung). Sie ist insbesondere scharf. Betrachte dazu Rademacher-verteilte ZVen.

(b) Für normalverteilte ZVen $X \sim \mathcal{N}(m, \nu)$ ist die momentenerzeugende Funktion $M_X(\lambda) = \exp\left(m\lambda + \frac{\lambda^2}{2}\nu\right)$, wie man durch direkte Rechnung sieht. Also gilt

$$X \sim \mathcal{N}(0, \nu) \Longrightarrow X \in \mathcal{G}(\nu)$$

(c) In der Tat, Bedingung (2.9) kann sogar als Vergleich mit Gauß-ZV verstanden werden:

$$X = Y - E(Y) \in \mathcal{G}(\nu) \quad \Leftrightarrow \quad \psi_X \leq \psi_{\mathcal{N}(0,\nu)} \text{ auf } \mathbb{R}$$

(d) Verträglichkeit mit der Faltung:
Sind $X_1, \ldots, X_n$ unabhängige ZVen mit $X_i \in \mathcal{G}(\nu_i)$ für $i = 1, \ldots, n$, so gilt folgende Stabilitätseigenschaft für $Z = \sum_{i=1}^n X_i \in \mathcal{G}(\sum_{i=1}^n \nu_i)$. Nachweis erfolgt z.B. mit dem Additionssatz der Varianz und der Stabilität der KEF:

$$\psi_Z = \sum_{i=1}^n \psi_{X_i} \leq \sum_{i=1}^n \psi_{\mathcal{N}(0,\nu_i)} = \psi_{\mathcal{N}(0,\sum_{i=1}^n \nu_i)}.$$

(e) Inklusionen

$$\lambda \leq \tilde{\lambda} \Longrightarrow \mathcal{G}(\lambda) \subset \mathcal{G}(\tilde{\lambda})$$

Definition 2.7. Ein messbares $X \colon \Omega \to \{-1, 1\}$ mit $P(X = 1) = P(X = -1) = \frac{1}{2}$ heißt *Rademacher-Zufallsvariable.*

Statt über Ungleichungen für KEF ψ_X kann man die Eigenschaft sub-gaußsch auch über Ungleichungen für Momente oder über Ungleichungen von Abfall bei ∞ (engl. *tail-probability*) charakterisiert werden.

Bemerkung 2.8 (Äquivalente *tail*-Charakterisierungen von sub-gaußsch). Sei $X \in \mathcal{G}(\nu)$. Chernoff liefert für $t > 0$

$$\max\{P(X > t), P(X < -t)\} \leq e^{-\frac{t^2}{2\nu}}.$$

Für $X \in \mathcal{G}(\nu)$ gilt nämlich $\psi_X^*(t) \geq \psi_{\mathcal{N}(0,\nu)}^*(t)$ (Übung) und damit

$$P(X > t) \leq e^{-\psi_X^*(t)} \leq e^{-\psi_{\mathcal{N}(0,\nu)}^*(t)} = e^{-t^2/2\nu}$$

Analoge Rechnung für $P(-X > t)$ wegen $-X \in \mathcal{G}(\nu)$ liefert Behauptung.

Satz 2.9.
Sei $X \in \mathcal{L}^1$ eine zentrierte ZVe.
(a) Gibt es ein $\nu > 0$, so dass $\forall\, s > 0$

$$\max\{P(X > s), P(X < -s)\} \leq e^{-\frac{s^2}{2\nu}}$$

gilt, so folgt $\forall\, q \in \mathbb{N}$:

$$E(X^{2q}) \leq 2q!(2\nu)^q \leq q!(4\nu)^q. \tag{2.10}$$

(b) Existiert ein $C \in (0, \infty)$ mit

$$\forall\ q \in \mathbb{N} : E(X^{2q}) \leq q!C^q, \tag{2.11}$$

so gilt $X \in \mathcal{G}(4C)$. Insbesondere folgt daraus

$$\max\{P(X > s), P(X < -s)\} \leq e^{-\frac{s^2}{8C}}. \tag{2.12}$$

Theorem 2.9 kann als Hilfsmittel dienen, um die Voraussetzung der Bernstein-Ungleichung 2.22 nachzurechnen, wie es im Beweis vom Johnson-Lindenstrauß-Lemma 2.26 in 2.9 getan wird.

Beweis. zu (a): Gilt $X \in \mathcal{G}(\nu)$, so ist $Y = \frac{1}{\sqrt{\nu}} X \in \mathcal{G}(1)$, denn

$$\psi_Y(\lambda) = \ln\left(E\left(e^{\lambda Y}\right)\right) = \ln\left(E\left(e^{\frac{\lambda X}{\sqrt{\nu}}}\right)\right) = \psi_X\left(\frac{\lambda}{\sqrt{\nu}}\right) \leq \frac{\frac{\lambda^2}{\nu} \cdot \nu}{2} = \frac{\lambda^2}{2}.$$

Betrachte zunächst den Fall $\nu = 1$ und beginne auf der linken Seite von (2.10):

$$E\left(Y^{2q}\right) = \int_0^\infty P\left(|Y|^{2q} > x\right)\, \mathrm{d}x.$$

Erste Substitution mit $y = x^{\frac{1}{2q}}$ ergibt

$$= 2q \int_0^\infty P(|Y| > y) \cdot y^{2q-1}\, \mathrm{d}y.$$

Anwendung der Voraussetzung ermöglicht folgende Abschätzung:

$$\leq 4q \int_0^\infty e^{-\frac{y^2}{2}} \cdot y^{2q-1} \, \mathrm{d}y.$$

Mit zweiter Substitution $t = \frac{y^2}{2}$ erhalten wir

$$= 4q \int_0^\infty e^{-t} \cdot (2t)^{q-\frac{1}{2}} \cdot (2t)^{-\frac{1}{2}} \, \mathrm{d}t.$$

Diese Substitution ist nützlich, da das Integral explizite Darstellung hat (vgl. Bronstein):

$$E\left(Y^{2q}\right) \leq 2^{q+1} q!.$$

Für allgemeinere Varianzen ν folgt:

$$E\left(X^{2q}\right) = E\left(\left(\sqrt{\nu}Y\right)^{2q}\right) = \nu^q E\left(Y^{2q}\right) \leq 2^{q+1}\nu^q q! \leq 2 \cdot 2^q \nu^q q! = (4\nu)^q q!.$$

zu (b): Es gelte $E\left(X^{2q}\right) \leq q!\, C^q$.
Sei $\tilde{X}$ eine unabhängig, identisch verteilte Kopie von X.
$\Rightarrow$ $X - \tilde{X}$ ist symmetrisch verteilt, d.h. $P(X - \tilde{X} > s) = P(\tilde{X} - X > s) \; \forall\, s \in \mathbb{R}$. Wir nutzen den Satz zur monotone Konvergenz und dass die ungeraden Momente verschwinden:

$$\begin{aligned}
&E\left(e^{\lambda X}\right) \cdot E\left(e^{-\lambda X}\right) \overset{\text{id. vert.}}{=} E\left(e^{\lambda X}\right) \cdot E\left(e^{-\lambda \tilde{X}}\right) \overset{\text{unabh.}}{=} E\left(e^{-\lambda(X-\tilde{X})}\right) \\
&= E\left(\sum_{q \in \mathbb{N}_0} \left[\frac{\lambda^{2q}(X-\tilde{X})^{2q}}{(2q)!} + \frac{\lambda^{2q+1}(X-\tilde{X})^{2q+1}}{(2q+1)!}\right]\right) \\
&= \sum_{q \in \mathbb{N}_0} \left[E\left(\frac{\lambda^{2q}(X-\tilde{X})^{2q}}{(2q)!}\right) + \underbrace{E\left(\frac{\lambda^{2q+1}(X-\tilde{X})^{2q+1}}{(2q+1)!}\right)}_{=0 \text{ wegen Symmetrie}}\right] \\
&= \sum_{q \in \mathbb{N}_0} \frac{\lambda^{2q} E\left(\left(X-\tilde{X}\right)^{2q}\right)}{(2q)!} \; \forall\, \lambda \in \mathbb{R}
\end{aligned}$$

Frage/Übung: Existieren überhaupt die ungeraden Momente? Sind sie summierbar?
Wegen Konvexität von $x \mapsto x^{2q} = x^m$ für $m \in 2\mathbb{N}$ folgt (vgl. auch Abb. ??):

$$(a-b)^m \leq 2^{m-1}(a^m - b^m) \leq 2^{m-1}(a^m + b^m)$$

$$\Rightarrow \quad E\left((X-\tilde{X})^{2q}\right) \leq 2^{2q-1}\left(E\left(X^{2q}\right) + E(\tilde{X}^{2q})\right) \overset{\text{id. vert.}}{=} 2^{2q} E\left(X^{2q}\right)$$

$$\Rightarrow \quad E\left(e^{\lambda X}\right) \cdot E\left(e^{-\lambda X}\right) \overset{\text{s.o.}}{=} \sum_{q\in\mathbb{N}_0} \frac{\lambda^{2q} E\left((X-\tilde{X})^{2q}\right)}{(2q)!}$$

$$\leq \sum_{q\in\mathbb{N}_0} \frac{\lambda^{2q}}{(2q)!} 2^{2q} \underbrace{E\left(X^{2q}\right)}_{\leq q! C^q \text{ nach Vor.}}$$

$$\leq \sum_{q\in\mathbb{N}_0} 2^{2q}\lambda^{2q} C^q \frac{q!}{(2q)!}$$

Nebenrechung.: $\frac{(2q)!}{q!} = \prod_{j=1}^{q}(q+j) \geq \prod_{j=1}^{q} 2j = 2^q \cdot q!$

$$\Rightarrow \quad E\left(e^{\lambda X}\right) \cdot E\left(e^{-\lambda X}\right) \overset{\text{s.o.}}{\leq} \sum_{q\in\mathbb{N}_0} 2^{2q}\lambda^{2q} C^q \frac{q!}{(2q)!}$$

$$\overset{\text{N.R.}}{\leq} \sum_{q\in\mathbb{N}_0} 2^{2q}\lambda^{2q} C^q \frac{1}{2^q q!}$$

$$= \sum_{q\in\mathbb{N}_0} \frac{2^q \lambda^{2q} C^q}{q!} \quad = \quad e^{2C\lambda^2}.$$

Da X zentriert, folgt mit der Jensen-Ungleichung:

$$\Rightarrow \quad M_X(\lambda) \leq E\left(e^{\lambda X}\right) \cdot \underbrace{E\left(e^{-\lambda X}\right)}_{\geq 1} \leq e^{2C\lambda^2}$$

$$\Rightarrow \quad \text{Also: } \psi_X(\lambda) \leq \frac{4C}{2}\lambda^2 \quad \Rightarrow \quad X \in \mathcal{G}(4C)$$

□

Werfen wir noch einen zweiten Blick auf das *tail*-Verhalten:

Lemma 2.10. *Die Momentenbedingung* (2.11) *ist äquivalent zu folgender Bedingung :*

$$\exists \alpha > 0, \text{ so dass } E\left(e^{\alpha X^2}\right) \leq 2. \tag{2.13}$$

Interpretation: Da exp(.) schnell wächst, muss αX^2 schnell abfallen, falls (2.13) gilt.

Beweis. Nach Voraussetzung und Konvergenzsatz

$$2 \geq \sum_{k \in \mathbb{N}_0} \frac{\alpha^k E\left(X^{2k}\right)}{k!}$$

$$\Leftrightarrow \quad 1 \geq \sum_{k \in \mathbb{N}} \frac{\alpha^k E\left(X^{2k}\right)}{k!}$$

Alle Summanden nichtnegativ $\Rightarrow$ $\forall\, k \in \mathbb{N}$ gilt: $E(X^{2k}) \leq \alpha^{-k} k!$.
Satz 2.9 Teil b) $\Rightarrow$ $X \in \mathcal{G}\left(\frac{4}{\alpha}\right)$.
Gegenrichtung: mit Satz 2.9 Teil a)

$$X \in \mathcal{G}(\nu) \quad \overset{\text{Satz 2.9}}{\Rightarrow} \quad E\left(X^{2q}\right) \leq C^q q! \text{ mit } C = 4\nu.$$

Setze: $\alpha = \frac{1}{2C} = \frac{1}{8\nu}$

$$\Rightarrow \quad E\left(e^{\alpha X^2}\right) = \sum_{q \in \mathbb{N}_0} \frac{\alpha^q E\left(X^{2q}\right)}{q!}$$

$$\leq \sum_{q \in \mathbb{N}_0} \left(\frac{1}{2C}\right)^q \frac{C^q q!}{q!} = \sum_{q \in \mathbb{N}_0} \left(\frac{1}{2}\right)^q = 2.$$

□

Quantitative Variante der Charakterisierung (2.13)? Es sei $\alpha > 0$ und X zentriert. Dann:

$$X \in \mathcal{G}\left(\frac{1}{8\alpha}\right) \Rightarrow E\left(e^{\alpha X^2}\right) \leq 2 \Rightarrow X \in \mathcal{G}\left(\frac{4}{\alpha}\right)$$

Beschränkte zentrierte ZVen sind sub-gaußsch:

Lemma 2.11 (Hoeffding-Lemma)**.**
Sei Y eine $[a,b]$-wertige (schreibe zukünftig $Y \in [a,b]$) zentrierte ZVe. Sei $\psi_Y(\lambda) = \ln(E(e^{\lambda Y}))$.

$$\Rightarrow \quad \psi_Y''(\lambda) \leq \frac{(b-a)^2}{4} \text{ und}$$

$$Y \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$$

Das Hoeffding-Lemma 2.11 wird beim Beweis der Hoeffding-Ungleichung 2.18 in 2.6 benötigt.

Beweis.

$$\left|Y - \frac{b+a}{2}\right| \leq \frac{b-a}{2} \quad \Rightarrow \quad \mathrm{Var}(Y) = \mathrm{Var}\left(Y - \frac{b+a}{2}\right)$$
$$\leq \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{4}.$$

Sei P_Y Verteilung von Y und $P_\lambda(\mathrm{dx}) = e^{-\psi_Y(\lambda)} e^{\lambda x} P_Y(\mathrm{dx})$ modifiziertes absolutstetiges Maß. Sei $Z \in \mathbb{R}$ mit Verteilung $P_Z = P_\lambda$.
Frage/Übung: Ist überhaupt P_λ ein W-Mass?
Da $Z \in [a,b]$ gilt ebenso:

$$\mathrm{Var}(Z) \leq \left(\frac{b-a}{2}\right)^2.$$

Direkte Rechnung ergibt:

$$\begin{aligned}
\psi_Y''(\lambda) &= \frac{\mathrm{d}^2}{\mathrm{d}^2\lambda} \ln(M_Y(\lambda)) \\
&= e^{-\psi_Y(\lambda)} E\left(Y^2 e^{\lambda Y}\right) - e^{-2\psi_Y(\lambda)} \left(E\left(Y e^{\lambda Y}\right)\right)^2 \\
&= e^{-\psi_y(\lambda)} \int_{\mathbb{R}} y^2 e^{\lambda y} \,\mathrm{d}P_Y(y) - \left(e^{-\psi_Y(\lambda)} \int_{\mathbb{R}} y e^{\lambda y} \,\mathrm{d}P_Y(y)\right)^2 \\
&= \int_{\mathbb{R}} y^2 \,\mathrm{d}P_\lambda(y) - \left(\int_{\mathbb{R}} y \,\mathrm{d}P_\lambda(y)\right)^2 \\
&= \mathrm{Var}(Z) \leq \frac{(b-a)^2}{4} \,\forall\, \lambda \in \mathbb{R}.
\end{aligned}$$

Da Y zentriert: $\psi_Y(0) = \psi_Y'(0) = 0$ und $\psi \in \mathcal{C}^2(0,\infty) \cap \mathcal{C}^1[0,\infty)$. Taylorentwicklung mit Lagrange-Restglied liefert:

$$\exists\, \theta \in [0,\lambda] : \psi_Y(\lambda) = \psi_Y(0) + \lambda \psi_Y'(0) + \frac{\lambda^2}{2} \psi_Y''(\theta)$$
$$\leq 0 + 0 + \frac{\lambda^2 (b-a)^2}{8} \quad \Rightarrow \quad Y \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$$

□

Beispiel: Ungleichung des Lemmas ist scharf: Führe sogenannte *Rademacher-Zufallsvariable* $X \colon \Omega \to \{-1,1\}$ ein mit $P(X=1) = P(X=-1) = \frac{1}{2}$. Es gilt $X \in [-1,1]$ mit $a = -1$ und $b = 1$. Wie in Lemma 2.11 zeigen wir:

$$\Rightarrow \quad \psi_X''(\lambda) \leq \frac{(b-a)^2}{4} = 1 \,\forall \lambda \geq 0$$

Nutzen nun die charakteristische Eigenschaft der KEF. $\lambda = 0$ einsetzen liefert: $\mathrm{Var}(X) = \psi_X''(0) \leq 1$.

Nun berechnen wir die Varianz exakt:

$$\mathrm{Var}(X) = E\left(X^2\right) = \frac{1}{2}(-1)^2 + \frac{1}{2}(-1)^2 = 1.$$

$\rightarrow$ Ungleichung lässt sich nicht mehr verbessern.

2.4. **Sub-Gamma-Zufallsvariablen.**

Einige wichtige Verteilungen haben Dichten, die schnell abfallen bei $\pm\infty$, jedoch nicht ganz so schnell wie $\mathcal{G}(\nu)$. Daher führen wir ein:

Definition 2.12. Sei $X \in \mathcal{L}^1$ mit $E(X) = 0$. X heißt *sub-Γ-verteilt von rechts* mit Varianzfaktor $\nu > 0$ und Skalenparameter $c \geq 0$, in Symbolen

$$X \in \Gamma_+(\nu, c) :\Leftrightarrow \begin{cases} \forall \lambda \in \left(0, \frac{1}{c}\right) \text{ gilt } \psi_X(\lambda) \leq \dfrac{\lambda^2 \nu}{2(1 - c\lambda)}, & \text{falls } c > 0 \\ \forall \lambda \in (0, \infty) \text{ gilt } \psi_X(\lambda) \leq \dfrac{\lambda^2 \nu}{2}, & \text{falls } c = 0. \end{cases}$$

Die Klasse $\Gamma_-(\nu, c)$ führen wir ein, indem wir setzen:

$$X \in \Gamma_-(\nu, c) \;\Leftrightarrow\; -X \in \Gamma_+(\nu, c),$$

d.h.: X ist sub-Γ von links mit Varianzfaktor ν und Skalenparameter $c :\Leftrightarrow -X \in \Gamma_+(\nu, c) \Leftrightarrow X \in \Gamma_-(\nu, c)$.
X heißt sub-Γ-verteilt mit Varianzfaktor ν und Skalenparameter $c \Leftrightarrow X \in \Gamma(\nu, c) := \Gamma_+(\nu, c) \cap \Gamma_-(\nu, c)$.

Das Konzept von sub-gamma findet im Kapitel 2.5 Verwendung, kann (dort) aber auch weggelassen werden. Es soll ein alternatives, flexibleres Kriterium zur sub-gauß-Eigenschaft bereitstellen.

Einige Bemerkungen:

(a) Insbesondere $\Gamma(\nu, 0) = \mathcal{G}(\nu)$

(b) Sei ZVe Y gammaverteilt mit Parameter $a, b \geq 0$, $X := Y - E(Y)$ zentrierte Version. Y hat die Dichte

$$\begin{aligned}
&\forall x \geq 0 : f(x) = \frac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a} \\
\Rightarrow\ & E(Y) = ab, \mathrm{Var}(Y) = ab^2 \text{ und} \\
& E\left(e^{\lambda X}\right) = \int_0^\infty e^{\lambda(y-ab)} f(y) dy = e^{-\lambda ab - a\ln(1-\lambda b)} \\
\Rightarrow\ & \forall \lambda \in \left(0, \frac{1}{c}\right) : \psi_X(\lambda) = a(-\lambda b - \ln(1-\lambda b)) \\
& \overset{\text{NR.}}{\leq} \frac{\lambda^2 \nu}{2(1-c\lambda)}, \text{ mit } \nu = ab^2, c = b
\end{aligned}$$

Nebenrechung: Für $x := \lambda b \in (0,1)$ gilt mit der Reihenentwicklung des Logarithmus:

$$\begin{aligned}
-\ln(1-x) - x = \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \ldots &= \frac{x^2}{2}\left(1 + \frac{2}{3}x + \frac{2}{4}x^2 + \ldots\right) \\
&\leq \frac{x^2}{2}\left(1 + x + x^2 + \ldots\right) \leq \frac{x^2}{2}\frac{1}{1-x}
\end{aligned}$$

Also sind Γ-ZVen sub-$\Gamma(ab^2, b)$. Allerdings ist X nicht symmetrisch verteilt. Die Verteilung von $-X$ fällt sogar schneller ab, da ja

$$\psi_{-X}(\lambda) = \ln E\left(e^{-\lambda X}\right) = \ln E\left(e^{(-\lambda)X}\right) = a\left(\lambda b - \ln(1+\lambda b)\right) \leq \frac{\lambda^2}{2} ab^2 = \frac{\lambda^2}{2}\nu$$

wegen

$$\begin{aligned}
y - \ln(1-y) &= y - y + \frac{y^2}{2} - \frac{y^3}{3} + \frac{y^4}{4} - \ldots \\
&= \frac{y^2}{2} - \left(\frac{y^3}{3} - \frac{y^4}{4}\right) - \left(\frac{y^5}{5} - \frac{y^6}{6}\right) - \ldots \leq \frac{y^2}{2} \quad \text{für } y \in (0,1).
\end{aligned}$$

Also: $X \in \Gamma_-(ab^2, 0) \subset \Gamma_-(ab^2, b)$ und damit $X \in \Gamma(ab^2, b)$.

Um das tail-Verhalten von sub-gamma ZV zu verstehen, untersuchen wir die Fenchel-Legendre-Duale von $\psi(\lambda) = \frac{\lambda^2\nu}{2(1-c\lambda)}$. Setze dazu $h_1(u) = 1 + u - \sqrt{1+2u}$ für $u > 0$. In der Übung wird gezeigt, dass

$$\psi^*(t) = \sup_{\lambda \in \left(0, \frac{1}{c}\right)} \left(t\lambda - \frac{\lambda^2\nu}{2(1-c\lambda)}\right) = \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right), \tag{2.14}$$

dass die Funktion $h_1 : (0,\infty) \to (0,\infty)$ strikt monoton wachsend, und bijektiv mit $h_1^{-1}(u) = u + \sqrt{2u}$ ist. Damit ergibt sich:

$$
\begin{aligned}
(\psi^*)^{-1}(u) &= \left(\mathcal{M}_{c/\nu}\right)^{-1} \circ h_1^{-1} \circ \left(\mathcal{M}_{\nu/c^2}\right)^{-1}(u) \\
&= \frac{\nu}{c} h_1^{-1}\left(\frac{c^2 u}{\nu}\right) = \frac{\nu}{c}\left(\frac{c^2 u}{\nu} + \sqrt{2c^2 u/\nu}\right) = \sqrt{2\nu u} + cu.
\end{aligned} \tag{2.15}
$$

Die Chernoff-Schranken lauten:

(i) $X \in \Gamma_+(\nu, c) \Rightarrow \forall t > 0$ gilt $P(X > t) \leq \exp\left(-\frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right)\right)$.

Mit der Substitution $s := \psi^*(t) = \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right)$ und der berechteten Formel für die Inverse erhalten wir die oftmals praktischere äquivalente Darstellung

$$\forall s > 0 \text{ gilt } P(X > \sqrt{2\nu s} + cs) \leq e^{-s}.$$

(ii) $X \in \Gamma(\nu, c) \Rightarrow \forall t > 0$:

$$\max\{P(X > \sqrt{2\nu t} + ct), P(-X > \sqrt{2\nu t} + ct)\} \leq e^{-t}.$$

Es gilt wieder (fast) eine Äquivalenz zu einer Momentenbedingung:

Satz 2.13.
Sei $X \in \mathcal{L}^1$ eine zentrierte Zufallsvariable.
(a) Gilt für ein $\nu > 0$ für jedes $t > 0$:

$$\max\left\{P(X > \sqrt{2\nu t} + ct), P(X < -(\sqrt{2\nu t} + ct))\right\} \leq e^{-t}, \tag{2.16}$$

so folgt für jedes $q \in \mathbb{N}$

$$E(X^{2q}) \leq q!(8\nu)^q + (2q)!(4C)^{2q}.$$

(b) Gilt umgekehrt für zwei Parameter $A, B \geq 0$ und jedes $q \in \mathbb{N}$

$$E(X^{2q}) \leq q! A^q + (2q)! B^{2q}, \tag{2.17}$$

so ist bereits $X \in \Gamma(4(A + B^2), 2B)$.

Man beachte die Analogien zu Kapitel 2.3, u.a. Satz 2.9.

2.5. **Eine Maximal-Ungleichung.**

In Teil (C) der Motivation wollten wir das Supremum einer Familie von ZVen abschätzen. Konkrete Schranke wird in diesem Kapitel für den Erwartungswert des Supremums hergeleitet.
Seien ZVen $Z_1, \ldots, Z_N \in \mathbb{R}$ mit $Z_i \in \mathcal{G}(\nu)$ für $i = 1, \ldots, N$ und $\nu > 0$.

Ziel: Schätze $E\left(\max_{i=1,\dots,N} Z_i\right)$ nach oben ab.
Idee: Nutze sub-gaußsche Eigenschaft der ZVen wie folgt aus.

$$\begin{aligned}\exp\left(\lambda E\left(\max_{i=1,\dots,N} Z_i\right)\right) &\overset{\text{Jensen}}{\leq} E\left(\exp\left(\lambda \max_{i=1,\dots,N} Z_i\right)\right)\\ &\leq E\left(\max_{i=1,\dots,N} \exp\left(\lambda Z_i\right)\right)\\ &\leq E\left(\sum_{i=1}^{N} \exp\left(\lambda Z_i\right)\right)\\ &= \sum_{i=1}^{N} M_{Z_i}(\lambda) \leq N \cdot \exp\left(\frac{\lambda^2 \nu}{2}\right).\end{aligned} \tag{2.18}$$

Durch Logarithmieren erhalten wir eine Abschätzung in gewünschter Form:

$$E\left(\max_{i=1,\dots,N} Z_i\right) \leq \frac{\ln(N) + \frac{\lambda^2\nu}{2}}{\lambda}.$$

Wähle nun

$$\lambda = \sqrt{2(\ln N)/\nu}. \tag{2.19}$$

Wir erhalten:

$$E\left(\max_{i=1,\dots,N} Z_i\right) \leq \frac{\sqrt{\ln(N)}}{\sqrt{2/\nu}} + \frac{\sqrt{\nu \ln(N)}}{\sqrt{2}} = \sqrt{2\nu \ln(N)}. \tag{2.20}$$

Fragen:

- Kann man λ geschickter wählen als in (2.19)?
- Ist die Schranke in (2.20) optimal?

Zur zweiten Frage: Sind $Z_1, \dots, Z_N \sim \mathcal{N}(0,1)$ unabhängig, folgt aus dem obigen:

$$\frac{E\left(\max_{i=1,\dots,N} Z_i\right)}{\sqrt{2\nu \ln(N)}} \leq 1.$$

Optimalität würde bedeuten, dass sogar:

$$\frac{E\left(\max_{i=1,\dots,N} Z_i\right)}{\sqrt{2\nu \ln(N)}} = 1.$$

Für dieses Beispiel gilt zumindest (Übung):

$$\lim_{N\to\infty} \frac{E\left(\max\limits_{i=1,\dots,N} Z_i\right)}{\sqrt{2\nu\ln(N)}} = 1.$$

2.5.1. *Ähnliche Resultate gelten auch für sub-Γ-Zufallsvariablen.*
Dazu holen wir etwas aus und entwickeln allgemeinere Resultate. Die folgenden Resultate dieses Kapitels sind anwendbar auf Klassen von Verteilungen, die geeignet dominiert werden. Die sub-gauß oder sub-Γ-ZVen sind lediglich Beispiele für solche Klassen, so dass auch ohne Kapitel 2.4 nachfolgende Ausführungen (bis auf Korollar 2.16) verstanden werden können.
Untersuchung von Eigenschaftern der abstrakten Legendre-Fenchel-Dualen.

Lemma 2.14.
Seien $b \in [0,\infty)$ *und* $\psi\colon [0,b) \to \mathbb{R}$ *konvex und* $\mathcal{C}^1([0,\infty))$ *mit* $\psi(0) = \psi'(0) = 0$. *Für* $t \geq 0$ *setzen wir:*

$$\psi^*(t) := \sup_{\lambda\in[0,b)} (\lambda t - \psi(\lambda)).$$

Dann ist ψ^* *auf* $[0,\infty)$ *nichtnegativ, monoton wachsend, konvex und unbeschränkt. Die* verallgemeinerte Inverse (Pseudo-Inverse):

$$(\psi^*)^{-1}(y) := \inf\{t \geq 0 \mid \psi^*(t) > y\}$$

erfüllt die Gleichung: $(\psi^*)^{-1}(y) = \inf\limits_{\lambda\in(0,b)} \left(\dfrac{y+\psi(\lambda)}{\lambda}\right)$.

Die Voraussetzungen sind für die KEF ψ_X einer zentrierten ZV X erfüllt.
Man beachte inhaltliche Analogien zum Kapitel 2.2 .

Beweis. $t \mapsto \lambda t - \psi(\lambda)$ ist affin-linear, konvex und isoton für $\lambda \geq 0$.
Übung $\Rightarrow$ ψ^* auch konvex und isoton als Supremum solcher Funktionen.

$$\psi^*(0) = \sup_{0<\lambda<b} \big(0 - \psi(\lambda)\big) = -\inf_{\lambda\in(0,b)} \psi(\lambda) = 0,$$

Denn:ψ konvex $\Rightarrow$ ψ' isoton, mit $\psi'(0) = 0$. Also ist ψ' nichtnegativ.
$\Rightarrow$ ψ isoton, mit $\psi(0) = 0$ und somit ist ψ nichtnegativ.
Damit ist auch ψ^* nichtnegativ.

Sei nun $\lambda_0 \in (0,b)$ fix. Da für jedes $t \geq 0$

$$\psi^*(t) \geq \sup_{\lambda\in[0,b)} (\lambda t - \psi(\lambda)) \geq \lambda_0 t - \psi(\lambda_0)$$

vorliegt, ist ψ^* unbeschränkt und $\{t \geq 0 \mid \psi^*(t) > y\} \neq \emptyset$, $\forall\, y \geq 0$.
Setze $u := \inf_{\lambda\in(0,b)} \left(\frac{y+\psi(\lambda)}{\lambda}\right)$, so gilt $\forall\, t \geq 0$:

$$\begin{aligned} u \geq t \quad &\Leftrightarrow \quad \forall\, \lambda \in (0,b) :\ \frac{y+\psi(\lambda)}{\lambda} \geq t \\ &\Leftrightarrow \quad \forall\, \lambda \in (0,b) :\ y \geq \lambda t - \psi(\lambda) \\ &\Leftrightarrow \quad y \geq \sup_{\lambda\in(0,b)} (\lambda t - \psi(\lambda)) = \psi^*(t) \end{aligned}$$

Komplementbildung liefert äquivalente Aussage dazu:

$$u < t \quad \Leftrightarrow \quad \psi^*(t) > y.$$

Also: $(u,\infty) = \{t \geq 0 \mid \psi^*(t) > y\}$
$\Rightarrow \quad u = \inf\big((u,\infty)\big) = \inf\{t \geq 0 \mid \psi^*(t) > y\} = \underbrace{(\psi^*)^{-1}(t).}_{\text{verallg. Inverse}}$ □

Vergleichen Sie die obige Vorgehensweise mit der aus der Stochastik bekannten Definition/Konstruktion der Quantil-Transformierten.
Anwendung von Lemma 2.14:

Satz 2.15 (Maximal-Ungleichung)**.**
Seien $Z_1, \ldots, Z_N \in \mathbb{R}$ ZVen, $b \in (0,\infty)$, $\psi \in C^1\big([0,b)\big)$ konvex mit $\psi(0) = \psi'(0) = 0$, so dass

$$\psi_{Z_i}(\lambda) = \ln(M_{Z_i}(\lambda)) \leq \psi(\lambda) \quad \text{für } \lambda \in [0,b) \text{ und } i = 1, \ldots, N.$$

Dann folgt

$$\text{für } i = 1, \ldots, N : Z_i \in \mathcal{L}^1 \text{ und } E\left(\max_{i=1,\ldots,N} Z_i\right) \leq (\psi^*)^{-1}(\ln(N)).$$

Es wird keine Unabhängigkeit verlangt!

Beweis. Mit Jensen-Ungleichung ist bekannt für $\lambda \in (0,b)$:

$$\begin{aligned} \exp\left(\lambda E\left(\max_{i=1,\ldots,N} Z_i\right)\right) &\leq E\left(\exp\left(\lambda \max_{i=1,\ldots,N} Z_i\right)\right) \\ &\leq E\left(\left(\max_{i=1,\ldots,N} e^{\lambda Z_i}\right)\right) \leq \sum_{i=1}^{N} E\left(e^{\lambda Z_i}\right) \overset{\text{Vorr.}}{\leq} N e^{\psi(\lambda)} \end{aligned}$$

$$\Rightarrow \ \forall\, \lambda \in (0,b): \ \lambda E\left(\max_{i=1,\dots,N} Z_i\right) \leq \psi(\lambda) + \ln(N)$$

$$\Rightarrow \ \forall\, \lambda \in (0,b): \ E\left(\max_{i=1,\dots,N} Z_i\right) \leq \frac{\psi(\lambda) + \ln(N)}{\lambda}$$

$$\Leftrightarrow \ E\left(\max_{i=1,\dots,N} Z_i\right) \leq \inf_{\lambda\in(0,b)} \frac{\psi(\lambda) + \ln(N)}{\lambda} \overset{2.14}{=} (\psi^*)^{-1}(\ln(N))$$

□

Korollar 2.16.
Seien $Z_1, \dots, Z_n \in \mathcal{L}^1$ zentrierte ZVen. Es gelten:

(a) *Falls $Z_i \in \mathcal{G}(\nu)$ für $i = 1, \dots, N$, so:*

$$E\left(\max_{i=1,\dots,N} Z_i\right) \leq \sqrt{2\nu \ln(N)}.$$

(b) *Falls $Z_i \in \Gamma_+(\nu, c)$ für $i = 1, \dots, N$, so:*

$$E\left(\max_{i=1,\dots,N} Z_i\right) \leq \sqrt{2\nu \ln(N)} + c \ln(N).$$

Beweis. zu (a): siehe Idee/Rechnung (2.18) am Anfang von Kapitel 2.5.
zu (b): Einsetzen von (2.15) in $(\psi^*)^{-1}(\ln(N))$ liefert die angegebene obere Schranke. □

Beispiel 2.17 (χ^2-Verteilung).
Sind $Y_1, \dots, Y_k \sim \mathcal{N}(0,1)$ unabhängige ZVen, so ist $X := \sum_{i=1}^k Y_i^2$ χ^2-verteilt mit k Freiheitsgraden. Die Dichte von X ist

$$f(x) = \frac{x^{k/2-1} e^{-x/2}}{\Gamma(k/2)\sqrt[k]{2}},$$

d.h. Dichte der Γ-Verteilung mit $a = k/2, b = 2$. Insbesondere $X - E(X) = X - k \in \Gamma_+(2k, 2) \cap \Gamma_-(2k, 0)$. Für $X_1, \dots, X_N \sim \chi^2$ mit k Freiheitsgraden impliziert Korollar 2.16

$$E\left(\max_{i=1,\dots,N} X_i - k\right) \leq 2\sqrt{k \ln(N)} + 2\ln(N) \square$$

2.6. **Hoeffding-Ungleichung.**
Nächstes Ziel: Schranken für Wahrscheinlichkeiten für große Werte von Summen von unabhängigen ZVen. Seien $X_1, \dots, X_n \in \mathbb{R}$ unabhängige $\mathcal{L}^1$-ZVen, so dass ein echtes Intervall $I \subset \mathbb{R}$ existiert mit $M_{X_i}(\lambda) = E(e^{\lambda X_i}) < \infty$ für $i = 1, \dots, n, \lambda \in I$. Für $S := \sum_{i=1}^n (X_i - E(X_i))$ gilt:

$$\forall\, \lambda \in I: \ \psi(\lambda) = \sum_{i=1}^n \ln(E(e^{\lambda(X_i - E(X_i))})).$$

Sind die X_i sogar beschränkt, genauer: $X_i \in [a_i, b_i]$ für $i = 1, \ldots, n$, so gilt nach Hoeffding-Lemma 2.11:

$$\begin{aligned} \psi''_{X_i - E(X_i)}(\lambda) \leq \frac{(b_i - a_i)^2}{4} \quad &\Rightarrow \quad \psi_{X_i - E(X_i)}(\lambda) \leq \frac{\lambda^2 (b_i - a_i)^2}{8} \\ (2.8) \Rightarrow \quad &\psi_S(\lambda) \leq \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2 \\ \Rightarrow \quad &S \in \mathcal{G}\left(\frac{1}{4} \sum_{i=1}^{n} (b_i - a_i)^2 \right) \end{aligned}$$

und somit

Satz 2.18 (Hoeffding-Ungleichung)**.**
Seien $X_1 \in [a_1, b_1], \ldots, X_n \in [a_n, b_n]$ unabhängige ZVen mit zentrierter Summe S. Dann gilt $\forall\, t \geq 0$:

$$P(S \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right)$$

Beweis. s.o. und Bemerkung 2.8. □

Bemerkung/Beispiel: Wähle $X_i = \alpha_i \varepsilon_i$, wobei $\varepsilon_1, \ldots, \varepsilon_n$ unabh. Rademacher-ZVen:

$$\overset{\text{Satz 2.18}}{\Rightarrow} \quad P(S \geq t) \leq \exp\left(-\frac{2t^2}{4 \sum_{i=1}^{n} \alpha_i^2} \right)$$

Da $\operatorname{Var}(S) = \operatorname{Var}(\alpha_1 \epsilon_1 + \ldots + \alpha_n \epsilon_n) = \sum_i \alpha_i^2 \operatorname{Var}(\epsilon_i) = \sum_i \alpha_i^2$ identifizieren wir

$$\exp\left(-\frac{2t^2}{4 \sum_{i=1}^{n} \alpha_i^2} \right) = \exp\left(-\frac{t^2}{2 \operatorname{Var}(S)} \right)$$

$$\text{Im Allgemeinen gilt aber: } \operatorname{Var}(S) < \frac{1}{4} \sum_{i=1}^{n} (b_i - a_i)^2.$$

In diesen Fällen gibt es feinere Ungleichungen: die *Bennett-* und *Bernsteinungleichung.*

Bemerkung 2.19 (Extremalitätseigenschjaft der Rademacher-ZVen). Wie lässt sich die Varianz beschränkter ZVen maximieren?
→ schiebe Werte so weit wie möglich nach außen, an die Ränder des Wertebereiches
→ Rademacher-ZVen sind gerade diejenigen ZVen, die auf einem Intervall $[-\alpha_i, \alpha_i]$ die Varianz maximieren (zum Wert $\sum_{i=1}^{n} \alpha_i^2$).

2.7. **Bennett-Ungleichung.**
Wie zuvor: $X_1, \dots, X_n$ unabhängige ZVen. $S := \sum_{i=1}^n (X_i - E(X_i))$.

$$\psi_S(\lambda) = \sum_{i=1}^n \Big(\ln(E(e^{\lambda X_i})) - \lambda E(X_i) \Big) \tag{2.21}$$

$$\leq \sum_{i=1}^n E(e^{\lambda X_i} - \lambda X_i - 1) \tag{2.22}$$

da ja $\ln u \leq u - 1$ für $u > 0$.

(2.21) und (2.22) sind Startpunkte für Bennett- bzw. Bernstein-Ungleichung.

Satz 2.20 (Bennett-Ungleichung)**.**
Sei $b > 0$. Seien $X_1, \dots, X_n \in \mathcal{L}^2$ unabhängige ZVen, so dass $X_i \leq b$ fast sicher für $i = 1, \dots, n$ (einseitige Beschränktheit).
Sei $\nu = \sum_{i=1}^n E(X_i^2)$. Dann gilt:

$$\psi_S(\lambda) \leq n \ln \Big(1 + \frac{\nu}{nb^2} \varphi(b\lambda) \Big) \leq \frac{\nu}{b^2} \varphi(b\lambda) \tag{2.23}$$

mit $\varphi(u) := e^u - u - 1$ für $u \in \mathbb{R}$. Ferner gilt:

$$P(S \geq t) \leq \exp\left(-\frac{\nu}{b^2} h\left(\frac{bt}{\nu} \right) \right),$$

wobei $h(u) := (1+u)\ln(1+u) - u$ für $u \geq 0$.

Die Bennett-Ungleichung wird mit der Bernstein-Ungleichung in Kapitel 2.8 verglichen. Ansonsten wird die Bennett-Ungleichung im weiteren Verlauf des Skripts nicht verwendet.

Beweis. O.E. sei $b = 1$ (skaliere sonst nach). Die Abbildung

$\mathbb{R} \setminus \{0\} \ni u \mapsto \varphi(u)/u^2$ ist isoton und stetig auf $\mathbb{R}$ fortsetzbar (Übung).

Sei $\lambda > 0$. Nach Voraussetzung ist $\lambda X_i \leq \lambda b = \lambda$. Also gilt $\forall i = 1, \dots, n$

$$e^{\lambda X_i} - \lambda X_i - 1 \leq X_i^2 (e^\lambda - \lambda - 1),$$
$$\Rightarrow \quad E(e^{\lambda X_i}) \leq 1 + \lambda E(X_i) + E(X_i^2)\varphi(\lambda).$$

Mit Aufsummieren und (2.21) folgt:

$$\psi_S(\lambda) \overset{(2.21)}{=} \sum_{i=1}^{n} \left(\ln(E(e^{\lambda X_i})) - \lambda E(X_i) \right)$$

$$\leq \frac{n}{n} \sum_{i=1}^{n} \left(\ln(1 + \lambda E(X_i) + E(X_i^2)\varphi(\lambda)) - \lambda E(X_i) \right)$$

(2.24)

$$\overset{konkav}{\leq} n \left(\ln \left(1 + \lambda \frac{\sum_{i=1}^{n} E(X_i)}{n} + \underbrace{\frac{\sum_{i=1}^{n} E(X_i^2)}{n}}_{=\nu/n} \varphi(\lambda) \right) - \lambda \frac{\sum_{i=1}^{n} E(X_i)}{n} \right)$$

$$\overset{\ln(x+1)\leq x}{\leq} \nu\varphi(\lambda).$$

Um die andere in (2.23) behauptete Schranke zu erhalten, schätzen wir (2.24) ab mit Hilfe von:

$$\text{Für } x \geq 1, a > 0 : 1 + \frac{a}{x} \leq 1 + a \leq e^a$$

$$\Rightarrow \text{ Für } a, b > 0 : \ \ln(1 + a + b) - a \leq \ln(1 + b).$$

In Kapitel 2.2 wurde bereits gesehen:

$$\nu\varphi(\lambda) \text{ ist KEF von } Y = X - E(X) \text{ für } X \sim Poi(\nu)$$

Aus der Ungleichung zwischen zwei KEFs ergibt sich eine zwischen den entsprechnden Cramer-Transformierten:

$$\psi_S^*(t) \geq \psi_Y^*(t) = \nu h\left(\frac{t}{\nu}\right) \overset{\text{Chernoff}}{\Rightarrow} P(S \geq t) \leq \exp\left(-\psi_S^*(t)\right) \leq \exp\left(-\nu h\left(\frac{t}{\nu}\right)\right)$$

□

Bemerkung 2.21. In einer Übungsaufgabe wird gezeigt

$$h(u) = (1+u)\ln(1+u) - u \geq \frac{u^2}{2\left(1 + \frac{u}{3}\right)}$$

$$\overset{\text{Bennett}}{\Rightarrow} \underbrace{P(S \geq t) \leq \exp\left(-\frac{t^2}{2\left(\nu + \frac{bt}{3}\right)}\right)}_{\text{Bernstein-Ungleichung}}$$

Das ist die *Bernstein-Ungleichung*. Für $\nu \gg tb$ sind Bennett und Bernstein im Wesentlichen äquivalent. Für $t \gg \frac{\nu}{b}$ verliert Bernstein gegenüber Bennett einen $\ln(t)$-Faktor im Exponenten.
→ Bernstein lässt sich aber allgemeiner beweisen (für schwächere Annahmen an ZVen)!

2.8. **Bernstein-Ungleichung.**

Satz 2.22 (Bernstein-Ungleichung)**.** *Seien $X_1, \dots, X_n$ unabhängige ZVen mit $\nu, c > 0$, so dass $\sum_{i=1}^n E(X_i^2) \le \nu$ und $\sum_{i=1}^n E((X_i)_+)^q \le \frac{q!}{2}\nu c^{q-2}$ für alle $q \in \mathbb{N}_{\ge 3}$. Dann gilt $\forall\, \lambda \in \left(0, \frac{1}{c}\right)$, $\forall\, t > 0$:*

$$\psi_S(\lambda) \le \frac{\nu\lambda^2}{2(1-c\lambda)}, \quad \psi_S^*(t) \ge \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right),$$

wobei $h_1(u) = 1 + u - \sqrt{1+2u}$, $u > 0$. Insbesondere gilt $\forall\, t > 0$:

$$P(S \ge \sqrt{2\nu t} + ct) \le e^{-t}.$$

Die Bernstein-Ungleichung wird in Kapitel 2.9 beim Johnson-Lindenstrauß-Lemma verwendet.

Beweis. Im Beweis der Bennett-Ungleichung wurde $\varphi(u) := e^u - u - 1$ definiert. Es gilt $\varphi(u) \le \frac{u^2}{2}$ für $u \le 0$ (Übung). Also folgt für $\lambda > 0, i = 1, \dots, n$:

$$\begin{aligned}\varphi(\lambda X_i) &\le \begin{cases} \sum_{q=3}^\infty \frac{\lambda^q (X_i)^q}{q!}, & \text{falls } X \ge 0 \\ \frac{\lambda^2 (X_i)^2}{2}, & \text{falls } X \le 0. \end{cases} \\ &\le \frac{\lambda^2 (X_i)^2}{2} + \sum_{q=3}^\infty \frac{\lambda^q (X_i)_+^q}{q!}\end{aligned}$$

$$\Rightarrow \quad E(\varphi(\lambda X_i)) \le \frac{\lambda^2 E(X_i^2)}{2} + \sum_{q=3}^\infty \frac{\lambda^q E((X_i)_+^q)}{q!}$$

$$\Rightarrow \quad \sum_{i=1}^n E(\varphi(\lambda X_i)) \le \frac{\nu}{2} \sum_{q=2}^\infty \lambda^q c^{q-2}. \quad \text{nach Vorauss.}$$

Letzte Summe ist endlich, falls $\lambda \in \left(0, \frac{1}{c}\right)$. Mit (2.22) ergibt sich

$$\psi_S(\lambda) \le \sum_{i=1}^n E(\varphi(\lambda X_i)) \le \frac{\nu}{2} \sum_{q=2}^\infty \lambda^q c^{q-2} = \frac{\nu\lambda^2}{2(1-c\lambda)}.$$

(Der letzte Schritt beleuchtet, warum der Bruch ein Bezugswert ist.) Damit folgt mit (2.15) insgesamt

$$\psi_S^*(t) \ge \sup_{\lambda \in \left(0, \frac{1}{c}\right)} \left(t\lambda - \frac{\lambda^2\nu}{2(1-c\lambda)}\right) = \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right)$$

Die Abschätzung an $P(S \ge t)$ folgt mit der Bemerkung über h_1^* vor Satz 2.13 in Kapitel 2.4, vgl. Übung. □

Korollar 2.23.
Seien $X_1, \dots, X_n$ unabhängige ZVen wie in Theorem 2.22. Dann gilt $\forall\, t > 0$:

$$P(S \geq t) \leq \exp\left(-\frac{t^2}{2(\nu + ct)}\right)$$

Beweis. Direkte Folgerung aus $h_1(u) \geq \frac{u^2}{2(1+u)}$ (Übungsaufgabe). □

Man kann zeigen, dass dies für $X_i \leq b$ die gleiche Schranke liefert, wie am Ende von Kapitel 2.7 aus Bennett hergeleitet wurde (Übungsaufgabe).

Beispiel 2.24 (Gaußsches Chaos der Ordnung Zwei).
Sei $X = (X_1, \dots, X_n)$ mit unabhängigen Komponenten $X_j \sim \mathcal{N}(0,1)$, also $X \sim \mathcal{N}(0, Id_n)$. Sei $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ symmetrisch mit $a_{jj} = 0$ für $j = 1, \dots, n$, insbesondere $\mathrm{Tr}(A) = \sum_{j=1}^n a_{jj} = 0$.
Definiere ZVe Z als quadratische Form: $Z = X^\top AX = \sum_{i,j=1}^n X_i a_{ij} X_j$.

$$\Rightarrow \quad E(Z) = \sum_{i,j=1, i\neq j}^{n} a_{ij} \underbrace{E(X_iX_j)}_{E(X_i)E(X_j)=0} + \sum_{i=1}^{n} \underbrace{a_{ii}}_{=0} E(X_iX_i) = 0.$$

Frage:
Wie stark schwankt die ZVe Z um den Mittelwert 0? ($\hat{=}$ Konzentration).

Da A symmetrisch, existiert $B \in \mathbb{R}^{n\times n}$ orthogonal mit $A = B^\top DB$, wobei

$$D = \begin{pmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \mu_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \mu_4 \end{pmatrix}$$ Diagonalmatrix mit Eigenwerten $\mu_1, \dots, \mu_n$ von A.

Sei $B = (b_{ij})_{i,j=1}^n$. Setze $Y_i = \sum_{j=1}^n b_{ij} X_j$ für $i = 1, \dots, n$.

$$\begin{aligned} \Rightarrow \quad \sum_{i=1}^n \mu_i Y_i^2 &= \sum_{i=1}^n \mu_i (\sum_{j=1}^n b_{ij} X_j)^2 = \sum_{i=1}^n \mu_i (BX)_i^2 \\ &= \langle X, B^\top DBX \rangle = X^\top AX = Z. \end{aligned}$$

Ist diese andere Darstellung von Z über Y_i besser?
Da $X \sim \mathcal{N}(0, Id_n)$ und B orthogonal, ist auch $Y = (Y_1, \dots, Y_n) \sim \mathcal{N}(0, Id_n)$ wegen Rotationsinvarianz.

$$\begin{aligned} &\Rightarrow \quad P_X = P_Y,\ P_{(X_1^2,\dots,X_n^2)} = P_{(Y_1^2,\dots,Y_n^2)}. \\ &\Rightarrow \quad P_Z = P_{\sum_{i=1}^n \mu_i Y_i^2} = P_{\sum_{i=1}^n \mu_i X_i^2}. \end{aligned}$$

Da $0 = \mathrm{Tr}(A) = \sum_{i=1}^n \mu_i$ gilt:

$$\sum_{i=1}^n \mu_i X_i^2 = \sum_{i=1}^n \mu_i (X_i^2 - 1).$$

Eigenschaft von X_i^2 : $E(X_i^2) = \mathrm{Var}(X_i) = 1 \quad \Rightarrow \quad X_i^2 - 1$ zentriert
$X_i^2 \sim \chi^2$-verteilt mit einem Freiheitsgrad, also insbesondere Γ-verteilt mit Parameter $a = \frac{1}{2}$ und $b = 2$. Nach Bsp. 2.17 in Kapitel 2.4 gilt:

$$\begin{aligned}\psi_{X_i^2-1}(\lambda) &= \frac{1}{2}\left[-\ln(1-2\lambda) - 2\lambda\right] \\ &\leq 2\frac{\lambda^2}{2(1-2\lambda)} = \frac{\lambda^2}{1-2\lambda}.\end{aligned}$$

Also folgt für die KEF ψ_Z von Z ähnlich wie (2.8) (in Kapitel 2.6) wegen Unabhängigkeit:

$$\begin{aligned}\psi_Z(\lambda) &= \psi_{\sum_{i=1}^n \mu_i (X_i^2-1)}(\lambda) = \sum_{i=1}^n \psi_{\mu_i(X_i^2-1)}(\lambda) \\ &= \sum_{i=1}^n \psi_{X_i^2-1}(\mu_i\lambda) = \frac{1}{2}\sum_{i=1}^n (-\ln(1-2\mu_i\lambda) - 2\mu_i\lambda) \\ &\leq \sum_{i=1}^n \frac{\mu_i^2\lambda^2}{1-2\mu_i\lambda}, \text{ sofern } \lambda \in (0, (2\max_{i=1,\ldots,n} \mu_i)^{-1}) =: J\end{aligned}$$

Nun gilt $\forall\, j = 1, \ldots, n$

$$\mu_i \leq |\mu_i| \leq \|A\| = \sup_{\|x\|=1} \|Ax\|.$$

Sei $\|A\|_{\mathrm{HS}} := \sqrt{\sum_{i=1}^n \mu_i^2}$ die *Hilbert-Schmidt-* oder *Frobenius-Norm* von A. Dann folgt für $\lambda \in J$:

$$\psi_Z(\lambda) \leq \sum_{i=1}^n \frac{\mu_i^2\lambda^2}{1-2\lambda\|A\|} = \frac{\lambda^2\|A\|_{\mathrm{HS}}^2}{1-2\lambda\|A\|}.$$

In der Notation der sub-Γ-Verteilungen ist $\nu = 2\|A\|_{\mathrm{HS}}^2$, $c = 2\|A\|$ und nach Kapitel 2.4 gilt:

$P(Z > 2\|A\|_{\mathrm{HS}}\sqrt{t} + 2\|A\|t) \leq e^{-t}$ (vgl. Satz 2.13), oder

$P(Z > t) < \exp\left(-\frac{\|A\|_{\mathrm{HS}}^2}{2\|A\|} h_1\left(\frac{\|A\|t}{\|A\|_{\mathrm{HS}}^2}\right)\right)$ oder

$P(Z > t) \leq \exp\left(-\frac{t^2}{4(\|A\|_{\mathrm{HS}}^2 + t\|A\|)}\right)$ Übung wie in Kor. 2.23,

wobei h_1 die schon bekannte Entropie-Funktion ist.

2.9. **Johnson-Lindenstrauss-Lemma.**
Folgende Ausführungen werden in einem viel allgemeineren Rahmen in Kapitel 5.6 ausgeführt.

Definition 2.25. Sei $\mathcal{H}$ ein separabler Hilbertraum (z.B. $\mathcal{H} = \mathbb{R}^D$ mit Dimension $D \in \mathbb{N}$). Für gegebenes $\varepsilon \in (0,1)$ und $A \subset \mathcal{H}$ heißt $f\colon \mathcal{H} \to \mathbb{R}^d$ *ε-Isometrie auf A*, falls für alle $a, b \in A$ gilt:

$$(1-\varepsilon)\|a-b\|_{\mathcal{H}} \leq \|f(a)-f(b)\|_{\mathbb{R}^d} \leq (1+\varepsilon)\|a-b\|_{\mathcal{H}} \tag{2.25}$$

Falls $\mathcal{H}$ hochdimensional, A große Kardinalität hat und $\varepsilon \in (0,1)$ und $d \in \mathbb{N}$ klein sind, ist völlig unklar, ob ein solches f existiert.

Das Johnson-Lindenstrauss-Lemma besagt, dass es eine universelle Konstante κ gibt, so dass für jedes A mit Kardinalität $\operatorname{card}(A) = n < \infty$ und

$$d \geq \frac{\kappa}{\varepsilon^2} \ln(n)$$

tatsächlich ein $f\colon \mathcal{H} \to \mathbb{R}^d$ mit Eigenschaft (2.25) existiert.

Allerdings kann man dieses f nicht explizit angeben, da es mit einem *Zufallsmechanismus* konstruiert wird. (Vergleiche die *probabilistic method* von Paul Erdős in der Kombinatorik und Graphentheorie.) Dafür ist es möglich f, linear zu wählen. Wir zeigen sogar, dass aus einem vorgegebenen Ensemble linearer Funktionen $f\colon \mathcal{H} \to \mathbb{R}^d$, die meisten (2.25) erfüllen.
Einfachheitshalber nehmen wir $\mathcal{H} = \mathbb{R}^D$ und typischerweise $D \gg d$ an.
Ansatz: Sei $W : \mathbb{R}^D \to \mathbb{R}^d$ linear und zufällig gewählt mit der Eigenschaft

$$\forall\, \alpha \in \mathbb{R}^D : \; E\left(\|W\alpha\|^2\right) = \|\alpha\|^2,$$

d.h. im Mittel haben wir eine exakte Isometrie. Gewünscht ist die Eigenschaft, dass die Zufallsvariable $\|W\alpha\|^2$ wenig streut.

Konstruktion von W
Seien $X_{ij}, i = 1,\dots,d$ und $j = 1,\dots,D$ unabhängig identisch verteilte Zufallsvariablen. Ferner seien $E(X_{ij}) = 0$ und $\operatorname{Var}(X_{ij}) = 1$. Für $\alpha = (\alpha_1,\dots,\alpha_D) \in \mathbb{R}^D$ und $i = 1,\dots,d$ setze

$$\tilde{W}_i(\alpha) = \sum_{j=1}^{D} \alpha_j X_{ij} \text{ und } W_\alpha = \left(\frac{1}{\sqrt{d}}\tilde{W}_i(\alpha)\right)_{i=1}^{d}. \tag{2.26}$$

Wegen der Unabhängigkeit gilt für für jedes $i = 1, \ldots, d$:

$$E\left(\tilde{W}_i(\alpha)^2\right) = E\left(\left(\sum_{j=1}^{D} \alpha_j X_{ij}\right)^2\right) = \sum_{i,j=1}^{D} \alpha_i \alpha_j E(X_{ij}^2)\delta_{ij},$$

wobei δ_{ij} das Kronecker-Delta ist. Daraus folgt für alle $\alpha \in \mathbb{R}^D$

$$E\left(\|W_\alpha\|^2\right) = E\left(\frac{1}{d}\sum_{i=1}^{d}(\tilde{W}_i(\alpha))^2\right) = \frac{1}{d}\sum_{i=1}^{d}\|\alpha\|^2 = \|\alpha\|^2.$$

Wir wollen nun solche Zufallsvariablen X_{ij} wählen, die sich typischerweise so verhalten, wie ihr Mittelwert. Das ist eine typische Konzentrationsbedingung, wie sie z.B. für sub-gaußsche Zufallsvariablen erfüllt ist.

Satz 2.26 (Johnson-Lindenstrauss-Lemma)**.**
Seien $A \subset \mathbb{R}^D$ mit card(A) $= n \in \mathbb{N}$, sowie $\varepsilon, \delta \in (0,1)$. Für $\nu \geq 1$ seien $X_{ij} \in \mathcal{G}(\nu), i = 1, \ldots, d$ und $j = 1, \ldots, D$ unabhängig identisch verteilt und W wie in (2.26)*.*

$$\begin{aligned}&\textit{Sobald } d \geq 100 \cdot \frac{\nu^2}{\varepsilon^2} \ln\left(\frac{n}{\sqrt{\delta}}\right) \textit{ gilt:}\\ &P(W \textit{ ist eine } \varepsilon\textit{-Isometrie auf } A) \geq 1 - 2\delta.\end{aligned}$$

Beweis. Sei $S \subset \mathbb{R}^d$ die Einheitssphäre und $T \subset S$ definiert durch

$$T = \left\{ \frac{a-b}{\|a-b\|} \mid a, b \in A, a \neq b \right\}$$

Wir wollen beweisen, dass mit hoher Wahrscheinlichkeit gilt:

$$\max_{\alpha \in T} \left| \|W\alpha\|^2 - 1 \right| \leq \varepsilon.$$

Denn dann folgt wegen Linearität $\forall\, a, b \in A$:

$$\begin{aligned}&\left| \|Wa - Wb\|^2 - \|a-b\|^2 \right| \leq \varepsilon \|a-b\|^2\\ \Rightarrow\ &(1-\varepsilon)\|a-b\|^2 \leq \|Wa - Wb\|^2 \leq (1+\varepsilon)\|a-b\|^2.\end{aligned}$$

Für jedes $\alpha \in S$ und $i \leq d$ gilt:

$$\begin{aligned} E(\exp(\lambda \tilde{W}_i(\alpha))) &= E(\exp(\lambda \sum_{j=1}^{D} \alpha_j X_{ij})) \\ &= \prod_{j=1}^{D} E(\exp(\lambda \alpha_j X_{ij})) \leq \prod_{j=1}^{D} \exp\left(\frac{\lambda^2 \alpha_j^2 \nu}{2}\right) \\ &= \exp\left(\frac{\lambda^2}{2} \nu \sum_{j=1}^{D} \alpha_j^2\right) = \exp\left(\frac{\lambda^2 \nu \|\alpha\|^2}{2}\right) = \exp\left(\frac{\lambda^2 \nu}{2}\right), \end{aligned}$$

also $\tilde{W}_i(\alpha) \in \mathcal{G}(\nu)$.
Insbesondere impliziert Satz 2.9 $\forall\, k \in \mathbb{N}_{\geq 2} :\ E(\tilde{W}_i(\alpha)^{2k}) \leq \frac{k!}{2}(4\nu)^k$.
Weiterhin sind alle Komponenten $\tilde{W}_1(\alpha), \ldots, \tilde{W}_d(\alpha)$ unabhängig $\forall\, \alpha \in S$.
Damit können wir die Bernstein-Ungleichung 2.22 anwenden:

$$\Rightarrow \quad \forall\, \alpha \in T, t > 0 : P\left(\left|\sum_{i=1}^{d} (\tilde{W}_i(\alpha)^2 - 1)\right| \geq 4\nu\sqrt{2dt} + 4\nu t\right) \leq 2e^{-t}.$$

$$\begin{aligned} \Rightarrow \quad & P\left(\max_{\alpha \in T} \left|\sum_{i=1}^{d} (\tilde{W}_i(\alpha)^2 - 1)\right| \geq 4\nu\sqrt{2dt} + 4\nu t\right) \\ &= P\left(\bigcup_{\alpha \in T} \left\{\left|\sum_{i=1}^{d} (\tilde{W}_i(\alpha)^2 - 1)\right| \geq 4\nu\sqrt{2dt} + 4\nu t\right\}\right) \\ &\leq \sum_{\alpha \in T} P\left(\left|\sum_{i=1}^{d} (\tilde{W}_i(\alpha)^2 - 1)\right| \geq 4\nu\sqrt{2dt} + 4\nu t\right) \\ &\leq 2e^{-t}|T| \leq 2e^{-t}n^2 \end{aligned}$$

Wähle $t = \ln\left(\frac{n^2}{\delta}\right) = 2\ln\left(\frac{n}{\sqrt{\delta}}\right)$, dann:

$$P\left(\max_{\alpha \in T} \left|\sum_{i=1}^{d} (\tilde{W}_i(\alpha)^2 - 1)\right| \geq 8\nu\sqrt{d\ln\left(\frac{n}{\sqrt{\delta}}\right)} + 8\nu\ln\left(\frac{n}{\sqrt{\delta}}\right)\right) \leq 2\delta$$

$$\Leftrightarrow \quad P\left(\max_{\alpha \in T} \left|\|W\alpha\|^2 - 1\right| \geq 8\nu\left(\sqrt{\frac{\ln(n/\sqrt{\delta})}{d}} + \frac{\ln(n/\delta)}{d}\right)\right) \leq 2\delta$$

Falls $d \geq 100\frac{\nu^2}{\varepsilon^2}\ln\left(\frac{n}{\sqrt{\delta}}\right)$ gewählt wird, folgt

$$8\nu\left[\sqrt{\frac{\ln(n/\sqrt{\delta})}{d}}+\frac{\ln(n/\sqrt{\delta})}{d}\right]\leq\frac{4\varepsilon}{5}+\frac{2\varepsilon^2}{25\nu}\leq\frac{20\varepsilon}{25}+\frac{2\varepsilon}{25}\leq\varepsilon,$$

da nach Annahme $\nu \geq 1$. Also haben wir tatsächlich gezeigt:

$$P\left(\sup_{\alpha\in T}\left|\|W\alpha\|^2-1\right|\leq\varepsilon\right)\geq 1-2\delta$$

□

Übung: Welche Konstanten verbessern sich, falls $X_{ij} \sim \mathcal{N}(0,1)$?

2.10. **Assoziations- und Korrelationsungleichungen.**

Dieser Abschnitt wird bei der Janson-Ungleichung und der Perkolationstheorie (optional) angewendet. Für das sonstige weitere Verständnis kann dieses Kapitel übersprungen werden.

Sind X, Y unabh. ZV, gelten praktische Rechenregeln für Produkte von X und Y oder von Funktionen davon. Unter welchen Annahmen kann man etwas von diesen Rechenregeln ‚retten', falls X und Y nicht unabh. sind?

Satz 2.27 (Čebyšev-Assoziationsungleichung)**.**
Seien $f, g\colon \mathbb{R} \to \mathbb{R}$ isotone Funktionen, ZVen $X \in \mathbb{R}$ und $Y \in \mathbb{R}_+$, so dass $f, g \in \mathcal{L}^2(P_X)$. Dann gilt

$$E(Y)E(Yf(X)g(X)) \geq E(Yf(X))E(Yg(X))$$

Bemerkung: Sei $Y \equiv 1$. Dann impliziert Satz 2.27:

$$E(f(X)g(X)) \geq E(f(X))E(g(X)) \;\Leftrightarrow\; E(f(X)g(X)) - E(f(X))E(g(X)) \geq 0$$
$$\overset{f,g\in\mathcal{L}^2}{\Leftrightarrow}\; \mathrm{Cov}(f(X), g(X)) \geq 0.$$

Ist f antiton und g weiterhin isoton, gilt

$$E(Y)E(Yf(X)g(X)) \leq E(Yf(X))E(Yg(X)).$$

Beweis. Sei der Vektor $(X', Y') : \Omega \to \mathbb{R} \times \mathbb{R}_+$ unabhängige Kopie von (X, Y) mit $P_{(X,Y)} = P_{(X',Y')}$. Sind f, g monoton wachsend, so gilt:

$$\begin{aligned}
& (f(X) - f(X'))(g(X) - g(X')) \geq 0 \\
\Rightarrow\ & YY'(f(X) - f(X'))(g(X) - g(X')) \geq 0 \\
\Rightarrow\ & E(YY'(f(X) - f(X'))(g(X) - g(X'))) \geq 0 \\
\overset{\text{Linearität}}{\Leftrightarrow}\ & E(YY'f(X)g(X) + YY'f(X')g(X')) \\
& \quad \geq E(YY'f(X)g(X') + YY'f(X')g(X)) \\
\overset{\text{unabh.}}{\Leftrightarrow}\ & E(Y')E(Yf(X)g(X)) + E(Y)E(Y'f(X')g(X')) \\
& \quad \geq E(Yf(X))E(Y'g(X')) + E(Yg(X))E(Y'f(X')) \\
\overset{\text{id. vert.}}{\Leftrightarrow}\ & 2E(Y)E(Yf(X)g(X)) \geq 2E(Yf(X))E(Yg(X)).
\end{aligned}$$

□

Der Beweis illustriert die Methode der *unabhängigen Kopie*. Führe dazu unabhängig, identisch verteilte ZVen ein und im zweiten Schritt durch Erwartungswerte auf die ursprünglichen ZVen zurück.
Es gibt auch eine multivariate Variante dieser Ungleichung. Dazu:

Definition 2.28. $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt *isoton* (oder *n-isoton*) :⇔

$\forall\, i \in \{1, \ldots, n\}$ und $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) \in \mathbb{R}^{n-1}$ ist

$\mathbb{R} \ni x \mapsto f(y_1, \ldots, y_{i-1}, x, y_{i+1}, \ldots, y_n)$ monoton wachsend bzw. isoton,

d.h. f isoton in jeder Koordinate.
f heißt *antiton* (oder *n-antiton*), falls $-f$ isoton ist.

Satz 2.29 (Harris-Ungleichung)**.**
Seien $f, g : \mathbb{R}^n \to \mathbb{R}$ n-isotone Funktionen und unabhängige ZVen $X_1, \ldots, X_n \in \mathbb{R}$. Setze $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$. Dann gilt

$$E(f(X)g(X)) \geq E(f(X))E(g(X)). \tag{2.27}$$

Bemerkung 2.30 (FKG-Ungleichung). Die Aussage (2.27) gilt auch im Fall $n = \infty$, d.h. falls $X_k, k \in \mathbb{N}$, eine Folge unabhängiger ZV, $f, g\colon \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ isoton in jedem Argument sind und $f(X), g(X)$ endliche Varianz besitzen. Diese Aussage wird mit Hilfe eines Martingal-Konverngezsatzes bewiesen.

Bemerkung 2.31. Da die X_i unabhängig sind, integrieren wir bezüglich eines Produktmaßes. Dennoch ist es in solchen Situationen sinnvoll, bedingte Erwartungen zu benutzen, um Schreibarbeit zu sparen.

Illustration für $n = 3$: Sei $f(X_1, X_2, X_3) = f(X)$. Nach dem Trafo-Satz gilt:

$$\begin{aligned} E(f(X) \mid X_1) &= \int_{\mathbb{R}} \int_{\mathbb{R}} f(X_1, t, s) \mathrm{d}P_{X_2}(t) \mathrm{d}P_{X_3}(s), \\ E(f(X) \mid X_1, X_2) &= \int_{\mathbb{R}} f(X_1, X_2, s) \mathrm{d}P_{X_3}(s) \textit{und} \\ E(E(f(X) \mid X_1, X_2)) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(r, t, s) \mathrm{d}P_{X_3}(s) \right) \mathrm{d}P_{X_2}(t) \mathrm{d}P_{X_1}(r) \\ &= E(f(X)), \end{aligned}$$

was sich auch aus der Turmeigenschaft für bedingte Erwartungen herleiten lässt.

Beweis. Zeige (2.27) in Satz 2.29 durch vollständige Induktion.
Induktionsanker: Fall $n = 1$ folgt aus Bemerkung zu Satz 2.27.
Induktionsschritt: Nehme an, dass (2.27) bekannt ist für alle $n < k$. Für das 1-dim. Integral bezüglich P_{X_k} gilt wieder wegen Satz 2.27:

$$E(f(X)g(X) \mid X_1, \ldots, X_{k-1}) \geq E(f(X) \mid X_1, \ldots, X_{k-1}) E(g(X) \mid X_1, \ldots, X_{k-1}),$$

denn bei eingefrorenen $X_1, \ldots, X_{k-1}$ sind f und g isotone Funktionen im $k$ten Argument. Turmeigenschaft und Monotonie des Integrals ergeben:

$$\begin{aligned} E(f(X)g(X)) &= E(E(f(X)g(X) \mid X_1, \ldots, X_{k-1})) \\ &\geq E\left[E(f(X) \mid X_1, \ldots, X_{k-1}) E(g(X) \mid X_1, \ldots, X_{k-1})\right] \end{aligned}$$

Nun ist

$$\begin{aligned} f_{k-1} \colon (X_1, \ldots, X_{k-1}) &\mapsto E(f(X) \mid X_1, \ldots, X_{k-1}) \\ &\overset{\text{unabh.}}{=} \int_{\mathbb{R}} f(X_1, \ldots, X_{k-1}, t) \mathrm{d}P_{X_k}(t) \quad (k-1)\text{-isoton}, \end{aligned}$$

ebenso wie

$$(X_1, \ldots, X_k) \mapsto E(g(X) \mid X_1, \ldots, X_{k-1}).$$

Wende Induktionsvorraussetzung (IV) an und erhalte:

$$E\left[\underbrace{E(f(X) \mid X_1, \ldots, X_{k-1})}\, E(g(X) \mid X_1, \ldots, X_{k-1})\right]$$

$$\begin{aligned} &\overset{\text{(IV)}}{\geq} E[E(f(X) \mid X_1, \ldots, X_{k-1})] E[E(g(X) \mid X_1, \ldots, X_{k-1})] \\ &\overset{\text{Turmeig.}}{=} E(f(X)) E(g(X)). \end{aligned}$$

□

Bemerkung 2.32 (Frage/Überlegung:). Funktionieren solche Aussage auch für andere Klassen von Funktionen? Z.B.:

- sphärisch-symmetrische,
- radial-abfallende

Funktionen? Betrachte dazu folgenden Satz.

Satz 2.33 (Gaußsche-Korrelationsungleichung)**.**
Sei $P = \mathcal{N}(0, C)$, mit C positiv definit, ein zentriertes Gaußmaß auf $\mathbb{R}^d$ und A, B zwei (bzgl. Spiegelung am Ursprung) symmetrische[2]*, konvexe Mengen.*

$$\Rightarrow \; P(A \cap B) \geq P(A)P(B).$$

Falls $P(A) > 0$, gilt das besser interpretierbare

$$P(B \mid A) \geq P(B)$$

Diesen Satz gibt es als Vermutung seit den 60ern motiviert durch zwei Arbeiten in 1955 und 1959. Beweis und Veröffentlichung von Thomas Royen im Jahr 2014. Seit 2017 gibt es weiteren, strukturierteren Beweis von Latala et. al..

2.11. **Anwendung der Harris-Ungleichung: Janson-Ungleichung.** Wir betrachten folgende kombinatorische Situation, die man in der Theoretischen Informatik, der probabilistischen Methode in der Kombinatorik und dem Studium von zufälligen Graphen antrifft.
Seien $n \in \mathbb{N}$, $[n] := \{1, \ldots, n\}$, $X_1, \ldots, X_n \in \{0, 1\}$ unabh. ZV mit

$$p_k := P(X_k = 1) = 1 - P(X_k = 0) \text{ für } k \in [n] \text{ und } p_1, \ldots, p_n \in [0, 1]$$

Für $A \subset [n]$ setzte

$$Y_A = \prod_{i \in A} X_i$$

und $\mathcal{I} \subset \mathcal{P}([n])$

$$Z := \sum_{A \in \mathcal{I}} Y_A$$

was ein Polynom in den 0/1-Variablen $X_1, \ldots, X_n$ ist.

[2] *d.h.* $x \in A \Rightarrow -x \in A$

Bemerkung 2.34 (Übung). Seien $f, g\colon \mathbb{R}^n \to \mathbb{R}$ zwei n-antitone Funktionen. Leiten Sie aus der Harris-Ungleichung

$$E(f(X)g(X)|Y_A = 1) \geq E(f(X)|Y_A = 1)E(g(X)|Y_A = 1)$$

her. Hierbei ist $X = (X_1, \ldots, X_n)$.

Offensichtlich gilt für $A, B \in \mathcal{I}$ mit $A \cap B = \emptyset$

$$E(Y_A Y_B) = E(\prod_{i\in A} X_i \prod_{k\in B} X_k) = \prod_{i\in A}\prod_{k\in B} E(X_i)E(X_k) = E(Y_A)E(Y_B)$$

Daher reduziert sich die

$$\mathrm{Var}(Z) = E(Z^2) - E(Z)^2 = \sum_{A,B\in\mathcal{I}} E(Y_A Y_B) - \sum_{A,B\in\mathcal{I}} E(Y_A)E(Y_B)$$

(2.28)

$$= \sum_{A,B\in\mathcal{I}, A\cap B\neq\emptyset} [E(Y_A Y_B) - E(Y_A)E(Y_B)] \leq \sum_{A,B\in\mathcal{I}, A\cap B\neq\emptyset} E(Y_A Y_B) \quad =: \Delta$$

Čebyšev liefert die symmetrische Schranke

$$P(|Z - EZ| > t) \leq \frac{\Delta}{t}$$

Auch wenn die Summanden von Z nicht unabhängig sind, gibt es eine (zumindest einseitige) exponentielle *tail*-Schranke.

$$Y_A = \prod_{i\in A} X_i, \quad Z = \sum_{A\in\mathcal{I}} Y_A, \quad \Delta = \sum_{A,B\in\mathcal{I}, A\cap B\neq\emptyset} E(Y_A Y_B)$$

Satz 2.35. *Seien $\mathcal{I} \subset \mathcal{P}([n])$ sowie Z und Δ wie soeben definiert. Dann gilt für alle $\lambda \leq 0$*

$$\psi_{Z-EZ}(\lambda) \leq \varphi\left(\frac{\lambda\Delta}{EZ}\right)\frac{(EZ)^2}{\Delta}, \qquad \text{mit } \varphi(x) = e^x - x - 1$$

Insbesondere gilt für $t \in [0, EZ]$

$$P(Z - EZ < -t) \leq \exp\left(\frac{-t^2}{2\Delta}\right)$$

Beweis. Für die KEF von $Z - EZ$ gilt

$$\psi'(\lambda) = \frac{E(Ze^{\lambda Z})}{E(e^{\lambda Z})} - E(Z) = \sum_{A\in\mathcal{I}} \frac{E(Y_A e^{\lambda Z})}{E(e^{\lambda Z})} - E(Z)$$

Wie wollen jeden A-Summanden auf der rechten Seite einzeln abschätzen. Zu jedem $A \in \mathcal{I}$ setze

$$U_A = \sum_{B \in \mathcal{I}, A \cap B \neq \emptyset} Y_B$$
$$Z_A = \sum_{B \in \mathcal{I}, A \cap B = \emptyset} Y_B,$$

so dass $Z = U_A + Z_A \geq Z_A$ für jedes $A \in \mathcal{I}$. Wegen der Fallunterscheidungsformel

$$E(Y_A e^{\lambda Z}) = E(Y_A e^{\lambda Z} | Y_A = 1) P(Y_A = 1) + 0 = E(e^{\lambda Z} | Y_A = 1) E(Y_A)$$

reicht es, die bedingte Erwartung von oben abzuschätzen. Da λ negativ ist, sind $X \mapsto e^{\lambda U_A}$ und $X \mapsto e^{\lambda Z_A}$ antitone Funktionen. Es folgt

$$\begin{aligned} E(e^{\lambda Z} | Y_A = 1) &= E(e^{\lambda U_A} e^{\lambda Z_A} | Y_A = 1) \\ \text{(Harris)} \quad &\geq E(e^{\lambda U_A} | Y_A = 1) E(e^{\lambda Z_A} | Y_A = 1) \\ (Z_A, Y_A \text{unabhängig}) \quad &= E(e^{\lambda U_A} | Y_A = 1) E(e^{\lambda Z_A}) \\ (Z \geq Z_A) \quad &\geq E(e^{\lambda U_A} | Y_A = 1) E(e^{\lambda Z}) \\ \text{(Jensen)} \quad &\geq e^{\lambda E(U_A | Y_A = 1)} E(e^{\lambda Z}) \end{aligned} \tag{2.29}$$

Damit erhalten wir

$$\begin{aligned} \frac{E(Z e^{\lambda Z})}{E(Z)} &= \sum_{A \in \mathcal{I}} \frac{E(Y_A e^{\lambda Z})}{E(Z)} = \sum_{A \in \mathcal{I}} \frac{E(e^{\lambda Z} | Y_A = 1) E(Y_A)}{E(Z)} \\ &\geq E(e^{\lambda Z}) \underbrace{\sum_{A \in \mathcal{I}} \frac{E(Y_A)}{E(Z)}} e^{\lambda E(U_A | Y_A = 1)} \end{aligned}$$

$$\text{(Jensen)} \geq E(e^{\lambda Z}) \exp\left[\lambda \sum_{A \in \mathcal{I}} \frac{E(Y_A)}{E(Z)} E(U_A | Y_A = 1)\right]$$

Nun wollen wir die bedingte Erwartung loswerden.

$$\begin{aligned} \Delta &= \sum_{A,B \in \mathcal{I}, A \cap B \neq \emptyset} E(Y_A Y_B) = \sum_{A \in \mathcal{I}} E(Y_A U_A) \\ &= \sum_{A \in \mathcal{I}} E(Y_A U_A | Y_A = 1) E(Y_A) + E(Y_A U_A | Y_A = 0) P(Y_A = 0) \\ &= \sum_{A \in \mathcal{I}} E(U_A | Y_A = 1) E(Y_A) \end{aligned}$$

Umstellen ergibt

$$\frac{E(Ze^{\lambda Z})}{E(e^{\lambda Z})} \geq E(Z) \exp\left[\lambda \frac{\Delta}{E(Z)}\right]$$

Damit folgt für die Ableitung der KEF

$$\frac{E(Ze^{\lambda Z})}{E(e^{\lambda Z})} - E(Z) \geq E(Z)\left[\exp\left(\lambda \frac{\Delta}{E(Z)}\right) - 1\right]$$

und wegen $\psi(0) = 0$ für die KEF selbst :

$$\begin{aligned}
\psi(\lambda) &= \psi(0) - \int_\lambda^0 \psi'(t)\mathrm{d}t \leq -E(Z)\int_\lambda^0 (e^{t\Delta/E(Z)} - 1)\mathrm{d}t \\
&= -E(Z)\int_{\lambda\Delta/E(Z)}^0 (e^s - 1)\frac{E(Z)}{\Delta}\mathrm{d}s = -\frac{(EZ)^2}{\Delta}\left[e^s - s\right]_{\lambda\Delta/E(Z)}^0 \\
&= \frac{(EZ)^2}{\Delta}\varphi\left(\frac{\lambda\Delta}{E(Z)}\right) \\
&= \frac{\lambda^2}{2}\Delta
\end{aligned}$$

Somit folgt auch $P(Z - EZ \leq -t) \leq e^{-t^2/(2\nu)}$ mit $\nu = \Delta$. □

Bemerkung 2.36. In Anwendungen will man oft zeigen, dass es Vektoren $X = (X_1, \ldots, X_n)$ gibt, die gewisse Klauseln erfüllen. Ja, man will sogar zeigen, dass dies für ein zufälig ausgewähltes X mit hoher W'keit zutrifft. Hier gilt:

- $Y_A = 1 \Leftrightarrow A$-te Bedigung ist erfüllt und
- $Z = \sum_{A\in\mathcal{I}} Y_A > 0 \Leftrightarrow$ mindestens eine der Klauseln in $\mathcal{I}$ ist erüllt.

Wir schätzen die W'keit des Komplements $\{Z = 0\}$ von $\{Z > 0\}$ ab (da ja Z nichtnegativ). Janson liefert:

$$P(Z = 0) = P(Z \leq EZ - EZ) \leq \exp\left(-\frac{(EZ)^2}{\Delta}\right).$$

Ob diese Abschätzung gut ist, ergibt sich aus der Abhängigkeit der Größen $(EZ)^2$ und Δ von den Modellparametern.

Beispiel 2.37 (Dreiecke in zufälligen Graphen). Betrachte n Vertices und verbinde sie alle mit Kanten: Das sind dann $m = \binom{n}{2}$ Stück. Der entstandene Graph heißt *vollständiger Graph auf/mit n Ecken.*
Seien $X_1, \ldots, X_m$ iid Bernoulli-ZV mit $P(X_i = 1) = p \in [0, 1]$. Falls $X_i = 0$, entfernen wir die entsprechende Kante aus dem Graphen. Das sich ergebende

Ensemble von Teilgraphen wir mit $G(n,p)$ bezeichent. Ein Element davon wird als *Erdős-Renyi-Graph* bezeichnet.
Er *enthält ein Dreick*, falls es drei Ecken u,v,w im vollständigen Graphen gibt, so dass die Kanten-ZV $X_{(u,v)}, X_{(v,w)}, X_{(w,u)}$ der ensprechenden Kanten $(u,v),(v,w),(w,u)$ alle den Wert Eins annehmen, d.h. die drei Kanten ‚überleben' den Zufallsprozess.
Sei $A = \{(u,v),(v,w),(w,u)\} \in \mathcal{P}([m])$. Dann bedeutet $Y_A = 1$, dass das entsprechende Dreick in dem Erdős-Renyi-Graphen vorhanden ist. Sei $\mathcal{I} \subset \mathcal{P}([m])$ die Menge aller solchen A-s. Dann ist $Z = \sum_{A\in\mathcal{I}}$ die Anzahl der Dreiecke im Erdős-Renyi-Graphen.
Man rechnet als Übung nach:

$$EZ = \binom{n}{3}p^3 \tag{2.30}$$

$$\mathrm{Var}(Z) = \binom{n}{3}(p^3 - p^6) + 2\binom{n}{4}\binom{4}{2}(p^5 - p^6) \tag{2.31}$$

$$\Delta = \binom{n}{3}p^3 + 2\binom{n}{4}\binom{4}{2}p^5 \tag{2.32}$$

und mit Janson:

$$P(Z=0) \le \exp\left(-\frac{\binom{n}{3}p^2}{2(1+3np^2)}\right). \tag{2.33}$$

Beachte, dass $\binom{n}{3} \sim n^3p^2$ für großes n. Also wächst der Exponent bei festem p wie n^2.

2.12. **Anwendung der Harris-Ungleichung: Perkolation.** Gegeben:

- Gitter $\mathbb{Z}^d$
- $p \in [0,1]$
- Kante e wird mit Wahrscheinlichkeit p beibehalten bzw. mit Wahrscheinlichkeit $(1-p)$ entfernt, unabh. von allen anderen Kanten.
- Modell entspricht Produktmaß von Bernoulli-Verteilungen mit Parameter p bzgl. $\mathbb{Z}^d$, in Zeichen: $\mathcal{B}_p^{\otimes\mathbb{Z}^d}$

Es entsteht ein zufälliger Untergraph auf $\mathbb{Z}^d$ (vgl. Abbildung ?? im Fall $d = 2$).
Man interessiert sich für die erzeugten Graphenstrukturen ohne die entfernten Kanten: Welche Eigenschaften besitzen die Cluster, d.h. die Zusammenhangskomponenten?

Das *0-1-Gesetz von Kolmogorov* liefert: Entweder

ABBILDUNG 3. Graph.

(1) es existiert ein ∞-Cluster fast sicher
oder
(2) es existiert ein ∞-Cluster fast sicher nicht.

$\overset{\text{Monotonie}}{\Rightarrow}$ Es existiert ein kritischer Wert $p_c \in (0,1)$, so dass

$p > p_c \Rightarrow$ (1)
$p < p_c \Rightarrow$ (2)

Bei $p = p_c$ unklar, was passiert!

Frage: Wir kann ich für einen passenden Wert $p \in [0,1]$ zeigen, dass f.s. ein unendlicher Cluster existiert?
→ Ein möglicher Zugang über (immer größere) Würfel bzw. Rechtecke bzw. Quader $\subset \mathbb{Z}^d$. Unendliche Cluster sind bei endlichen Unterstrukturen natürlich nie möglich, also brauchen wir eine modifizierte Sichtweise.

Strategie in $\mathbb{Z}^2$: Mit welcher W'keit gibt es einen durchgängigen Pfad zwischen dem linken und rechten Rändern eines Rechtecks?
→ Studiere dieses Verhalten in Abhängigkeit der Länge und Breite des Rechtecks.
→ Lasse die Größe der Box dann gegen unendlich laufen, um asymptotisches Perkolationsverhalten zu verstehen.

Was hat das mit der Harris-Ungleichung zu tun?
Bei der Frage, ob ‚Perkolation' bereits auf einem endlichem großen Rechteck sichtbar ist (vgl. Abbildung ??) spielen Ereignisse der folgenden Bauart

$A = \{$Es gibt einen durchgängigen Weg von der linken zur rechten Kante.$\}$
und
$B = \{$Es gibt einen durchgängigen Weg von der unteren zur oberen Kante.$\}$
eine Rolle (vgl. Abbildung ??). Wie stehen die Wahrscheinlichkeiten der Ereignisse A, B zueinander?

Klar ist: Man kann nicht mit Unabhängigkeit folgern: $P(A \cap B) = P(A)P(B)$,

Wegen Monotonie liefert Anwendung der Harris-Ungleichung:

$$P(A \cap B) \geq P(A)P(B).$$

Was hat das mit Konzentrationsungleichungen zu tun?
Russo-Seymour-Welsh Theorie:„Auf größeren Skalen sind Durchquerungswahrscheinlichkeiten größer“.
Mittels einer induktiven Zerlegung von grossen Rechtecken in kleinere kann man folgende Konzentrationsungleichung für unabhängige Kantenperkolation auf $\mathbb{Z}^2$ zeigen.

Definition 2.38.

$$\Lambda_{L,n} := \mathbb{Z}^2 \cap ([-nL, nL] \times [-L, L])$$

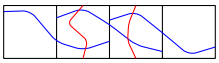
$$\Omega_{L,n} := \{\exists \text{ aktiver Pfad, der linken und rechten Rand von } \Lambda_{L,n} \text{ verbindet}\},$$

$$R_{L,n}(p) := P_p(\Omega_{L,n}).$$

Ereigniss $\Omega_{L,n}$ bzw. ZV $\mathbb{1}_{\Omega_{L,n}} \colon \{0,1\}^{\Lambda_{L,n}} \to \{0,1\}$ ist isoton im Sinne von Definition (2.28) und dessen Wahrscheinlichkeit ist Polynom in p.

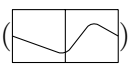
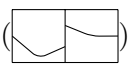
Lemma 2.39 (Reskalierung)**.**
Sei $p \in [0,1]$. Angenommen es existiert ein $L \in \mathbb{N}$ und $c \leq \frac{1}{16}$, s.d. $R_{L,2}(p) \geq 1 - c \cdot e^{-1}$. Dann gilt $\forall k \in \mathbb{N}: \; R_{2^k L,2}(p) \geq 1 - c \cdot e^{-2^k}$

Beweis. Zerlege Rechteck $\Lambda_{L,4}$ in vier Quadrate vom Typ $\Lambda_L := \Lambda_{L,1} \Rightarrow$

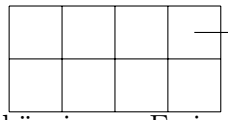
$$R_{L,4}(p) = \mathbb{P}_p\Big($$  $$\Big) \geq \mathbb{P}_p\Big($$  $$\Big)$$

$$\overset{\text{Harris}}{\geq} \Big(\mathbb{P}_p\Big($$  $$\Big)\Big)^3 \Big(\mathbb{P}_p\Big($$  $$\Big)\Big)^2$$

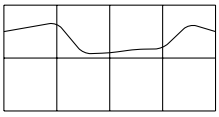
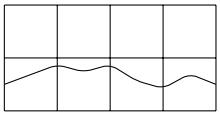
$$= R_{L,2}(p)^3 R_L(p)^2$$

Wegen $R_{L,2} = \mathbb{P}_p($  $) \leq \mathbb{P}_p($  $) = R_L^2$ gilt:

$$R_{L,4} \geq (R_{L,2})^4 \geq \Big(1 - \frac{c}{e}\Big)^4 \geq 1 - \frac{4c}{e}$$

Zerlege $\Lambda_{2L,2}$ in zwei parallele Streifen oben und unten  Λ_L

Streifen disjunkt $\Rightarrow$ Ereignisse im oberen Streifen unabhängig von Ereignissen im unteren.

$$R_{2L,2} \geq \mathbb{P}_p \left(\text{  oder  } \right)$$

$\Rightarrow$ gehe zu Komplementen über, damit "$\cap$" erscheint.

$$\Rightarrow 1 - R_{2L,2} \leq (1 - R_{L,4})^2 \leq \left(\frac{4c}{e}\right)^2 = \frac{16c^2}{e^2} \leq \frac{c}{e^2}$$

da $c \leq \frac{1}{16}$.
Fertig für $k = 1$. Weiter induktiv. □

2.13. **Negativ assoziierte ZV.** Betrachte reelwertige ZV $X_i, i \in [n]$.

Definition 2.40 (Negative Assoziation)**.** Die ZV $X_i, i \in [n]$ heißen *negativ assoziiert*, falls für beliebige disjunkte $I, J \subset [n]$ und für beliebige isotone $f\colon \mathbb{R}^I \to \mathbb{R}, g\colon \mathbb{R}^J \to \mathbb{R}$

$$E\left(f(X_I)g(X_J)\right) \leq E\left(f(X_I)\right) E\left(g(X_J)\right)$$

gilt. Hierbei haben wir die Abkürzung $X_I = (X_i)_{i \in I}$ benutzt.

Geben Sie ein Beispiel für neg. assoz. ZV an!
Welche Eigenschaften folgen aus dieser Definition?

Bemerkung 2.41. Seien die ZV $X_i, i \in [n]$ negativ assoziiert. Dann:

(1) Für jedes Paar $i \neq j \in [n]$ gilt: $E(X_iX_j) \leq E(X_i)E(X_j)$
(2) Sind $f_i, i \in [n]$, isotone, nichtnegative Funktionen so gilt

$$E\left(\prod_{i\in[n]} f_i(X_i)\right) \leq \prod_{i\in[n]} E\left(f_i(X_i)\right),$$

(3) insbesondere

$$E\left(\exp(\lambda \sum_{i\in[n]} X_i)\right) \leq \prod_{i\in[n]} E\left(e^{\lambda E X_i}\right) \quad \text{für } \lambda \geq 0.$$

[2]See: Concentration of Measure for the Analysis of Randomised Algorithms, by: Devdatt P. Dubhashi and Alessandro Panconesi

Nun können wir Ideen der Chernoff und der Hoeffding-Schranke auf diese Situation übertragen.

Lemma 2.42. *Seien $a_i \leq b_i \in \mathbb{R}$, $X_i \in [a_i, b_i]$, $i \in [n]$, zentrierte, negativ assoziierte ZV und $S = \sum_{i=1}^n X_i$. Dann gilt für $\lambda \geq 0$*

$$\psi_S(\lambda) \leq \frac{\lambda^2}{2} \sum_{i=1}^n \frac{(b_i - a_i)^2}{4}$$

$$P(S > t) \leq e^{-\frac{t^2}{2\nu}} \quad \text{für } t \geq 0.$$

Beweis.

$$\psi_S(\Lambda) = \ln E\left(\exp\left(\lambda \sum_{i=1}^n X_i\right)\right) \tag{2.34}$$

$$\leq \ln\left(\prod_{i=1}^n E\left(e^{\lambda X_i}\right)\right) \tag{2.35}$$

$$= \sum_{i=1}^n \psi_{X_i}(\lambda) \tag{2.36}$$

$$\text{(Hoeffding)} = \frac{\lambda^2}{2} \sum_{i=1}^n \frac{(b_i - a_i)^2}{4} \tag{2.37}$$

also $S \in \mathcal{G}(\nu)$ mit $\nu = \sum_{i=1}^n \frac{(b_i - a_i)^2}{4}$.
Mit der Chernoff-Methode und Argumenten wie in Bemerkung 2.8 folgt

$$P(S > t) \leq e^{-\frac{t^2}{2\nu}}$$

□

Fragen:

- Kann man die Annahmen in dem Lemma abschwächen?
- Falls $X_1, \ldots, X_n$ negativ assoz., sind dann auch die zentrierten Versionen $X_1 - E(X_1), \ldots, X_n - E(X_n)$ negativ assoz.?
- Falls $X_1, \ldots, X_n$ negativ assoz., sind dann auch die zentrierten Versionen $-X_1, \ldots, -X_n$ negativ assoz.?

2.14. **Minkowski-Ungleichung.**

Die Minkowski-Ungleichung wird im Beweis der Bonami–Beckner-Ungleichung über Hyperkontraktivität verwendet. Bekannte Minkowski-Ungleichung: $X, Y \in \mathcal{L}^q$, dann gilt

$$E(|X + Y|^q))^{\frac{1}{q}} \leq E(|X|^q)^{\frac{1}{q}} + E(|X|^q)^{\frac{1}{q}}.$$

Satz 2.43 (Minkowski-Ungleichung)**.**
Seien $X\colon \Omega \to E, Y : \Omega \to F$ *unabhängige ZVen und*

$$f\colon E \times F \to \mathbb{R},\ \textit{messbar}\ \ \textit{und}\ Z = f(X,Y) : \Omega \to \mathbb{R} \quad P_X\textit{-f.s.}\ P_Y\textit{-intbar}$$

Für $q \geq 1$ *gilt (auch für* $q = \infty$*):*

$$E_X(|E_Y(Z)|^q)^{\frac{1}{q}} \leq E_Y((E_X(|Z|^q))^{\frac{1}{q}}),\ \textit{wobei}$$
$$E_X(Z) = E(Z \mid Y) = \int_{\mathbb{R}} f(t,Y)\mathrm{d}P_X(t)\ \textit{die Integration bzgl. } P_X.$$

Also ist

$$\left(\int_{\mathbb{R}}\int_{\mathbb{R}} |f(t,s)|^q \mathrm{d}P_X(t)\mathrm{d}P_Y(s)\right)^{\frac{1}{q}} \leq \int_{\mathbb{R}} \left(\int_{\mathbb{R}} |f(t,s)|^q \mathrm{d}P_X(t)\right)^{\frac{1}{q}} \mathrm{d}P_Y(s).$$

Für den Fall $q = 1$ werden lediglich Betragsstriche reingezogen.
Frage: Gibt es einen Zusammenhang zur klassischen Minkowski-Ungleichung?
Seien dazu

- $F = \{1,2\}$, $P_Y(\{1\}) = P_Y(\{2\}) = \frac{1}{2}$
- $X = (X_1, X_2)$, $f(X,1) = X_1$, $f(X,2) = X_2$

$$\Rightarrow \left(E_X\left(\left|\frac{1}{2}(X_1+X_2)\right|^q\right)\right)^{\frac{1}{q}} \leq \frac{1}{2}\left((E_X(|X_1|^q))^{\frac{1}{q}} + (E_X(|X_2|^q))^{\frac{1}{q}})\right)$$
$$\leq \frac{1}{2}\left(\int_{\mathbb{R}} |X_1(e)|^q \mathrm{d}P_X(e)\right)^{\frac{1}{q}} + \frac{1}{2}\left(\int_{\mathbb{R}} |X_2(e)|^q \mathrm{d}P_X(e)\right)^{\frac{1}{q}}$$

Beweis. Für $q = 1$: Wende Fubini und Dreiecksungleichung an.
Für $q = \infty$: $E_X(\text{ess-sup}(|E_Y(Z)|)) \leq E_Y(E_X(\text{ess-sup}|Z|))$.
Nun $q = (1,\infty)$: o.E. sei $Z \geq 0$, daher $|Z| = Z$. Sei U unabhängige Kopie von Y und unabhängig von X.

$$E_X(E_Y(Z)^{q-1+1}) = E_X\left[E_U(f(X,U))^{q-1} E_Y(f(X,Y))\right]$$

$$= E_Y\left[E_X((E_U(f(X,U))^{q-1} f(X,Y)))\right]$$
$$\leq E_Y\left[(E_X(E_U(f(X,U)))^q)^{\frac{q-1}{q}} (E_X(f(X,Y)^q))^{\frac{1}{q}}\right]$$
$$= [E_X(E_U(f(X,U))^q)]^{\frac{q-1}{q}}\ E_Y\left[(E_X(f(X,Y)^q))^{\frac{1}{q}}\right]$$
$$= [E_X((E_Y(Z))^q)]^{1-\frac{1}{q}}\ E_Y\left[(E_X(Z^q))^{\frac{1}{q}}\right]$$

Dividiere durch $(E_X(E_Y(Z))^q)^{1-\frac{1}{q}}$, dann folgt

$$(E_X(E_Y(Z)^q))^{\frac{1}{q}} \leq E_Y((E_X(Z^q))^{\frac{1}{q}})$$

□

3. Schranken an die Varianz

Wir interessieren uns nun für ZVen der Form $Z = f(X_1, \ldots, X_n)$, wobei $X_1, \ldots, X_n : \Omega \to \mathcal{X}$ unabhängige ZVen seien (nicht notwendig identisch verteilt) und $\mathcal{X}$ ein messbarer Raum und $f \colon \mathcal{X}^n \to \mathbb{R}$ eine messbare Abbildung. Ferner lautet die Generalannahme in diesem Kapitel $Z \in \mathcal{L}^2$, da wir an Abschätzungen für die Varianz interessiert sind.

Beispiel: Sei $Z = \sum_{i=1}^n X_i$.

In diesem Fall erzielt der Beweis des Gesetztes der großen Zahlen bereits eine Konzentrationsungleichung mithilfe der Čebyšev-Ungleichung. Jetzt möchten wir allgemeinere f betrachten. Theorem 3.1 in Kapitel 3.1 findet in allen Unterkapitel von 3.2 bis 3.8 Verwendung. In Kapitel 3.9 wird ein alternativer Beweis für Theorem 3.1 angegeben. Die Kapitel 3.2 bis 3.8 lassen sich fast unabhängig voneinander lesen. Lediglich unter den Kapiteln 3.2, 3.3 und Bsp. 3.14 gibt es kleine Bezüge.

3.1. **Efron-Stein-Ungleichung.**

Zur Motivation betrachte:

$$Z = \sum_{i=1}^n X_i$$

Es gilt: $X_1, \ldots, X_n \in \mathcal{L}^2$ unabhängig $\Rightarrow$ $X_1, \ldots, X_n$ unkorrelliert $\Rightarrow$

$$\mathrm{Var}(Z) = \sum_{i=1}^n \mathrm{Var}(X_i).$$

Idee um zu Verallgemeinern: Drücke $Z - E(Z)$ für allgemeines $Z = f(X_1, \ldots, X_n) \in \mathcal{L}^2$ als Summe von *Martingaldifferenzen* bzgl. der Doob-Filtrierung aus. Notation dazu: Sei Y intbare ZVe und

$$E_i(Y) := E(Y \mid X_1, \ldots, X_i) \text{ für } i = 1, \ldots, n,$$
$$\text{wobei } E_0(Y) = E(Y).$$

Setze $\Delta_i := E_i(Z) - E_{i-1}(Z)$ für $i = 1, \ldots, n$, so dass

$$Z - E(Z) = \sum_{i=1}^{n} \Delta_i \text{ (Teleskopsumme).}$$

$$\Rightarrow \ \mathrm{Var}(Z) = E\left((Z - E(Z))^2\right) = E\left(\left(\sum_{i=1}^{n} \Delta_i\right)^2\right)$$
$$= \sum_{i=1}^{n} E(\Delta_i^2) + 2\sum_{i=1}^{n}\sum_{j>i}^{n} E(\Delta_i \Delta_j).$$

Für $i \leq k$ gilt wegen der Turmeigenschaft:

$$E_i(E_k(Z)) = E(E_k(Z) \mid X_1, \ldots, X_i) = E\left[E(Z \mid X_1, \ldots, X_k) \mid X_1, \ldots, X_i\right] E(Z \mid X_1, \ldots, X_i)$$
$$= E_i(Z)$$

Für $i < j$ folgt damit:

$$E_i(\Delta_j) = E_i\left(E_j(Z) - E_{j-1}(Z)\right) = 0$$
$$\Rightarrow E_i(\Delta_i \Delta_j) = E_i(\Delta_i(E_i Z - E_{i-1} Z)) = \Delta_i E_i(\Delta_j) = 0$$
$$\Rightarrow E(E_i(\Delta_i \Delta_j)) = E(\Delta_i E_i(\Delta_j)) = 0.$$

Also gilt auch *ohne* jegliche Unabhängigkeit/Unkorreliertheit:

$$\mathrm{Var}(Z) = E\left(\left(\sum_{i=1}^{n} \Delta_i\right)^2\right) = \sum_{i=1}^{n} E(\Delta_i^2). \tag{3.1}$$

Nimmt man die Unabhängigkeit der $X_1, \ldots, X_n$ an, folgt wie in Kapitel 2.9 bereits einmal:

$$E_i(Z) = \int_{\mathcal{X}^{n-i}} f(X_1, \ldots, X_i, t_{i+1}, \ldots, t_n)\, \mathrm{d}P_{X_{i+1}}(t_{i+1}) \ \ldots \ \mathrm{d}P_{X_n}(t_n).$$

Ist $X^{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, so gilt analog

$$E^{(i)}(Z) := E\left(Z \mid X^{(i)}\right) = \int_{\mathcal{X}} f(X_1, \ldots, X_{i-1}, x, X_{i+1}, \ldots, X_n) \mathrm{d}P_{X_i}(x).$$

Mit Fubini folgt

$$E_i\left(E^{(i)}(Z)\right) = E_{i-1}(Z). \tag{3.2}$$

Dies wird im folgenden Beweis benutzt

Satz 3.1 (Efron-Stein-Ungleichung)**.**
Seien $X_1, \ldots, X_n$ unabhängige ZVen, $X = (X_1, \ldots, X_n)$ und $Z = f(X) \in \mathcal{L}^2$.

$$\Rightarrow \ \mathrm{Var}(Z) \leq \nu := \sum_{i=1}^{n} E\left((Z - E^{(i)}(Z))^2\right).$$

Ist $X' = (X'_1, \ldots, X'_n)$ unabhängige, identische Kopie zu X und $Z'_i = f(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n)$, so gilt

$$\begin{aligned}\nu &= \frac{1}{2}\sum_{i=1}^{n} E\left((Z - Z'_i)^2\right) = \sum_{i=1}^{n} E\left((Z - Z'_i)_+^2\right) \\ &= \sum_{i=1}^{n} E\left((Z - Z'_i)_-^2\right) = \inf_{Z_1,\ldots,Z_n} \sum_{i=1}^{n} E\left((Z - Z_i)^2\right).\end{aligned}$$

wobei das Infimum über alle n-Tupel $Z_1, \ldots, Z_n$ läuft, wobei jedes Z_i $X^{(i)}$-messbar und quadratintegrierbar ist.

Beweis. Mit (3.2) folgt direkt $\Delta_i = E_i(Z - E^{(i)}(Z))$. Verwende Jensen-Ungleichung für die bedingte Erwartung:

$$\begin{aligned}&\Delta_i^2 = (E_i(Z - E^{(i)}(Z)))^2 \leq E_i\left(\left(Z - E^{(i)}(Z)\right)^2\right) \\ \Rightarrow \ &\mathrm{Var}(Z) = \sum_{i=1}^{n} E(\Delta_i^2) \leq \sum_{i=1}^{n} E(E_i(Z - E^{(i)}(Z))^2) = \sum_{i=1}^{n} E((Z - E^{(i)}(Z))^2).\end{aligned}$$

Nun sind die Gleichheiten für ν noch zu zeigen:
Für eine ZVe Y setze

$$\mathrm{Var}^{(i)}(Y) = E((Y - E(Y \mid X^{(i)}))^2 \mid X^{(i)}) = E^{(i)}((Y - E^{(i)}(Y))^2)$$

Turmeigenschaft

$$\Rightarrow \nu = \sum_{i=1}^{n} E(E^{(i)}((Z - E^{(i)}(Z))^2)) = \sum_{i=1}^{n} E(\mathrm{Var}^{(i)}(Z)).$$

Für zwei unabhängig, identisch verteilte ZVen $W, Y \in \mathcal{L}^2$ gilt:

$$\begin{aligned}\operatorname{Var}(W) &= \frac{1}{2}\operatorname{Var}(W) + \frac{1}{2}\operatorname{Var}(Y) \\ &= \frac{1}{2}(E(W^2) - \underbrace{E(W)^2}_{=E(W)E(Y)} + E(Y^2) - \underbrace{E(Y)^2}_{=E(W)E(Y)}) \\ &= \frac{1}{2}(E(W^2) + E(Y^2) - 2E(Y)E(W)) \\ &= \frac{1}{2}E((W-Y)^2).\end{aligned}$$

Analog gilt:

$$\operatorname{Var}^{(i)}(Z) = \frac{1}{2}E^{(i)}((Z - Z_i')^2),$$

falls nur die 1-dim. Integration $E^{(i)}(.)$ ausgeführt und die Unabhängigkeit benutzt wird.

Vorüberlegung:: Die Verteilung von $Y - W$ ist symmetrisch, da für unabhängig, identisch verteilte ZVen $W, Y \in \mathbb{R}$ gilt

$$P_{(W,Y)} = P_W \otimes P_Y \overset{\text{id. vert.}}{=} P_Y \otimes P_W = P_{(Y,W)}$$

Sei nun $g(a,b) := a - b$, dann sind

$$\begin{aligned}&W - Y = g(W,Y) \text{ und } Y - W = g(Y,W). \\ \Rightarrow\ &P_{W-Y} = P_{(W,Y)} \circ g^{-1} = P_{(Y,W)} \circ g^{-1} = P_{Y-W} \\ \Rightarrow\ &\forall\, t \in \mathbb{R}\colon P(W - Y \geq t) = P(Y - W \geq t).\end{aligned}$$

Unter der Annahme $W, Y \in \mathcal{L}^2$ folgt:

$$\frac{1}{2}E((Y-W)^2) = \frac{1}{2}\int_{-\infty}^{\infty} t^2 \,\mathrm{d}P_{Y-W}(t) = \int_0^{\infty} t^2 \,\mathrm{d}P_{Y-W}(t) = \int_{-\infty}^{0} t^2 \,\mathrm{d}P_{Y-W}(t).$$

Bezüglich der 1-dim. $E^{(i)}$-Integration sind die ZVen Z und Z_i' unabhängig und identisch verteilt. Also:

$$\frac{1}{2}E^{(i)}((Z - Z_i')^2) = E^{(i)}((Z - Z_i')_+^2) = E^{(i)}((Z - Z_i')_-^2).$$

Zur letzten Gleichheit: Zeige ähnlich wie oben die Aussage für eine reellwertige ZVe und folgere sie im multivariaten Fall.

Es gilt für $Y \in \mathcal{L}^2$: $\operatorname{Var}(X) = E((X - E(X))^2) = \inf_{q \in \mathbb{R}} E((X-q)^2)$.

Das Infimum wird bei $q = E(X)$ angenommen. Daher gilt für jedes $i = 1, \ldots, n$

$$\mathrm{Var}^{(i)}(Z) = \inf_{q \in \mathbb{R}} E^{(i)}((Z-q)^2),$$

falls wir nur die 1-dim. $E^{(i)}$-Integration ausführen. Der Minimierer $q_{\min}$ ist eine Funktion der $(n-1)$ Variablen $X^{(i)}$. Wie oben ist der Minimierer gerade $q_{\min} = E^{(i)}(Z)$. Dieser ist $X^{(i)}$-messbar und in $\mathcal{L}^2$. Also darf dies bei der Klasse von Funktionen über das Infimum verwendet werden. □

Im Fall $Z = \sum_{i=1}^n X_i$ gilt:

$$Z - E^{(i)}(Z) = \sum_{j=1}^n X_j - \left(\sum_{j \neq i} X_j + E(X_i) \right) = X_i - E(X_i).$$

$$\Rightarrow \nu = \sum_{i=1}^n E((X_i - E(X_i))^2) = \sum_{i=1}^n \mathrm{Var}(X_i) \overset{\text{unabh.}}{=} \mathrm{Var}(Z).$$

Beispiel zeigt: Efron-Stein-Ungleichung ist scharfe Abschätzung.

Bemerkung 3.2 (Jackknife Resampling).
In der Statistik benutzt man Subpopulationen von Datensätzen, um Parameter besser zu schätzen. *Jackknife Resampling* ist eine Vorgängerversion des *Bootstraps*, bei dem aus einer Population von N Individuen (samples) $N-1$ Subpopulationen durch Weglassen eines Individuums erzeugt werden.

Seien $X_1, \ldots, X_n$ unabhängig, identisch verteilte ZVen mit Verteilung P_X. Der zu schätzende Parameter θ ist eine Funktion von P_X, den wir durch eine Funktion $Z = f_n(X_1, \ldots, X_n)$ der Daten schätzen wollen. Um die Qualität des Schätzers zu beschreiben, verwendet man u.a.

Bias: $E(Z) - \theta$
Mittlere quadratische Abweichung: $\mathrm{Var}(Z) + (E(Z) - \theta)^2$

Allerdings können wir $E(Z)$ und $\mathrm{Var}(Z)$ nicht explizit ausrechnen, da P_X unbekannt ist. Also benutzt man eine empirische Mittlung anhand der Daten.

Jackknife-Schätzer für den Bias:

$$(n-1)\left(\frac{1}{n} \sum_{i=1}^n (Z_i - Z) \right), \text{ wobei } Z_i := f_{n-1}\left(x^{(i)} \right).$$

$X^{(i)}$ heißt auch i-tes Jackknife-Sample und Z_i die i-te Jackknife-Replikation.

Jackknife-Schätzer für die Varianz:

$$\frac{n-1}{n}\sum_{i=1}^{n}(Z-Z_i)^2$$

Die Efron-Stein-Ungleichung besagt, dass dieser Schätzer immer einen nichtnegativen Bias besitzt. D.h. wir schätzen die Varianz eher zu hoch als zu tief ein. Bei statistischen Verfahren sind wir damit eher auf der sichern Seite.

Prinzip: Pessimismus > Optimismus (bei Varianzen)

Es werden in wachsender Allgemeinheit verschiedene Anwendungen der Efron-Stein Ungleichung vorgestellt.

3.2. **Funktionen mit beschränkter Differenz.**

Definition 3.3. $f\colon \mathcal{X}^n \to \mathbb{R}$ hat *beschränkte Differenzen* (oder die beschränkte Differenzeneigenschaft), falls es $c_1,\dots,c_n \geq 0$ gibt, so dass

$$\sup_{x_1,\dots,x_n,x_i'\in\mathcal{X}} |f(x_1,\dots,x_n)-f(x_1,\dots,x_i',\dots,x_n)| \leq c_i \,\forall\, i.$$

Korollar 3.4.
Hat f beschränkte Differenzen mit Konstanten $c_1,\dots,c_n$ und sind $X_1,\dots,X_n$ unabhängige ZVen mit Werten in $\mathcal{X}$, so gilt für $Z := f(X_1,\dots,X_n)$:

$$\mathrm{Var}(Z) \leq \frac{1}{4}\sum_{i=1}^{n} c_i^2.$$

Beweis. E-S-U besagt

$$\mathrm{Var}(Z) \leq \inf_{Z_1,\dots,Z_n} \sum_{i=1}^{n} E\left((Z-Z_i)^2\right).$$

In das Infimum dürfen wir die quadratintegrierbare und $X^{(i)}$-messbare 'Itervallmitte'

$$\begin{aligned} Z_i = \frac{1}{2}\Big(&\sup_{x\in\mathcal{X}} f(X_1,\dots,X_{i-1},x,X_{i+1},\dots,X_n) \\ &+ \inf_{y\in\mathcal{X}} f(X_1,\dots,X_{i-1},y,X_{i+1},\dots,X_n)\Big) \\ &\leq \frac{c_i}{2} \end{aligned}$$

einsetzen. Es folgt:

$$(Z-Z_i)^2 \leq \frac{c_1}{4} = (\text{ halbe Intervallänge })^2$$

□

Beispiel 3.5 (Längste gemeinsame Teilfolge). Seien $X_1, \ldots, X_n, Y_1, \ldots, Y_n \sim Ber(\frac{1}{2})$ unabhängige ZVen. Betrachte

$$\begin{aligned} Z &:= f(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \\ &:= \max\{k \mid X_{i_1} = Y_{j_1}, \ldots, X_{i_k} = Y_{j_k}, 1 \leq i_1 < \ldots < i_k \leq n, 1 \leq j_1 < \ldots < j_k \leq n\}. \end{aligned}$$

Z ist die Länge der längsten Teilfolge, die in beiden Folgen auftaucht. Beispiel:

$$\begin{aligned} &\text{(i) } X = (0,0,0,0,0), Y = (1,1,1,1,1) \;\Rightarrow\; Z = 0 \\ &\text{(ii) } X = (1,0,1,0,1), Y = (0,1,0,1,0) \;\Rightarrow\; Z = 4 \end{aligned}$$

Es ist bekannt, dass $\lim_{n\to\infty} E(Z)/n$ f.s. gegen eine Konstante γ konvergiert, deren Wert allerdings unbekannt ist. Vermutung:

$$\gamma := \lim_{n\to\infty} \frac{E(Z)}{n} = \frac{2}{1+\sqrt{2}}$$

Wie im Kontext vom üblichen Gesetz der großen Zahlen wollen wir das Konzentrationsphänomen über die $\mathrm{Var}(Z)$ verstehen. Ändert man nur eine einzige Ziffer in der Folge $X_1, \ldots, X_n, Y_1, \ldots, Y_n$, kann sich Z um höchstens Eins ändern, also hat Z beschränkte Differenzen mit $c_{i,x} = c_{i,y} = 1$.

$$\Rightarrow \quad \mathrm{Var}(Z) \leq \frac{1}{4} \sum_{i=1}^{2n} 1 = \frac{n}{2}$$

Beachte: EZ wächst proportional zu n. Die Čebyšev-Ungleichung liefert

$$P\left(\left|\frac{Z - E(Z)}{n}\right| \geq t\right) = P(|Z - E(Z)| \geq nt) \leq \frac{\frac{n}{2}}{n^2 t^2} = \frac{1}{2t^2}\frac{1}{n}$$

bzw.

$$P(|Z - E(Z)| \geq s\sqrt{n}) \leq \frac{\frac{n}{2}}{ns^2} = \frac{1}{2s^2}$$

Mit hoher Wahrscheinlichkeit fällt Z in ein am Erwartungswert zentriertes Intervall der Länge $c \times \sqrt{n}$.

Definition 3.6. Eine Folge $(Z_n)_{n\in\mathbb{N}}$ nichtnegativer ZVen heißt *relativ stabil*, wenn $\frac{Z_n}{E(Z_n)} \to 1$ in Wahrscheinlichkeit (Fluktuationen von Z_n um $E(Z_n)$ werden klein).

Um relative Stabilität z.z., ist es oft nützlich Schranken an $\mathrm{Var}(Z_n)$ z.z., denn

$$P\left(\left|\frac{Z_n}{E(Z_n)} - 1\right| \geq \varepsilon\right) = P\left(|Z_n - E(Z_n)| \geq \varepsilon E(Z_n)\right) \leq \frac{\mathrm{Var}(Z_n)}{\varepsilon^2 E(Z_n)^2}.$$

Beispiel 3.7. (Schätzen von Dichten durch Glättungskerne)
Sei $X_1, \ldots, X_n$ unabhängig, identisch verteilt gemäß unbekannter Dichte $\rho\colon \mathbb{R} \to [0, \infty)$. Intuitiv würde man anhand von n Daten $x_1, \ldots, x_n$ die Verteilung von X_1 durch das empirische Maß $\mu_n := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ schätzen. Dieses Maß hat allerdings keine Dichte. Um dem Rechnung zu tragen, ersetzt man das Punktmaß durch eine *Approximation der Eins*, eine stark konzentrierte W'Dichte. Wir schätzen ρ durch

$$\rho_n(x) := \rho_{n;(x_1,\ldots,x_n)}(x) := \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - x_i}{h_n}\right),$$

wobei $h_n > 0$ *Glättungsparameter* heißt und $K : \mathbb{R} \to [0, \infty)$, $\int_{\mathbb{R}} K(x)dx = 1$ die *Approximation der Eins* ist. Wir interessieren uns für die $\mathcal{L}^1$-Norm des Fehlers, d.h.

$$Z = f(X_1, \ldots, X_n) = \int_{\mathbb{R}} |\rho(x) - \rho_{n;(X_1,\ldots,X_n)}(x)|dx.$$

Bei folgender Differenz kürzen sich $2(n-1)$ Summanden weg:

$$\begin{aligned}
|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| &= \int_{\mathbb{R}} \big||\rho(x) - \rho_n(x)| - |\rho(x) - \rho_n(x)|\big|\, dx \\
&\leq \frac{1}{nh_n} \int_{\mathbb{R}} \left| K\left(\frac{x - x_i}{h_n}\right) - K\left(\frac{x - x_i'}{h_n}\right)\right| \\
&\leq \frac{1}{nh_n} \int_{\mathbb{R}} \left(K\left(\frac{x - x_i}{h_n}\right) + K\left(\frac{x - x_i'}{h_n}\right)\right) = \frac{2}{n}.
\end{aligned}$$

Mit Korollar 3.4 folgt unmittelbar

$$\mathrm{Var}(Z) \leq \frac{n}{4} \frac{2^2}{n^2} = \frac{1}{n}$$

Beispiel 3.8. (Supremum eines Rademacher-Prozesses)
Seien $(\alpha_{i,t})_{i=1,\ldots,n,t\in\mathcal{T}}$ reelle Zahlen und $X_1, \ldots, X_n$ unabhängig, identisch verteilte Rademacher-ZVen. Definiere

$$Z := \sup_{t\in\mathcal{T}} \sum_{i=1}^n \alpha_{i,t} X_i.$$

Z heißt *Rademacher-Mittel*. $E(Z)$ hängt stark von der Wahl der $\alpha_{i,t}$ ab. Aber ändert man ein X_i, dann ändert sich Z um höchstens $2 \sup_{t\in\mathcal{T}} |\alpha_{i,t}|$. Korollar 3.4 impliziert dann:

$$\mathrm{Var}(Z) \leq \sum_{i=1}^n \sup_{t\in\mathcal{T}} |\alpha_{i,t}|^2.$$

Dies Schranke kann mit der E-S-U noch verbessert werden.

Seien $X_1', \ldots, X_n'$ unabhängige Kopien von $X_1, \ldots, X_n$ und definiere

$$Z_i' := \sup_{t \in \mathcal{T}} \left(\left(\sum_{j: j \neq i} X_j \alpha_{j,t} \right) + X_i' \alpha_{i,t} \right).$$

Sei t^* der (zufällige) Index in $\mathcal{T}$, so dass

$$\sum_{j=1}^{n} X_j \alpha_{j,t^*} = \sup_{t \in \mathcal{T}} \sum_{j=1}^{n} X_j \alpha_{j,t}.$$

Nach Definition des Supremus:

$$Z_i' = \sup_{t \in \mathcal{T}} \left(\left(\sum_{j: j \neq i} X_j \alpha_{j,t} \right) + X_i' \alpha_{i,t} \right) \geq \left(\sum_{j: j \neq i} X_j \alpha_{i,t^*} \right) + X_i' \alpha_{i,t^*}$$

Dann gilt für jedes i:

$$Z - Z_i' = \sum_{j=1}^{n} X_j \alpha_{j,t^*} - Z_i' \leq (X_i - X_i') \alpha_{i,t^*}$$

Fallunterscheidung ergibt:

$$(Z - Z_i)_+^2 \leq (X_i - X_i')^2 \alpha_{i,t}^2.$$

Turmeigenschaft und Unabhängigkeit ergeben

$$\begin{aligned} E((Z - Z_i')_+^2) &\leq E(E((X_i - X_i')^2 \alpha_{i,t^*}^2 \mid X_1, \ldots, X_n)) \\ &\leq E(\alpha_{i,t^*}^2 E(X_i^2 - 2X_i X_i' + X_i'^2 \mid X_1, \ldots, X_n)) \\ &\leq E(\alpha_{i,t^*}^2 E(1 - 2X_i X_i' + 1 \mid X_1, \ldots, X_n)) \\ &= 2E(\alpha_{i,t^*}^2) \end{aligned}$$

Da T^* nicht von dem Laufindex i abhängt, impliziert E-S-U: $\mathrm{Var}(Z) \leq 2E\left(\sum_{i=1}^n \alpha_{i,t^*}^2\right) \leq 2E\left(\sup_{t \in \mathcal{T}} \sum_{i=1}^n \alpha_{i,t}^2\right)$. Nun steht das Supremum ausserhalb der Summe!

3.3. **Selbstbeschränkende Funktionen.**

Definition 3.9. Sei $\mathcal{X} \neq \emptyset$ eine Menge und $n \in \mathbb{N}$. Eine Funktion $f \colon \mathcal{X}^n \to \mathbb{R}_+$ heißt *selbstbeschränkend* oder *self-bounding*, falls
$\forall\, i \in \{1, \ldots, n\}\, \exists f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ mit $\forall\, x_1, \ldots, x_n \in \mathcal{X}$

(1)

$$0 \leq f(x_1, \ldots, x_n) - f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1.$$

(Approximation der n-dim. Funktion mit $(n-1)$-dim. Funktion mit maximalem Fehler 1)

(2) $\sum_{i=1}^{n}(f(x_1,\ldots,x_n) - f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)) \leq f(x_1,\ldots,x_n)$
($\hat{=}$ self-bounding).

Welche Folgerungen ergeben sich?
Für $i \in \{1,\ldots,n\}$, $x_1,\ldots,x_n,x_i' \in \mathcal{X}$ gilt:

$$\begin{aligned}
&|f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \\
\leq\ &|f(x_1,\ldots,x_n) - f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)| \\
+\ &|f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \\
\leq\ &2, \text{ also hat } f \text{ beschränkte Differenzen.}
\end{aligned}$$

Erwartungshaltung: Solche f liefern gute Konzentrationsungleichungen.

$$\begin{aligned}
&\sum_{i=1}^{n}(f(x_1,\ldots,x_n) - f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n))^2 \\
\overset{(1)}{\leq}\ &\sum_{i=1}^{n}(f(x_1,\ldots,x_n) - f_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)) \cdot 1 \\
\overset{(2)}{\leq}\ &f(x_1,\ldots,x_n)
\end{aligned}$$

Die vorherige Rechnung impliziert folgendes Korollar:

Korollar 3.10.
Seien $X_1, \dots, X_n : \Omega \to \mathcal{X}$ unabhängige ZVen. Sei $Z = f(X_1, \dots, X_n)$ für ein selbstbeschränkendes $f \colon \mathcal{X}^n \to \mathbb{R}_+$. Dann

$$\mathrm{Var}(Z) \leq E(Z).$$

Beweis. Hier ist Z_i quadratintbar und $X^{(i)}$ messbar:

$$\begin{aligned} \Rightarrow \ \mathrm{Var}(Z) &\overset{\text{Satz 3.1}}{\leq} \sum_{i=1}^{n} \inf_{Z_1,\dots,Z_n} E((Z - Z_i)^2) \\ &\leq \sum_{i=1}^{n} E((Z - f_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n))^2) \\ &\overset{\text{s.o.}}{\leq} E(f(X_1, \dots, X_n)) = E(Z) \end{aligned}$$

□

→ self-bounding-Eigenschaft schließt schwere Ränder (Konzentration des Maßes an den Rändern) aus. Kontrast mit Rademacher ZV.

Anwendung bei verschiedenen Modellen. Hat für jedes $n \in \mathbb{N}$ $Z(n) = f_n(X_1, \dots, X_n)$ die selbstbeschränkende Eigenschaft, so folgt $\forall\, \varepsilon > 0$:

$$\begin{aligned} P\left(\left|\frac{Z(n)}{E(Z(n))} - 1\right| > \varepsilon\right) &= P\left(|Z(n) - E(Z(n))| > \varepsilon E(Z(n))\right) \\ &\leq \frac{\mathrm{Var}(Z(n))}{\varepsilon^2 E(Z(n))^2} \leq \frac{1}{\varepsilon^2 E(Z(n))}. \end{aligned}$$

Falls $E(Z(n))) \to \infty$ für $n \to \infty$, so liegt relative Stabilität vor.
Es ergeben sich also zwei Vorgehensweisen, um die self-bounding property auszunutzen: Zwei Szenarien beim Abschätzen:

- $E(Z(n))$ von oben abschätzen.
- $\frac{1}{E(Z(n))}$ von oben abschätzen.

Greife eine Klasse von Funktionen heraus, die selbstbeschränkend sind:

Definition 3.11. Sei $\mathcal{X} \neq \emptyset$ Menge, $n \in \mathbb{N}$, $\Pi_i \subset \mathcal{X}^i$ für $i = 1, \dots, n$ und Π das n-Tupel: $(\Pi_1, \dots, \Pi_n)$. Für $m \leq n$ sagen wir, dass der Vektor $(x_1, \dots, x_m) \in \mathcal{X}^m$ die Eigenschaft Π hat $:\Leftrightarrow (x_1, \dots, x_m) \in \Pi_m$.
Π heißt *erblich* $:\Leftrightarrow$ Hat $(x_1, \dots, x_n)$ die Eigenschaft Π, dann auch jeder Teilvektor $(x_{i_1}, \dots, x_{i_k}) = (x_j)_{j \in I}$ mit $I = \{i_1, \dots, i_k\} \subset \{1, \dots, n\}$
Ist Π eine erbliche Eigenschaft, so heißt die Funktion f , die jedem $(x_1, \dots, x_m)$,

$m \leq n$, die Mächtigkeit der längsten Teilfolge $(x_{i_1}, \ldots, x_{i_k})$ von $(x_1, \ldots, x_m)$ zuordnet, die die Eigenschaft Π hat, die *Konfigurationsfunktion* von Π. Eigentlich umfaßt f gleich n Funktionen

$$f_1 \colon \mathcal{X} \to \mathbb{N}_0, f_2 \colon \mathcal{X}^2 \to \mathbb{N}_0, \ldots, f_n \colon \mathcal{X}^n \to \mathbb{N}_0$$

Bei der VC-Theorie in Kapitel und Bsp. 3.14 spielt die Konfigurationseigenschaft eine tragende Rolle.

Korollar 3.12.
Sei Π *eine erbliche Eigenschaft auf* $\mathcal{X} \neq \emptyset$, f *die Konfigurationsfunktion von* Π, $X_1, \ldots, X_n$ *unabhängige ZVen mit Werten in* $\mathcal{X}$ *und* $Z = f(X_1, \ldots, X_n)$.

$$\Rightarrow \ \mathrm{Var}(Z) \leq E(Z).$$

Beweis. Nach Korrolar 3.10 genügt es zu zeigen, dass f selbst-beschränkend ist. Wegen Erblichkeit gilt

$$Z_i = f_{n-1}(X^{(i)}) \leq f_n(X) \text{ und } Z \leq Z_i + 1 \quad \Rightarrow \ 0 \leq Z - Z_i \leq 1$$

Andererseits: Gilt für ein $(x_1, \ldots, x_n) \in \mathcal{X}^n$: $Z = f(x_1, \ldots, x_n) = k$, so existiert nach Definition eine Teilfolge $(x_{i_1}, \ldots, x_{i_k})$ von $(x_1, \ldots, x_n)$ mit $f(x_{i_1}, \ldots, x_{i_k}) = k$. Sei $I = \{i_1, \ldots, i_k\}$. Für $i \notin I$ gilt: $Z = Z_i$, denn beide haben $(x_j)_{j \in I}$ als Teilfolge.

$$\Rightarrow \ \sum_{i=1}^{n} (Z - Z_i) = \sum_{i \in I} (Z - Z_i) \leq \sum_{i \in I} 1 = |I| = k = Z.$$

Also ist f selbst-beschränkend. □

Zusammenfassung:

f Konfigurationsfunktion $\Rightarrow f$ ist self-bounding
$\Rightarrow \mathrm{Var}(Z) \leq E(Z)$ für $Z = f(X_1, \ldots, X_n)$.

Beispiel 3.13 (Anzahl verschiedener Werte in einer Stichprobe).
Seien $X_1, \ldots, X_n : \Omega \to \mathbb{N}$ unabhängig, identisch verteilte ZVen mit Verteilung $P(X_1 = k) = p_k \in [0, 1]$, so dass $\sum_{k \in \mathbb{N}} p_k = 1$. Ferner betrachte

$$Z_n = \#\text{ unterschiedliche Werte in } (x_1, \ldots, x_n).$$

Anders geschrieben heißt das

$$\begin{aligned} Z_n &= 1 + \sum_{i=2}^{n} \mathbb{1}_{\{x_i \neq x_1, \ldots, x_i \neq x_{i-1}\}} \\ &= \sum_{i=1}^{n} \mathbb{1}_{\{x_i \neq x_1, \ldots, x_i \neq x_{i-1}\}} \\ &= \sum_{i=1}^{n} \prod_{j=1}^{i-1} \mathbb{1}_{\{x_i \neq x_j\}}. \end{aligned}$$

$$\begin{aligned} \Rightarrow \; E(Z_n) &= \sum_{i=1}^{n} E\left(\prod_{j=1}^{i-1} \mathbb{1}_{\{X_i \neq X_j\}} \right) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{\infty} E(\mathbb{1}_{\{X_i \neq X_1\}} \ldots \mathbb{1}_{\{X_i \neq X_{i-1}\}} \mid X_i = k) P(X_i = k) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{\infty} (1 - p_k)^{i-1} p_k \end{aligned}$$

Es gilt:

$$\lim_{n \to \infty} \frac{E(Z_n)}{n} = 0 \text{ (Übung)},$$

d.h. Wiederholungen sind nicht vernachlässigbar, bremsen Wachstum von Z_n.

Welche Konzentrationsungleichungen können wir für Z_n zeigen?

(A) $Z_n = f(X_1, \ldots, X_n)$ ist Funktion mit beschränkten Differenzen mit Schranken $c_i = 1 \, \forall \, i = 1, \ldots, n$.

$$\overset{\text{Kor. } 3.4}{\Rightarrow} \; \mathrm{Var}(Z_n) \leq \frac{n}{4} \Rightarrow \; \frac{\mathrm{Var}(Z_n)}{4n} \quad \text{beschränkt}$$

$$\begin{aligned} & P\left(\left| \frac{Z_n}{n} \right| > \varepsilon \right) \leq \frac{E(Z_n)^2}{\varepsilon^2 n^2} \\ &= \frac{1}{n^2 \varepsilon^2} \left(E((Z_n - E(Z_n))^2) + E(Z_n)^2 \right) \\ &= \frac{\mathrm{Var}(Z_n)}{\varepsilon^2 n^2} + \frac{1}{\varepsilon^2} \left(\frac{E(Z_n)}{n} \right)^2 \\ &\leq \frac{1}{4\varepsilon^2 n} + o(n) = o(n) \end{aligned}$$

Somit liegt stochastische Konvergenz gegen 0 vor:

$$\frac{Z_n}{n} \to 0.$$

Ist das optimal?

(B) Sei $f(x_1, \ldots, x_n)$ eine Konfigurationsfunktion zur Eigenschaft

$$(x_1, \ldots, x_m) \in \Pi_m \ :\Leftrightarrow\ \forall\, x_1, \ldots, x_m \text{ unterschiedlich}$$

Π ist erblich.

$$\overset{\text{Kor. 3.12}}{\Rightarrow} \operatorname{Var}(Z_n) \le E(Z_n)$$
$$\Rightarrow \frac{\operatorname{Var}(Z_n)}{n} \le \frac{E(Z_n)}{n}.$$

Also ist der Quotient $\left(\frac{\operatorname{Var}(Z_n)}{n}\right)_{n\in\mathbb{N}}$ Nullfolge und nicht nur beschränkt.
(C) Kann man hier Janson-Ungleichung anwenden?

Beispiel 3.14 (Vapnik-Chervonenkis-Dimension). aus der statistischen Lerntheorie. Sei $\mathcal{X} \neq \emptyset$ beliebige Menge, $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$ (z.B. σ-Algebra) und $H \subset \mathcal{X}$ endlich ($\hat{=}$ sample). Sei $G \subset H$. Wir sagen

$$\mathcal{A} \text{ identifiziert } G :\Leftrightarrow\ \exists A \in \mathcal{A} : A \cap H = G$$
$$s(\mathcal{A}, H) := \sharp\{G \subset H \mid \mathcal{A} \text{ identifiziert } G\}$$
$$= \sharp\{A \cap H \mid A \in \mathcal{A}\} =: \sharp \operatorname{Tr}_A(H) \le 2^{\sharp H}$$

H ist vollständig identifiziert von $\mathcal{A} :\Leftrightarrow\ s(\mathcal{A}, H) = 2^{\sharp H}$,

d.h. jede Teilmenge von H wird identifiziert. Betrachte *VC-Wachstumsfunktion* von $\mathcal{A}$

$$n \mapsto s_n(\mathcal{A}) := \sup_{H \in \mathcal{F}_n} s(\mathcal{A}, H),$$

wobei $\mathcal{F}_n = \{H \in \mathcal{P}(\mathcal{X}) \mid \sharp H = n\}$. Die *VC-Dimension* von $\mathcal{A}$ ist

$$D_{\mathcal{A}} := \sup\{n \in \mathbb{N} \mid s_n(\mathcal{A}) = 2^n\}.$$

Die *VC-Dimension* von $\mathcal{A}$ bzgl. H ist

$$D_{\mathcal{A}}(H) := \sup\{n \in \mathbb{N} \mid \text{ es existiert } G \subset H \text{ mit } \sharp G = n,$$
$$\text{das von } \mathcal{A} \text{ vollständig identifiziert wird}\}.$$

Sei nun $\mathcal{A}$ fixiert. Falls $X_1, \ldots, X_n \colon \Omega \to \mathcal{X}$ ZVen, so bezeichnen wir mit

$$D(X) = D(X_1, \ldots, X_n) = D(H) \quad \text{mit} \quad H := \{X_1(\omega), \ldots, X_n(\omega)\}.$$

Das ist eine Konfigurationsfunktion zur Eigenschaft Π „vollständig identifizierbar“. Diese ist erblich. Unter der Annahme das $X_1, \ldots, X_n$ unabhängig sind, folgt also mit Korollar 3.12

$$\mathrm{Var}(D(X)) \leq E(D(X)).$$

3.4. **Exkurs: Ursprung der VC-Theorie:** Seien $F\colon \mathbb{R} \to [0,1]$ eine Verteilungsfunktion und $X_i \sim F$ für $i = 1, \ldots, n$, unabhängig. Sei $F_n(t) = \frac{1}{n}\sharp\{i \mid X_i \leq t\}$ die empirische Verteilungsfunktion von $X_1, \ldots, X_n$.

$$\overset{\text{Hoeffding}}{\Rightarrow} \forall\, t \in \mathbb{R}, \varepsilon > 0, n \in \mathbb{N} : P(|F_n(t) - F(t)| \geq \varepsilon) \leq 2e^{-n\varepsilon^2/2}. \tag{3.3}$$

Satz von *Glivenko-Cantelli* dagegen besagt:

$$\forall\, n \in \mathbb{N}, \varepsilon > 0 : P(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \geq \varepsilon) \to 0, \tag{3.4}$$

Man sieht, dass dies eine ‚Uniformisierung‘ von (3.3) ist. Als Einstieg in die weitere Diskussion ist es günstig, letztere Aussage zu

$$\forall\, n \in \mathbb{N}, \varepsilon > 0 : P(\|P_n - P\|_{\mathcal{A}} \geq \varepsilon) \to 0,$$

umzuschreiben, wobei

$$P_n(A) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{X_i \in A\}}.$$

das *empirisches Maß*, $\|P_n - P\|_{\mathcal{A}} := \sup_{A\in\mathcal{A}} \|P_n(A) - P(A)\|$ und $\mathcal{A}$ das Mengensystem $\mathcal{A} := \{(-\infty, t] \mid t \in \mathbb{R}\}$ ist.
Es stellt sich die Frage, ob solch eine Konvergenz auch für andere Mengensysteme $\mathcal{A}$ gilt? Z.B. für $\mathcal{A} = \mathcal{B}(\mathbb{R})$? Durch welche Größe wird dann $\|F_n - F\|_\infty$ ersetzt? Konvergiert die Differenz auch mit dem neuen Konvergenzbegriff gegen 0?

Satz 3.15 (Theorem von Vapnik-Chervonenkis)**.** *Sei $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$. Für alle $n \in \mathbb{N}$ und $\varepsilon > 0$ gilt*

$$P(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \varepsilon) \leq 8 \underbrace{s_n(\mathcal{A})}_{\text{siehe oben}} e^{-\frac{n\varepsilon^2}{32}}$$

Allerdings: Was weiß man überhaupt über $s_n(\mathcal{A})$?

Satz 3.16 (Theorem von Sauer)**.**

$$\forall\, n \geq D_{\mathcal{A}} \text{ gilt: } s_n(\mathcal{A}) \leq (n+1)^{D_{\mathcal{A}}}$$

Das Theorem von Sauer besagt: Sobald das maximal mögliche exponentielle Wachstum abbricht, ist es sogar nur noch polynomial und der Exponent is die VC-Dimension.

Theoreme von Vapnik-Chervonenkis und Sauer kombiniert ergeben:

$$\forall\, n \geq D_A:\ P(\|P_n - P\|_{\mathcal{A}} > \varepsilon) \leq 8 s_n(\mathcal{A}) e^{-\frac{n\varepsilon^2}{32}} \leq 8(n+1)^{D_{\mathcal{A}}} e^{-\frac{n\varepsilon^2}{32}}$$

Bemerkung 3.17 (Besinnen uns nochmal:). Was ist eigentlich der formale Rahmen vom Satz von Glivenko-Cantelli?

$$\|F_n - F\|_\infty = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{} 0 \text{ f.s.}$$

für i.i.d. $X_1, \ldots, X_n \sim F$ und die zugehörige empirische Verteilungsfunktion

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, X_i]}(t).$$

Frage: Fast sicher bezüglich welchem Maß? (P oder P_X oder P_n?)
O.E. setzt man $\Omega = \mathbb{R}$, $\mathcal{A} = \mathcal{B}$, $X = id$ und $P = P_X$.

Allgemeiner: Sei $(S, \mathcal{S})$ Messraum, $\mathcal{M}_1(S, \mathcal{S})$ Raum der W'maße auf $(S, \mathcal{S})$ und i.i.d. ZVen $X_1, \ldots, X_n : \Omega \to S$ mit P_X als Verteilung.
(Häufig wird zur Vereinfachung $\Omega = S$ und $P = P_X$ gesetzt.)
empirisches Maß: $P_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i)$ für $A \in \mathcal{S}$.
Sei

$$\mathcal{C} \subset \mathcal{S}$$

oder

$$\mathcal{F} \subset \{f \colon S \to \mathbb{R} \text{ messbar, beschränkt}\}$$

und definiere ZV (falls messbar)

$$\|P_n - P_X\|_{\mathcal{C}} := \sup_{A\in\mathcal{C}} |P_n(A) - P_X(A)|$$

bzw.

$$\|P_n - P_X\|_{\mathcal{F}} := \sup_{A \in \mathcal{C}} |P_n(f) - P_X(f)|$$

$$\text{wobei } \mu(f) = \int f \, d\mu$$

Vorsicht! Nicht notwendigerweise messbar, da $\mathcal{C}, \mathcal{F}$ überabzählbar sein kann. Man braucht im Allgemeinen Zusatzannahmen an $\mathcal{C}$ bzw. $\mathcal{F}$, wie z.B. Separabilitätseigenschaften!

Satz 3.18. *Sei $\mathcal{C} \subset \mathcal{S}$ wie oben und $P_X \in \mathcal{M}_1(S, \mathcal{S})$. Dann sind äquivalent:*

$$\begin{aligned} &(i) \quad \|P_n - P_X\|_{\mathcal{C}} \xrightarrow[n\to\infty]{} 0 \text{ } f.s. \\ &(ii) \quad \|P_n - P_X\|_{\mathcal{C}} \xrightarrow[n\to\infty]{} 0 \text{ } stoch. \\ &(iii) \quad E(\|P_n - P_X\|_{\mathcal{C}}) \xrightarrow[n\to\infty]{} 0. \end{aligned}$$

Definition 3.19. Wir definieren $\mathcal{C} \subset \mathcal{S}$ als *Glivenko-Cantelli-Klasse* (GC-Klasse) für P_X, falls eine (und damit alle) der drei obigen Bedingungen erfüllt ist.

$\mathcal{C} \subset \mathcal{S}$ heißt *universelle GC-Klasse* $:\Leftrightarrow$ $\mathcal{C}$ ist eine GC-Klasse für jedes $P_X \in \mathcal{M}_1$.

$\mathcal{C} \subset \mathcal{S}$ heißt *uniforme* GC-Klasse $:\Leftrightarrow \sup_{P_X \in \mathcal{M}_1} E(\|P_n - P_X\|_{\mathcal{C}}) \xrightarrow[n\to\infty]{} 0$ (Konvergenz glm. bzgl. Verteilungen).

$\mathcal{C} \subset \mathcal{S}$ heißt *VC-Klasse* $:\Leftrightarrow$ VC-Dimension $D_{\mathcal{C}} < \infty$.

Satz 3.20 (Vapnik-Chervonenkis)**.** *$\mathcal{C} \subset \mathcal{S}$ ist genau dann eine uniforme GC-Klasse, wenn es eine VC-Klasse ist.*

Bemerkenswert: Zusammenhang zwischen Wahrscheinlichkeits- und Maßtheorie zu Komplexitäts- und Mengentheorie. (Etwas anschaulich formuliert: Maßtheoretische Strukturen für Suprema von überabzählbaren Mengen instabil.)

Beispiel 3.21. Sei $\mathcal{A} := \{F \sqsubset S\} = \{$ endliche Teilmengen von $S\}$ und P ein W'Mass auf $\mathcal{S}$ ohne Atome. Für $A_n := \{X_1, \ldots, X_n\}$ gilt $P(A_n) = 0$ und $P_n(A_n) = P_n^\omega(A_n^\omega) = 1$. $\mathcal{A}$ ist keine GC-Klasse für P.

Beispiel 3.22. $S = \mathbb{R}$, $\mathcal{C} = \{(-\infty, t] \mid t \in \mathbb{R}\}$.
GC-Theorem impliziert: $\mathcal{C}$ ist GC-Klasse.

In der Statistik ist folgende Aussage im Kontext des Kolmogorov-Smirnov-Test wichtig:

Satz 3.23 (Satz von Kolmogorov:)**.**

$$\sup_{P_X \in \mathcal{M}_1(\mathbb{R}, \mathcal{B})} \|P_n - P_X\|_{\mathcal{C}} \sim \frac{1}{\sqrt{n}}$$

Insbesondere ist $\mathcal{C}$ eine uniforme GC-Klasse, sogar mit expliziter Fehlerrate!

Verschärfung:

$$\sqrt{n}\|P_n - P_X\|_{\mathcal{C}} \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{t\in[0,1]} |B(F(t))|,$$

wobei B die Brownsche Brücke ist.

Insbesondere gilt für stetiges F via Umparametrisierung durch das Simulationslemma/Quantiltransformation: Invarianzprinzip von Donsker für empirische Verteilungen

$$\sqrt{n}\|P_n - P_X\|_{\mathcal{C}} \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{t\in[0,1]} |B(t)|.$$

Definition 3.24. Ist $\mathcal{C} \subset \mathcal{S}$ und $\mathcal{M} \subset \mathcal{M}_1(S, \mathcal{S})$, so heißt $(\mathcal{C}, \mathcal{M})$ ein *GC-Paar*, falls

$$E(\|P_n - P_X\|_{\mathcal{C}}) \xrightarrow[n\to\infty]{} 0$$

für jedes $P_X \in \mathcal{M}$.

Falls $\mathcal{C}$ keine universelle GC-Klasse ist, macht es Sinn zu prüfen, ob man sich beider konkreten Anwendung auf eine Klasse $\mathcal{M}$ von Maßen beschränken kann, so dass $(\mathcal{C}, \mathcal{M})$ ein GC-Paar sind. Manchmal hat man ja a-priori Informationen über die in Frage kommenden Maße.
Analog für $\mathcal{F} \subset \{f\colon \to \mathbb{R}$ messbar, beschränkt$\}$.

Beispiel 3.25 (keine universelle GC-Klasse). Seien $X_j \sim \mathcal{N}(0,1)$, $j \in \mathbb{N}$, i. i. d. standardnormalverteilte ZV und $Y_j := -X_j$. Betrachten die ZVektoren $Z_j = (X_j, Y_j)^\top \in \mathbb{R}^2$ und deren empirische Verteilungen

$$P_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{Z_j} = \frac{1}{n}\sum_{j=1}^{n} \delta_{(X_j,-X_j)}, n \in \mathbb{N}$$

Die Testfuntion

$$g := g_\omega := \mathbb{1}_{\{(x,y)\in\mathbb{R}^2 | x+y<0\}} + \mathbb{1}_{\{(X_j(\omega),Y_j(\omega))|j\in\mathbb{N}\}} \colon \mathbb{R}^2 \to [0,1]$$

ist in jeder Koordinate antiton. Es gilt

$$P(g) - P_n(g) = P(g) - \frac{1}{n}\sum_{j=1}^{n} g(Z_j) = 0 - 1.$$

Insbesondere gilt für

$$\mathcal{F}_M := \{f \colon \mathbb{R}^2 \to \mathbb{R}, \text{ antiton in beiden Koordinaten und } \|f\|_\infty < M\}$$

$$\|P - P_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \|P_n(f) - P_X(f)\| = 1 \nrightarrow 0. \tag{3.5}$$

Definition 3.26.

- Eine Teilmenge $G \subset \mathbb{R}^2$ heißt *strikt monotoner Graph*, falls es eine strikt monotone Funktion $g\colon \mathbb{R} \to \mathbb{R}$ gibt mit $G = \{(x, g(x)) \mid x \in \mathbb{R}\}$.
- Sei $\mathcal{M} \subset \mathcal{M}_1(\mathbb{R}^2, \mathcal{B})$ die Menge der Maße P mit

(3.6) $$P_c(G) = 0 \quad \text{für jeden strikt monotonen Graphen } G,$$

wobei P_c den kontinuierlichen Anteil von P bezeichnet, also

$$P_c := P - \sum_{\substack{x \in \mathbb{R}^2, \\ P(\{x\})>0}} \delta_x$$

Satz 3.27 (De Hardt, Wright)**.**

(a) $(\mathcal{F}_M, \mathcal{M})$ *sind ein GC-Paar.*

(b) Gibt es für $P \in \mathcal{M}_1$ *einen strikt monotonen Graphen mit* $P_c(G) > 0$*, dann gilt* $\|P - P_n\|_{\mathcal{F}_M} \not\to 0$.

(a) beschreibt eine hinreichende Bedingung für Konvergenz,
(b) eine notwendige.

Bemerkung 3.28. Satz zeigt, dass im höher dimensionalen Fall Phänomene auftreten, die im eindimensionalen unmöglich sind. Ab Dimension 2 bleibt alles aber im Prinzip gleich. Es gibt analoge Resultate für Dimension > 2.

Vergleiche mit Portemanteau-Theorem:

Satz 3.29 (Portemanteau-Theorem)**.** *Seien* $n \in \mathbb{N}$*,* S *topologischer Raum mit Borel-Sigma-Algebra* $\mathcal{S}$*, und* $\mu, \mu_n, n \in \mathbb{N}$*, Wahrscheinlichkeitsmaße auf* $(S, \mathcal{S})$*. Dann sind äquivalent:*

(i) $\mu_n \xrightarrow{w} \mu$*;*

(ii) für alle gleichmäßig stetigen und beschränkten f *gilt:* $\int f d\mu_n \xrightarrow[n\to\infty]{} \int f d\mu_n$*;*

(iii) für alle Lipschitz-stetigen und beschränkten f *gilt:* $\int f d\mu_n \xrightarrow[n\to\infty]{} \int f d\mu_n$*;*

(iv) für alle messbaren und beschränkten f *mit* $\mu($*Punkte, an denen* f *unstetig ist*$) = 0$ *gilt:* $\int f d\mu_n \xrightarrow[n\to\infty]{} \int f d\mu$*;*

(v) für alle abgeschlossenen Mengen $A \subset S$ *gilt:* $\limsup_{n\to\infty} \mu_n(A) \le \mu(A)$*;*

(vi) für alle offenen Teilmengen $U \subset S$ *gilt:* $\liminf_{n\to\infty} \mu_n(U) \ge \mu(U)$*;*

(vii) für alle $B \in \mathcal{B}(S)$ *mit* $\mu(\partial B) = 0$ *gilt:* $\lim_{n\to\infty} \mu_n(B) = \mu(B)$.

Strikt monoter Graph G spielt die Rolle von ∂B.

3.5. **Weitere Anwendungen von self-bounding.**

Beispiel 3.30 (Bedingte Rademacher-Mittelwerte). Seien $n \in \mathbb{N}$, $X_1, \ldots, X_n$, $\varepsilon_1, \ldots, \varepsilon_n$ unabhängig mit $X_i \in [-1,1]^d$ i.i.d., $X_i = (X_{i1}, \ldots, X_{id})^\top$ und $\varepsilon_i \in \{-1, 1\}$ Rademacher-ZVen.
Folgende Größe wird in der Lerntheorie benutzt, um die Komplexität eines Modells zu beschreiben:

$$Z = E\left(\max_{j=1,\ldots,d} \sum_{i=1}^{n} \varepsilon_i X_{ij} \mid X_1, \ldots, X_n\right). \tag{3.7}$$

Aus dem *Faktorisierungslemma* folgt, dass es Abbildung $f\colon ([-1,1]^d)^n \to \mathbb{R}$ gibt, so dass

$$Z = f(X_1, \ldots, X_n) \quad \text{ist.} \tag{3.8}$$

Satz 3.31 (Faktorisierungslemma)**.**
Seien $(S, \mathcal{S})$ Messraum, $\Omega \neq \emptyset$ und $f\colon \Omega \to S$, $g\colon \Omega \to \mathbb{R}$ Abbildungen.
Dann gilt:
g ist $\sigma(f) - \mathcal{B}$-messbar $\Leftrightarrow$ Es existiert ein messbares $\varphi\colon (S, \mathcal{S}) \to \mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit $g = \varphi \circ f$.

Z hat beschränkte Differenzen:
Ersetzt man X_i durch unabhängige Kopie $X_i' \in [-1,1]^d$, so kann sich $\sum_{i=1}^n \varepsilon_i X_{ij}$ um höchstens 2 ändern (für $j = 1, \ldots, d$). Setze also $c_1 = \ldots = c_j := 2$.
Korollar 3.4 impliziert: $\mathrm{Var}(Z) \leq n$.
Die ZV Z ist aber auch selbst-beschränkend:
Um dies z.z., müssen wir neben f auch $f_1, \ldots, f_n$ identifizieren:

$$Z = f(X_1, \ldots, X_n) \text{ wie oben },$$
$$Z_i = E\left(\max_{j=1,\ldots,d} \sum_{k=1, k\neq i}^{n} \varepsilon_i X_{kj} \mid X^{(i)}\right) = f_i(X^{(i)}).$$

Als Übungsaufgabe zeigt man:

$$0 \leq Z - Z_i \leq 1 \text{ und } \sum_{i=1}^{n} (Z - Z_i) \leq Z$$

Mit Korollar 3.10 folgt $\mathrm{Var}(Z) \leq E(Z)$.
In vielen Anwendungen gilt $E(Z) \leq C\sqrt{n}$ und somit $\mathrm{Var}(Z) \leq C\sqrt{n}$.

Beispiel 3.32 (First passage percolation:). Sei (V, E) ein Graph mit (abzählbarer) Kantenmenge $E = (e_i)_{i\in I}$ und uniform beschränkten Vertexgrad. Betrachte Kantengewichte

$$X_i \colon \Omega \to [0, \infty), \quad E(X_i^2) \le \sigma^2, \quad i \in I \qquad \text{unabhängige ZV}$$

Für einen Pfad γ zwischen Vertizes x und y in V definiere Gewicht/Gesamtlänge als

$$\ell(\gamma) = \sum_{i\in I, e_i\subset\gamma} X_i$$

und den Abstand zwischen x und y als

$$Z := Z(x, y) := \inf\{\ell(\gamma) \mid \gamma \text{ verbindet } x \text{ und } y\} \tag{3.9}$$

Dies ist offensichtlich auch ein ZV. Ersetzt man im obigen Ausdruck X_i durch eine unabhängige Kopie X_i', wid die entsprechende ZV mit Z_i' bezeichnet.

Stillschweigend nehmen wir im folgenden an, dass obige inf tatsächlich min sind. Dies ist z.B. der Fall, falls der Graph endlich viele Kanten enthält. $Z(x, y)$ kann man als die kürzeste Zeit ansehen, die nötig ist, um von x nach y zu kommen oder umgekehrt. Daher der Name *first passage percolation*.
Sei γ^* ein Pfad von x nach y, das das Infimum in (3.9) realisiert. Ersetzen wir das Gewicht X_i der i-ten Kante e_i durch X_i', so ändert sich die Gesamtlänge des Pfades γ^* nur, falls $e_i \subset \gamma$, also

$$(Z_i - Z_i')_- = (Z_i' - Z_i)_+ \le (X_i' - X_i)_+$$

sowie

$$\big((Z_i' - Z_i)_+\big)^2 \le \big((X_i' - X_i)_+\big)^2 \mathbb{1}_{\{e_i\subset\gamma^*\}} \le \big(X_i'\big)^2 \mathbb{1}_{\{e_i\subset\gamma^*\}}$$

Daher folgt unter Nutzung der Unabhängigkeit von X_i, X_i' aus der E-S-U:

$$\begin{aligned}
\operatorname{Var} Z &\le \sum_{i\in I} E\left[\big((Z_i - Z_i')_-\big)^2\right] \\
&\le \sum_{i\in I} E\left[\big(X_i'\big)^2 \mathbb{1}_{\{e_i\subset\gamma^*\}}\right] \\
&\le E\left[(X_i')^2\right] E\left[\sum_{i\in I} \mathbb{1}_{\{e_i\subset\gamma^*\}}\right] \\
&\le \sigma^2 E\left[\sharp\text{Kanten im optimalen Pfad } \gamma^*\right]
\end{aligned}$$

Übung: Für den Graphen $\mathbb{Z}^d$ und iid Gewichte X_i mit Werten in $[a,b]$, wobei $0 < a < b < \infty$, gilt:

$$\operatorname{Var} Z(x,y) \leq \frac{b}{a}\|y-x\|_1 := \frac{b}{a}\sum_{j=1}^{d}|y_j - x_j|$$

Dies ist eine deterministische Schranke, die nur von der Verteilung der ZV und dem ℓ^1-Abstand der betrachteten Punkte abhängt.
Tatsächlich wird $\operatorname{Var} Z(0, n\cdot e_1) \sim n^{2/3}$ vermutet.

Beispiel 3.33 (Größter Eigenwert einer zufälligen symmetrischen Matrix). Sei $A := (X_{i,j})_{i,j}$ symmetrische Matrix mit Koeffizienten

$$\begin{aligned} X_{i,j},\ (1 \leq i \leq j \leq n) \quad & \text{unabh. ZV,} \\ X_{i,j} \in [a,b],\ (1 \leq i \leq j \leq n) \quad & \text{und} \\ X_{i,j} = X_{j,i},\ (1 \leq i \leq n, 1 \leq j \leq n). & \end{aligned}$$

Damit sind alle Eigenwerte von A reell. Sei $Z := \lambda_{\max}(A)$ der maximale Eigenwert von A und $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ ein zugehöriger normierter Eigenvektor. Dann gilt

$$\lambda_{\max} = v^T A v = \sup_{u\in\mathbb{R}^n, \|u\|_2=1} u^T A u = \max_{u\in\mathbb{R}^n, \|u\|_2=1} u^T A u$$

Ersetzt man den Eintrag $X_{i,j}$ durch unabhängig Kopie $X'_{i,j}$ (und entsprechend $X_{j,i}$) erhält man $A'_{i,j}$ und $Z'_{i,j} = \lambda_{\max}(A'_{i,j})$. Dann ergeben sich

$$Z'_{i,j} = \max_{u\in\mathbb{R}^n, \|u\|_2=1} u^T A'_{i,j} u \geq v^T A'_{i,j} v$$

sowie

$$\begin{aligned} Z - Z'_{i,j} &\leq v^T A v - v^T A'_{i,j} v \\ &\leq 2\left|v_i (X_{i,j} - X'_{i,j}) v_j\right| \\ &\leq 2(b-a)\left|v_i v_j\right| \end{aligned}$$

Um $(Z - Z'_{i,j})_+$ abzuschätzen interessieren wir uns nur für den Fall $0 \leq Z - Z'_{i,j} \leq 2(b-a)|v_i v_j|$. Also folgt

$$\begin{aligned} \sum_{1\leq i\leq j\leq n} \left((Z - Z'_{i,j})_+\right)^2 &\leq 4(b-a)^2 \sum_{1\leq i\leq j\leq n} |v_i v_j|^2 \\ &= 4(b-a)^2 \left(\sum_{1\leq i\leq n} v_i^2\right)^2 \\ &= 4(b-a)^2 \end{aligned}$$

da v normiert. Einsetzen in die E-S-U ergibt

$$\mathrm{Var}(Z) \leq 4(b-a)^2$$

ein Schranke, die nicht von der Matrixgröße abhängt, und von der Verteilung der Einträge nur über den Durchmesser $(b-a)$. Die Einträge müssen unabh., nicht aber notwendig identisch verteilt sein.

3.6. **Eine konvexe Poincaré-Ungleichung.**

Definition 3.34. $f\colon [0,1]^n \to \mathbb{R}$ ist *separat konvex*, falls $\forall\, i = 1,\dots,n$: $\forall\, x_1,\dots,x_{i-1},x_{i+1},\dots,x_n$ die Funktion

$$x_i \mapsto f(x_1,\dots,x_i,\dots,x_n) \text{ konvex ist.}$$

Satz 3.35 (Konvexe Poincaré-Ungleichung)**.**
Seien $X_1,\dots,X_n \in [0,1]$ unabhängige ZVen. Sei $f\colon [0,1]^n \to \mathbb{R}$ separat konvex (und partiell differenzierbar). Dann gilt für $Z = f(X) = f(X_1,\dots,X_n)$

$$\mathrm{Var}(f(X)) \leq E(\|\nabla f(X)\|^2).$$

Prinzip: $\mathcal{L}^2$-Norm von f durch $\mathcal{L}^2$-Norm von ∇f abschätzen.
Beachte: Konvexität impliziert bereits Differenzierbarkeit fast überall. Die Aussage lässt sich abschwächen für schwache Gradienten mit den schwachen partiellen Ableitungen.
Ähnliche Ungleichungen werden in Kapitel 3.8 und 5.3 bewiesen. Diese sind unabhängig von Satz 3.35.

Beweis. Nutze variationelle Version der Darstellung von ν in der Efron-Stein-Ungleichung

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} E\left[(Z - Z_i)^2\right]$$

wobei $Z_i = \inf \tilde{Z}_i$ über alle $(X_1,\dots,X_{i-1},X_{i+1},\dots,X_n)$-messbaren $\mathcal{L}^2$-ZVen $\tilde{Z}_i$ läuft. Insbesondere ist $\tilde{Z}_i = \inf_{x_i\in[0,1]} f(X_1,\dots,x_i,\dots,X_n)$ eine zulässige Wahl. Sei $x' = x_i(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)$ ein Minimierer des Infimums (Existenz klar wegen Stetigkeit und Kompaktum).
Sei $\widetilde{X^{(i)}} = (X_1,\dots,X_{i-1},x',X_{i+1},\dots,X_n)$.

$$\begin{aligned}
&\Rightarrow \sum_{i=1}^{n}(Z-Z_i)^2 = \sum_{i=1}^{n}(f(X) - f(\widetilde{X^{(i)}}))^2 \\
&\overset{\text{konvex}}{\leq} \sum_{i=1}^{n}\left(\frac{\partial f(X)}{\partial x_i}\right)^2 \underbrace{(X-\widetilde{X^{(i)}})^2}_{\leq 1} \leq \sum_{i=1}^{n}\left(\frac{\partial f(X)}{\partial x_i}\right)^2 = \|\nabla f(X)\|^2
\end{aligned}$$

Abbildung 4. Konvexität

□

Beispiel 3.36 (Größter Singulärwert einer zufälligen Matrix). Sei $A \in \mathbb{R}^{m\times n}$ Martix mit unabhängigen zufälligen Koeffizienten $X_{ij} \in [0,1]$. Ähnlich wie bei einer symmetrischen Martix in Beispiel 3.33 wollen wir die Konzentration des größten Singulärwerts Z von A untersuchen. Dieser ist gegeben durch

$$Z = Z(A) = Z((X_{ij})_{ij}) = \sqrt{\lambda_{\max}(A^T A)}$$

Da $A^T A$ symmetrisch ist, ist $\lambda_{\max}(A^T A)$ wohldefiniert und reell. Wieder haben wir eine variationelle Darstellung

$$Z = \sqrt{\sup_{u\in\mathbb{R}^n, \|u\|_2=1} u^T A^T A u} = \sqrt{\max_{u\in\mathbb{R}^n, \|u\|_2=1} \|Au\|_2} =: \|A\|$$

wobei $\|A\|$ *Operator-* oder *Spektralnorm* von A genannt wird.
Für festes $u \in \mathbb{R}^n$ ist

$$[0,1] \ni X_{ij} \mapsto f_{X,u}(X_{ij}) := \|Au\| = \sqrt{\sum_{l=1}^{m} \left(\sum_{k=1}^{n} X_{l,k} u_k \right)^2}$$

eine konvexe Funktion von X_{ij}. Als Maximum konvexer Funktionen ist

$$[0,1] \ni X_{ij} \mapsto \max_{u\in\mathbb{R}^n, \|u\|_2=1} f_{X,u}(X_{ij})$$

ebenfalls konvex. Nun verwenden wir einen Satz aus der Linearen Algebra der die Störungstheorie von Matrizen betrifft.

Satz 3.37 (Satz von Liidiski)**.** *Seien A und $B \in \mathbb{R}^{m\times n}$ mit Singulärwerten*

$$s_{\max}(M) = s_1(M) \geq s_2(M) \geq \ldots \geq s_n(M), \quad M \in \{A, B\}$$

Dann gilt:

$$\begin{aligned}(s_1(B) - s_1(A))^2 &\leq \sum_{i=1}^{n} (s_i(B) - s_i(A))^2 \\ &\leq \sum_{i=1}^{n} s_i \left(B - A\right)^2 \\ &= \mathrm{Tr}\left((B-A)^T (B-A)\right) \\ &= \sum_{l=1}^{m}\sum_{l=1}^{n} (b_{k,l} - a_{k,l})^2\end{aligned}$$

was drei verschidenen Darstellungen der Hilbert-Schmidt-Norm (oder Frobenius-Norm) zum Quadrat sind.

Gilt nun

$$b_{k,l} = a_{k,l} + \varepsilon \delta_{k,i} \delta_{l,j}$$

so folgt

$$\frac{(s_1(B) - s_1(A))^2}{\varepsilon^2} \leq 1$$

und damit ist $s_1(A)$ als Funktion X_{ij} Lipschitz-stetig mit Lipschitz-Konstante gleich Eins. Daraus folgt, dass für Lebesgue-fast alle Werte in $[0,1]$

$$a_{ij} \mapsto s_1(A)$$

differenzierbar ist mit $\left|\frac{\partial}{\partial a_{ij}} s_1(A)\right| \leq 1$. In der Tat reicht diese abgeschwächte Bedingung, um eine konvexe Poincare-Ungleichung zu beweisen (Übung).
Da wir den Gradienten einer Funktion auf dem $\mathbb{R}^{m\cdot n}$ untersuchen müßen, erweitern wir unsere 1-dimensionalen Überlegungen. Satz von Liidski impliziert

$$(s_1(B) - s_1(A)) \leq \sqrt{\sum_{l=1}^{m}\sum_{l=1}^{n} (b_{k,l} - a_{k,l})^2} = \|B - A\|_{\mathbb{R}^{m\cdot n}}$$

wobei im letzten Ausdruck die Euklidische Norm auf $\mathbb{R}^{m\cdot n}$ sowie A und B als Vektoren in $\mathbb{R}^{m\cdot n}$ betrachtet werden. Mit anderen Worten ist

$$\mathbb{R}^{m\cdot n} \ni A \mapsto s_1(A)$$

Lipschitz-stetig mit Lipschitz-Konstante gleich Eins, woraus man (als Übung) folgert, das (Lebesgue fast überall)

$$\|\nabla f(X)\| = \left\| \left(\frac{\partial f}{\partial x_{ij}}(X) \right)_{ij} \right\| \leq 1$$

Somit folgt aus der konvexen Poincare-Ungleichung

$$\mathrm{Var}(Z) = \mathrm{Var}(f(X)) \leq E\left(\|\nabla f(X)\|^2\right) \leq E(1) = 1$$

Dies ist wieder von den Dimensionen n und m unabhängig!

3.7. **Anwendung der Efron-Stein-Ungleichung auf Tail-Events.** Bei der allgemeinen Markov-Ungleichung kann man statt der quadratischen — sofern exponentielle Momente existieren — auch exponentielle Funktionen einsetzen, um schärfere Schranken an das Abfallverhalten zu bekommen. In diesem Sinne wollen wir E-S-U verbessern! Hier nehmen wir für die Funktion $f\colon \mathcal{X}^n \to \mathbb{R}$ etwas weniger als die beschränkte Differenz-Bedingung an: Es existiere $\nu > 0$ mit

$$\sum_{i=1}^{n}((Z - Z_i')_+)^2 \leq \nu \text{ f.s., wobei wieder} \tag{3.10}$$
$$Z_i' = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n).$$

wobei X_i' eine unabh. Kopie von X_i ist. Es liegt also eine Art *kummulative* oder *gemittelte* beschränkte Differenz-Bedingung vor.
Beispiel: $\lambda_{\max}$ ist EW von symm. Matrix $\Rightarrow$ (3.10) gilt mit $\nu = 4(b-a)^2$
Insbesondere ist die Konstante unabhämgig von der Konfiguration $x \in \mathcal{X}^n$ und sogar von der Dimension n. Dann folgt mit E-S-U:

$$\mathrm{Var}(Z) \leq E\left(\sum_{i=1}^{n}((Z - Z_i')_+)^2\right) \leq \nu.$$

Für beliebige $\alpha \in (0,1)$ sei Q_α das α-Quantil von $Z = f(X) = f(X_1, \ldots, X_n)$, d.h.:

$$Q_\alpha := \inf\{t \in \mathbb{R} \mid P(Z \leq t) \geq \alpha\} = \inf\{t \in \mathbb{R} \mid F_Z(t) \geq \alpha\}.$$

Insbesondere ist $\mathcal{M}(Z) = Q_{\frac{1}{2}}$ der Median. Zu gegebener Funktion $f\colon \mathcal{X}^n \to \mathbb{R}$ betrachte Modifizierung $g = g_{a,b} : \mathcal{X}^n \to \mathbb{R}$ für $a < b \in \mathbb{R}$:

$$g(x) = \begin{cases} b & f(x) \geq b \\ f(x) & f(x) \in (a,b) \\ a & f(x) \leq a \end{cases}$$

Betrachte den Fall $\mathcal{M}(Z) \le a$, dann gilt: $p := P(Z \le a) \ge \frac{1}{2}$. Also ist

$$\begin{aligned} E(g(X)) &= E(g(X)\mathbb{1}_{\{f(X)\le a\}}) + E(g(X)\mathbb{1}_{\{f(X)>a\}}) \\ &\le aE(\mathbb{1}_{\{f(X)\le a\}}) + bE(\mathbb{1}_{\{f(X)>a\}}) \\ &= aP(Z \le a) + bP(Z > a) \\ &= ap + b(1-p) = b + p\underbrace{(a-b)}_{\text{negativ}} \\ &\le b + \frac{1}{2}(a-b) = \frac{a+b}{2}. \end{aligned}$$

Dann folgt für $Y = g(X)$:

$$\begin{aligned} \mathrm{Var}(Y) &= E((Y - E(Y))^2) \ge E\left((Y-E(Y))^2\mathbb{1}_{\{Y\ge b\}}\right) \\ &\ge (b - E(Y))^2 E(\mathbb{1}_{\{Y\ge b\}}) \ge \left(b - \frac{a+b}{2}\right)^2 P(Y \ge b) \\ &= \frac{(b-a)^2}{4} P(Z \ge b). \end{aligned}$$

Komplementäre Abschätzung der $\mathrm{Var}(Y)$ von oben mit E-S-U. Zu einem Punkt $x \in \mathcal{X}^n$ setze

$$\tilde{x}^{(i)} = (x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n).$$

Wir nutzen

$$f(x) \le a \quad \Rightarrow \quad a = g(x) \le g(\tilde{x}^{(i)}). \tag{3.11}$$

in der folgenden Ungleichungskette

$$\begin{aligned} \mathrm{Var}(g(X)) \le \sum_{i=1}^n E\left((g(X) - g(\tilde{X}^{(i)}))^2\right) &\overset{\text{Symmetrie}}{=} 2\sum_{i=1}^n E\left((g(X) - g(\tilde{X}^{(i)})_+)^2\right) \\ &\overset{(3.11)}{=} 2E\left(\sum_{i=1}^n \mathbb{1}_{\{f(X)>a\}}(g(X) - g(\tilde{X}^{(i)})_+)^2\right) \\ &\le 2\nu P(Z > a). \end{aligned}$$

wegen der Annahme (3.10). Kombiniere untere und obere Schranke:

$$\begin{aligned} &\frac{(b-a)^2}{4} P(Z \ge b) \le \mathrm{Var}(Y) \le 2\nu P(Z > a) \\ \Rightarrow\ & b - a \le \sqrt{8\nu \frac{P(Z>a)}{P(Z\ge b)}} \end{aligned}$$

Ziel: Schätze Abstände zwischen Quantilen ab:
Seien $0 \leq \delta \leq \gamma \leq \frac{1}{2}$. Setze:

$$a = Q_{1-\gamma}, b = Q_{1-\delta} \quad \text{insbesondere} \quad P(Z > a) \leq \gamma, P(Z > b) \leq \delta.$$

Aus dem gezeigten folgt für die Abstände zwischen Quantilen rechts von $\mathcal{M}(Z)$:

$$Q_{1-\delta} - Q_{1-\gamma} \leq \sqrt{\frac{8\nu\gamma}{\delta}}.$$

Um das *tail*-Verhalten zu verstehen, ist es sinnvoll $\gamma = 2^{-k} > \delta = 2^{-(k+1)}$ für $k \in \mathbb{N}$ zu setzen. Dann:

$$Q_{1-2^{-k-1}} - Q_{1-2^{-k}} \leq \sqrt{\frac{8\nu}{1/2}} = 4\sqrt{\nu}$$

D.h. die Abstände zwischen aufeinander folgenden Quantilen von exponentiell fallenden Wahrscheinlichkeiten sind beschränkt. Insbesondere:

$$Q_{1-2^{-m-1}} - Q_{1/2} = \sum_{k=1}^{m} (Q_{1-2^{-k-1}} - Q_{1-2^{-k}}) \leq 4m\sqrt{\nu},$$

also $Q_{1-2^{-m-1}} \leq \mathcal{M}(Z) + 4m\sqrt{\nu}$ und somit:

$$\forall\, m \in \mathbb{N}: \quad P(Z > \mathcal{M}(Z) + 4m\sqrt{\nu}) \leq P(Z > Q_{1-2^{-m-1}}) \leq 2^{-m-1}$$

Umformung ergibt:

$$\forall\, t \geq 0 : P(Z \geq \mathcal{M}(Z) + t) \leq 2^{-\frac{t}{4\sqrt{\nu}}}.$$

Optimal wäre sub-gaußscher Abfall $P(\dots) \leq 2^{-\frac{t^2}{\nu}}$. Ist erreichbar mit anderen Methoden.

Nun: zweite Methode um aus E-S-U Abschätzungen für Wahrscheinlichkeiten von Tail-Events zu erhalten: Führe wie bei Anwendung der Markov-Ungleichung Substitution mit exponentielle Funktion durch.
Für $\lambda > 0$ betrachte ZVe $Y := e^{\frac{\lambda Z}{2}}$.

$$\begin{aligned} \mathrm{Var}(Y) &= E(e^{\lambda Z}) - E(e^{\frac{\lambda Z}{2}})^2 \\ &\overset{\text{E-S-U}}{\leq} E\left(\sum_{i=1}^{n} \left(e^{\frac{\lambda Z}{2}} - e^{\frac{\lambda Z_i'}{2}}\right)_+^2\right) \leq (*) \end{aligned}$$

wobei Z_i' unabhängige, identische Kopien von Z_i sind. Anwendung des Mittelwertsatzes liefert

$$e^{\frac{\lambda z}{2}} - e^{\frac{\lambda w}{2}} \leq \frac{\lambda}{2} e^{\frac{\lambda \xi}{2}} (z - w)$$

für einen Wert ξ zwischen w und z. Da $x \mapsto \exp(x)$ isoton ist, gilt:

$$\begin{aligned} & (e^{\frac{\lambda z}{2}} - e^{\frac{\lambda w}{2}})_+ = \frac{\lambda}{2} e^{\frac{\lambda \xi}{2}} (z-w)_+ \leq \frac{\lambda}{2} e^{\frac{\lambda z}{2}} (z-w)_+ \\ \Rightarrow \quad & (e^{\frac{\lambda z}{2}} - e^{\frac{\lambda w}{2}})_+^2 \leq \frac{\lambda^2}{4} e^{\lambda z} (z-w)_+^2 \end{aligned}$$

Daraus ergibt sich

$$(*) \leq \frac{\lambda^2}{4} E\Big(e^{\lambda Z} \underbrace{\sum_{i=1}^{n} (Z - Z_i')_+^2}_{\leq \nu}\Big) \leq \frac{\nu \lambda^2}{4} E\left(e^{\lambda Z}\right)$$

wobei wir auch hier die globale Annahme (3.10) verwendet haben. Insgesamt liefert E-S-U

$$\begin{aligned} & \mathrm{Var}(Y) = E\left(e^{\lambda Z}\right) - E\left(e^{\frac{\lambda Z}{2}}\right)^2 \leq \frac{\nu \lambda^2}{4} E\left(e^{\lambda Z}\right) \\ \Leftrightarrow \quad & \left(1 - \frac{\nu \lambda^2}{4}\right) E\left(e^{\lambda Z}\right) \leq \left(E(e^{\frac{\lambda Z}{2}})\right)^2 \\ \Leftrightarrow \quad & \left(1 - \frac{\nu \lambda^2}{4}\right) E\left(e^{\lambda Z} e^{-\lambda E(Z)}\right) \leq \left(E(e^{\frac{\lambda Z}{2}})\right)^2 e^{-\lambda E(Z)} \\ \Leftrightarrow \quad & \left(1 - \frac{\nu \lambda^2}{4}\right) E\left(e^{\lambda (Z - E(Z))}\right) \leq \left(E\left(e^{\frac{\lambda}{2}(Z - E(Z))}\right)\right)^2 \\ \Leftrightarrow \quad & \left(1 - \frac{\nu \lambda^2}{4}\right) M(\lambda) \leq M\left(\frac{\lambda}{2}\right)^2, \end{aligned}$$

wobei F die MEF von $Z - E(Z)$ ist. Diese Ungleichung für die MEF lässt sich mittels folgenden Lemmas aus der Analysis ausnutzen:

Lemma 3.38. *(Lemma aus der Analysis)*
$g : (0,1) \to (0,\infty)$ erfülle folgende (sanfte) Regularitätsannahme

$$\lim_{x \to 0} \frac{g(x) - 1}{x} = 0.$$

Dann impliziert

$$g(x)(1-x^2) \leq g\left(\frac{x}{2}\right)^2 \text{ für alle } x \in (0,1),$$

$$\text{dass } g(x) \leq \left(\frac{1}{1-x^2}\right)^2 \text{ gilt.}$$

Beweis.

$$\forall\, x \in (0,1) : g(x)(1-x^2) \leq g\left(\frac{x}{2}\right)^2 \Leftrightarrow \quad \forall\, x \in (0,1) : g(x) \leq \frac{1}{1-x^2} g\left(\frac{x}{2}\right)^2.$$

Wegen $x \in (0,1) \Rightarrow \frac{x}{2^n} \in (0,1)$ kann man dies rekursiv anwenden

$$\forall\, k \in \mathbb{N} : g(x) \leq (g(x2^{-k}))^{2^k} \prod_{j=0}^{k-1} \left(1 - \left(x2^{-j}\right)^2\right)^{-2^j}$$

Für $k \to \infty$ folgt:

$$g(x) \leq \lim_{k\to\infty} (g(x2^{-k}))^{2^k} \prod_{j=0}^{k-1} \left(1 - \left(x2^{-j}\right)^2\right)^{-2^j}$$

Betrachte Limes für ersten Faktor separat.
Regularitätsannahme $g(x) = 1 + o(x)$ impliziert

$$\ln(g(x2^{-k})^{2^k}) = 2^k \ln(g(x2^{-k})) = 2^k \ln(1 + o(x2^{-k})) \leq 2^k o(x2^{-k}) \xrightarrow[k\to\infty]{} 0,$$

wobei in der letzten Ungleichung $\ln(1+x) \leq x$ eingeht. Es folgt $g(x2^{-k})^{2^k} \leq \exp(2^k o(x2^{-k})) \xrightarrow[k\to\infty]{} 1$. Damit folgt

(3.12)

$$\ln(g(x)) \leq \lim_{k\to\infty} \sum_{j=0}^{k-1} 2^j(-\ln(1-(x2^{-j})^2)) = \sum_{j=0}^{\infty} 2^j(-\ln(1-(x2^{-j})^2)).$$

Da $x \mapsto \ln(x)$ konkav, ist $-x \mapsto -\ln(x)$ konvex. Also:

$$u \mapsto -\frac{1}{u}\ln(1-u) \text{ isoton für } u \in (0,1)$$
$$\Rightarrow -\frac{1}{x^2 2^{-2j}}\ln(1-x^2 2^{-2j}) \leq -\frac{1}{x^2}\ln(1-x^2), \text{ da } x^2 \geq x^2 2^{-2j}.$$

Umformen liefert:

$$-\ln(1-(x2^{-j})^2) \leq 2^{-2j}(-\ln(1-x^2))$$

Einsetzen in (3.12) ergibt:

$$\ln(g(x)) \leq \sum_{j=0}^{\infty} 2^j 2^{-2j}(-\ln(1-x^2)) = \left(\sum_{j=0}^{\infty} 2^{-j}\right)(-\ln(1-x^2)) = -2\ln(1-x^2).$$

Exponentialfunktion anwenden, liefert

$$g(x) \leq \exp(-2\ln(1-x^2)) = \exp\left(\ln\left((1-x^2)^{-2}\right)\right)$$

die Behauptung. □

Wende das Lemma auf $M(\lambda) = E(e^{-\lambda(Z-E(Z))})$ an.

$$M(0) = 1, M'(\lambda) = E\left((Z-E(Z))e^{\lambda(Z-E(Z))}\right) \overset{\text{zentriert}}{\Rightarrow} M'(0) = 0,$$

also ist $M(\lambda) = 1 + o(\lambda)$ für $\lambda \to 0$. Setzen wir $g(x) = M\left(\frac{2x}{\sqrt{\nu}}\right)$, dann gilt für alle $\lambda \in \left(0, \frac{2}{\sqrt{\nu}}\right)$.

$$M(\lambda) = g\left(\lambda\frac{\sqrt{\nu}}{2}\right) \leq \left(\frac{1}{1-\left(\frac{\lambda\sqrt{\nu}}{2}\right)^2}\right)^2 = \left(1-\frac{\lambda^2\nu}{4}\right)^{-2}. \tag{3.13}$$

Insbesondere sehen wir, dass die E-S-U unter der Annahme (3.10) gewährleistet, dass Z exponentiell integrierbar ist. Weiterhin erhalten wir eine Abschätzung an W'keiten, dass Z große Werte annimmt:

$$\forall t > 0: \; P(Z - E(Z) \geq t) \overset{\text{Markov}}{\leq} E(e^{(Z-E(Z))\nu^{-\frac{1}{2}}})e^{-t\nu^{-\frac{1}{2}}}$$
$$\leq e^{-\frac{t}{\sqrt{\nu}}} M\left(\frac{1}{\sqrt{\nu}}\right) \leq e^{-\frac{t}{\sqrt{\nu}}} \underbrace{\left(1-\frac{1}{4}\right)^{-2}}_{\leq 2} \leq 2e^{-\frac{t}{\sqrt{\nu}}}.$$

Die Schranke hat die gleiche Struktur wie beim ersten Anwendung der EFU auf *tail*-Abschätzungen. Aber:

- hier haben wir das Ereignis: Überschreitung des Erwartungswerts
- dort haben wir das Ereignis: Überschreitung des Medians!

Nutze (3.13) etwas anders mit Hilfe von $-\ln(1-u) \leq u(1-u)^{-1}$:

$$\Rightarrow \forall \lambda \in \left(0, 2\nu^{-\frac{1}{2}}\right) : \ln(M(\lambda)) \leq \ln\left(1-\frac{\lambda^2\nu}{4}\right)(-2) \leq 2\frac{\lambda^2\nu}{4}\left(1-\frac{\lambda^2\nu}{4}\right)^{-1}$$
$$\Rightarrow M(\lambda) \leq e^{\frac{\lambda^2\nu}{2}\left(1-\frac{\lambda^2\nu}{4}\right)^{-1}}.$$

Also: $Z - E(Z)$ ist eine sub-Γ ZV (vgl. Kapitel 2.4) mit Varianzfaktor ν und Skalenparameter $c = \frac{\sqrt{\nu}}{2}$.
Aus Kapitel 2.4 folgt wieder eine Tail-Abschätzung:

$$P(Z - E(Z) \geq \sqrt{2\nu t} + ct) \leq e^{-t} \text{ für } t > 0. \tag{3.14}$$

Entscheidend: Wie groß ist der Wert c? Bei uns $c = \frac{\sqrt{\nu}}{2}$. Für nicht zu kleine t dominiert $ct = \frac{\sqrt{\nu}}{2}t$ den Term $\sqrt{2\nu t}$. Also haben wir dort nur exponentielles Verhalten und (3.14) ist keine sub-gaußsche Abschätzung.

3.8. **Gaußsche Poincaré-Ungleichung.**

Grundprinzipien

- Annahmen an f abschwächen
- Annahmen an ZVe dafür stärker

Satz 3.39. *(Poincaré-Ungleichung)*
Sei $X = (X_1, \dots, X_d) \sim \mathcal{N}(0, I_d)$, *also sind* $X_1, \dots, X_d$ *unabhängig, identisch verteilt. Sei* $f \colon \mathbb{R}^d \to \mathbb{R}$ *mit* $f \in \mathcal{C}_0^2(\mathbb{R}^d)$[3].

$$\textit{Dann: } \operatorname{Var}(f(X)) \leq E\left(\|\nabla f(X)\|^2\right).$$

Satz 3.39 wird in Kapitel 5.3 in Theorem 5.4 allgemeiner bewiesen.

Beweis. O.E. $E\left(\|\nabla f(X)\|^2\right) < \infty$. Zuerst 1-dimensional; höhere Dimensionen liefert E-S-U. Außerdem betrachte erst $\varepsilon_1, \dots, \varepsilon_n$ unabhängig, identisch Rademacher-verteilte ZVen, also $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = \frac{1}{2}$. Dann gilt nach ZGS

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \xrightarrow{\mathcal{D}} X \sim \mathcal{N}(0, \nu).$$

Also $\forall\, g \in \mathcal{C}_b(\mathbb{R})$:

$$\int g(t) P_{S_n}(dt) \overset{n\to\infty}{\to} \int g(t) dP_X(dt),$$

insbesondere $\operatorname{Var}(f(S_n)) \to \operatorname{Var}(f(X)) = \operatorname{Var}(Z)$, falls man $g = f^2$ waehlt.
Idee: Leite Schranke für linke Seite her und betrachte den Limes. Für jedes i betrachte Varianz bezüglich ε_i:

$$\operatorname{Var}^{(\varepsilon_i)}(f(S_n)) = \frac{1}{4} \left(\underbrace{f\left(S_n + \frac{1-\varepsilon_i}{\sqrt{n}}\right) - f\left(S_n - \frac{1+\varepsilon_i}{\sqrt{n}}\right)}_{=: D_{n,i}} \right)^2.$$

Die E-S-U liefert mit $Z = \tilde{f}(\varepsilon_1, \dots, \varepsilon_n) = f(S_n)$:

$$\operatorname{Var}(f(S_n)) \leq \frac{1}{4} \sum_{i=1}^{n} E(D_{n,i}^2).$$

Setze $k := \sup_{t \in \mathbb{R}} |f''(t)| < \infty$ nach Annahme $f \in \mathcal{C}_0^2$. Taylorentwicklung ergibt

$$\begin{aligned} |D_{n,i}| &\leq \frac{2}{\sqrt{n}} |f'(S_n)| + \frac{1}{2} \left(\frac{2}{\sqrt{n}}\right)^2 k, \text{ also} \\ \frac{1}{4} \sum_{i=1}^{n} D_{n,i}^2 &\leq \frac{1}{4} \sum_{i=1}^{n} \left(\frac{4}{n} f'(S_n)^2 + \frac{4k^2}{n^2} + \frac{8k}{n\sqrt{n}} |f'(S_n)| \right) \\ &\leq f'(S_n)^2 + \frac{k^2}{n} + \frac{2k}{\sqrt{n}} \underbrace{|f'(S_n)|}_{\leq \|f'\|_\infty} \end{aligned}$$

[3] Insbesondere impliziert die Annahme, dass supp f kompakt ist.

Mit dem ZGS folgt

$$\limsup_{n\to\infty} \frac{1}{4}\sum_{i=1}^{n} E(D_{n,i}^2) \le \limsup_{n\to\infty} E(f'(S_n)^2) + \limsup_{n\to\infty} E\left(\frac{k^2}{n} + \frac{4k}{\sqrt{n}}\|f'\|_\infty\right)$$
$$\overset{\text{ZGS}}{=} E(f'(X)^2).$$

Insgesamt: $\operatorname{Var}(f(X)) = \lim_{n\to\infty} \operatorname{Var}(f(S_n)) \le \lim_{n\to\infty} \frac{1}{4}\sum_{i=1}^{n} E(D_{n,i}^2) \le E(f'(X)^2)$.
Nun: $f\colon \mathbb{R}^d \to \mathbb{R}$ und $X = (X_1, \dots, X_d) \sim \mathcal{N}(0, I_d)$.
E-S-U liefert für $Z = f(X_1, \dots, X_n)$:

$$\begin{aligned}\operatorname{Var}(f(X)) &\le \sum_{i=1}^{d} E\left((Z - E^{(i)}(Z))^2\right) = \sum_{i=1}^{d} \operatorname{Var}^{(i)}(f(X)) \\ &\overset{\text{Beweisteil 1}}{\le} \sum_{i=1}^{d} E\left(\left(\frac{\partial f}{\partial x_i}(X_1, \dots, X_d)\right)^2\right) = E\left(\sum_{i=1}^{d}\left(\frac{\partial f}{\partial x_i}(X_1, \dots, X_d)\right)^2\right) \\ &= E\left(\|\nabla f(X)\|^2\right)\end{aligned}$$

□

3.9. **Beweis für die ES-Ungleichung mittels Dualität.**
Wir nutzen nun die Sichtweise der Dualität, um einen alternativen Beweis der E-S-U zu entwickeln. Diese Beweismethode wird später im allgemeineren Kontext angewendet.

Für $Y, T \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ gilt

$$0 \le \operatorname{Var}(Y - T) = \operatorname{Var}(Y) - 2\operatorname{Cov}(Y, T) + \operatorname{Var}(T)$$

und umgestellt

$$\operatorname{Var}(Y) \ge 2\operatorname{Cov}(Y, T) - \operatorname{Var}(T)$$

Diese Ungleichung kann nicht verschärft werden, denn für $T = Y$ gilt:

$$\operatorname{Var}(Y) = 2\operatorname{Cov}(Y, T) - \operatorname{Var}(T)$$

Also haben wir bewiesen:

Proposition 3.40. *Für jedes $Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ gilt:*

$$\operatorname{Var}(Y) = \max_{T \in \mathcal{L}^2(\Omega,\mathcal{A},P)} [2\operatorname{Cov}(Y, T) - \operatorname{Var}(T)] \tag{3.15}$$

Werden dies i. F. nutzen um Varianz von $Z = f(X_1, \ldots, X_n)$ mit unabhängigen $X_1, \ldots, X_n$ abzuschätzen. Verwenden zunächst wieder Teleskopsumme

$$Z^2 - (EZ)^2 = \sum_{i=1}^{n} \left[(E_i Z)^2 - (E_{i-1} Z)^2 \right], \quad \text{wobei}$$

Bemerkung 3.41 (Erinnerung).

$$\begin{aligned} E_i(Z) &= \int_{\mathcal{X}^{n-i}} f(X_1, \ldots, X_i, \xi_{i+1}, \ldots, \xi_n)\, P_{X_{i+1}}(\xi_{i+1}), \ldots, P_{X_n}(\xi_n) \\ E^{(i)}(Z) &= \int_{\mathcal{X}} f(X_1, \ldots, X_{i-1}, \xi_i, X_{i+1}, \ldots, X_n)\, P_{X_i}(\xi_i) \\ \Delta_i &= E_i(Z) - E_{i-1}(Z) \end{aligned}$$

Damit folgt

$$\mathrm{Var}(Z) = E\left[Z^2 - (EZ)^2\right] = \sum_{i=1}^{n} E\left[(E_i Z)^2 - (E_{i-1} Z)^2\right]$$

Erst im folgenden Schritt nutzen wir Orthogonalität. Da $E_{i-1}(Z) \perp \Delta_i$, folgt aus Pythagoras

$$\begin{aligned} E\left[(E_i Z)^2\right] &= E\left[(E_{i-1} Z)^2\right] + E\left[\Delta_i^2\right] \\ \Rightarrow E\left[\Delta_i^2\right] &= E\left[(E_i Z)^2\right] - E\left[(E_{i-1} Z)^2\right] \end{aligned}$$

Wie schon vorher bemerkt, gilt für unabh. ZV $X_1, \ldots, X_n$:

$$\forall\, i \in \{2, \ldots, n\} :\ E_{i-1}(Z) = E^{(i)}(E_i Z)$$

also auch

$$\begin{aligned} E\left[(E_i Z)^2 - (E_{i-1} Z)^2\right] &= E\left[(E_i Z)^2 - (E^{(i)}(E_i Z))^2\right] \\ = E\left[E^{(i)}\left[(E_i Z)^2 - (E^{(i)}(E_i Z))^2\right]\right] &= E\left[\mathrm{Var}^{(i)}\left(E_i Z\right)\right] \end{aligned}$$

Also haben wir (diesmal ohne Nutzung der Struktureigenschaft (3.1) der Martingaldifferenzen) bewiesen:

$$\mathrm{Var}(Z) = \sum_{i=1}^{n} E\left[\mathrm{Var}^{(i)}\left(E_i Z\right)\right]$$

Falls wir es nun schaffen, den E_i Operator an $\mathrm{Var}^{(i)}$ vorbeizuziehen, verschwindet er wegen der Turmeigenschaft. Mit einer Ungleichheit ist dies wegen Lemma 3.42 in der Tat öglich und wir erhalten

$$\sum_{i=1}^{n} E\left[\mathrm{Var}^{(i)}\left(E_i Z\right)\right] \leq \sum_{i=1}^{n} E\left[\mathrm{Var}^{(i)}\left(Z\right)\right]$$

also die E-S-U.

Lemma 3.42. *Ist $Z = f(X_1, \ldots, X_n)$ quadratintegrierbar mit unabhängigen $X_1, \ldots, X_n$, so gilt*

$$\forall\, i \in \{1, \ldots, n\}: \quad E\left[\mathrm{Var}^{(i)}\left(E_i Z\right)\right] \leq E\left[\mathrm{Var}^{(i)}\left(Z\right)\right]$$

Beweis. Setze $\mathrm{Cov}^{(i)}(Z,T) = E^{(i)}\left[(Z - E^{(i)}(Z))(T - E^{(i)}(T))\right]$. Dann folgt wie bei obiger Proposition

$$\text{(3.16)}\quad \mathrm{Var}^{(i)}(Y) \geq 2\,\mathrm{Cov}^{(i)}(Y,T) - \mathrm{Var}^{(i)}(T) \qquad \text{für bel. } Y, T \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$$

Andererseits gilt für jede $\sigma(X_1, \ldots, X_i)$-messbare ZV $T \in \mathcal{L}^2$:

$$\begin{aligned}
E\left[\mathrm{Cov}^{(i)}(Z,T)\right] &= E\left[E^{(i)}\left[(Z - E^{(i)}(Z))(T - E^{(i)}(T))\right]\right] \\
\text{Turm-E.} \quad &= E\left[(Z - E^{(i)}(Z))(T - E^{(i)}(T))\right] \\
\text{Turm-E.} \quad &= E\left[E_i\left[(Z - E^{(i)}(Z))(T - E^{(i)}(T))\right]\right] \\
T \text{ messbar} \quad &= E\left[(T - E^{(i)}(T)) E_i\left[(Z - E^{(i)}(Z))\right]\right] \\
E \text{ linear} \quad &= E\left[(T - E^{(i)}(T))\left[E_i(Z) - E_i(E^{(i)}(Z))\right]\right] \\
\text{unabh.} \quad &= E\left[(T - E^{(i)}(T))\left[E_i(Z) - E^{(i)}(E_i(Z))\right]\right] \\
\text{Turm-E.} \quad &= E\left[E^{(i)}\left[(T - E^{(i)}(T))(E_i(Z) - E^{(i)}(E_i(Z)))\right]\right] \\
&= E\left[\mathrm{Cov}^{(i)}\left(E_i(Z), T\right)\right]
\end{aligned}$$

insbesondere für $T = E_i(Z) \in \mathcal{L}^2(\sigma(X_1, \ldots, X_i), P)$ mit (3.16)

$$\begin{aligned}
E\left[\mathrm{Cov}^{(i)}(Z, E_i(Z))\right] &= E\left[\mathrm{Cov}^{(i)}\left(E_i(Z), E_i(Z)\right)\right] = E\left[\mathrm{Var}^{(i)}\left(E_i(Z)\right)\right] \\
&\geq E\left[2\,\mathrm{Cov}^{(i)}\left(E_i(Z), E_i(Z)\right) - \mathrm{Var}^{(i)}\left(E_i(Z)\right)\right] \\
&= E\left[\mathrm{Var}^{(i)}\left(E_i(Z)\right)\right]
\end{aligned}$$

□

4. Der Entropiebegriff und Informationsungleichungen

Der Begriff der Entropie ist zwar mathematisch schnell definiert, aber inhaltlich schwer zu fassen. Er mißt, wie viel Unsicherheit oder Zufall in einem System enthalten sind. Wir werden sehen, dass eine deterministische Verteilung, also $P = \delta_x$, minimale Entropie, und eine Gleichverteilung (wenn sie

auf dem vorliegedem Messraum definiert werden kann) maximale Entropie aufweist.

4.1. **Shannon-Entropie und relative Entropie.**
Zunächst einfaches Setting:

- $X\colon \Omega \to \mathcal{X}$ ZVe.
- $\mathcal{X}$ ist abzählbar (diskret).
- Gewichtsfunktion $p(x) = P(X = x)$, $x \in \mathcal{X}$.

Definition 4.1. Die *Shannon-Entropie* der ZVe X (bzw. Verteilung P_X):

$$H(X) := E(-\ln(p(X))) = -\sum_{x\in\mathcal{X}, p(x)>0} p(x)\ln(p(x)) \geq 0$$

Im Spezialfall $X \equiv x_0 \in \mathcal{X}$ wird der Wert Null angenommen:

$$H(X) = -\sum_{x\in\mathcal{X}, \delta_{x_0}(x)>0} \ln(\delta_{x_0}(x)) = -\ln(1) = 0$$

Sei $q\colon \mathcal{X} \to [0,1]$, $\sum_{x\in\mathcal{X}} q(x) = 1$ eine weitere Gewichtsfunktion zur Verteilung einer ZVe Y. Sei Z eine ZVe mit Gewichtsfunktion $\frac{1}{2}(p+q)$. Dann gilt:

$$H(Z) \geq \frac{1}{2}H(X) + \frac{1}{2}H(Y).$$

Das Mischen von Gewichtsfunktionen vergrößert also die Entropie, z.B.

$$H\left(\frac{1}{2}(\delta_y + \delta_x)\right) \geq \frac{1}{2}H(\delta_y) + \frac{1}{2}H(\delta_x).$$

was die Beschreibung vom Anfang des Kapitels bestätigt.
Allgemeiner gilt für jedes $t \in (0,1)$ und für eine ZVe Z mit Gewichtsfunktion $tp + (1-t)q$:

$$H(Z) \geq tH(X) + (1-t)H(Y).$$

Beweis. Die Abbildung $(0,\infty) \in x \mapsto -x\ln(x)$ ist konkav.

Für beliebiges $t \in (0,1)$ und $x \in \mathcal{X}$ gilt:

$$-\sum_{x\in\mathcal{X}} \Big(\big(tp(x) + (1-t)q(x)\big)\cdot \ln\big(tp(x) + (1-t)q(x)\big)\Big)$$

$$\geq -\left(t\sum_{x\in\mathcal{X}} p(x)\ln(p(x)) + (1-t)\sum_{x\in\mathcal{X}} q(x)\ln(q(x))\right) \text{ wegen Konkavität.}$$

□

Definition 4.2. Seien $\mathcal{X}$ höchstens abzählbar, P, Q Wahrscheinlichkeitsmaße auf $\mathcal{P}(\mathcal{X})$, p, q dazugehörige Gewichtsfunktionen. Die *Kulback-Leibler-Divergenz* oder *relative Entropie* von Q nach P ist definiert durch:

$$D(P\|Q) := \begin{cases} \sum_{x\in\mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) & \text{falls } P \ll Q \\ \infty & \text{sonst} \end{cases}$$

Die relative Entropie erinnert an einen Abstandsbegriff. Allerdings ist sie keine echte Metrik, da keine Symmetrie vorliegt. Offensichtlich gilt wegen $\ln(y) \leq y - 1$ für $y > 0$ zumindest:

$$\begin{aligned} D(P\|Q) &= -\sum_{x\in\mathcal{X}, p(x)>0} p(x) \ln\left(\frac{q(x)}{p(x)}\right) \\ &\geq -\sum_{x\in\mathcal{X}, p(x)>0} p(x)\left(\frac{q(x)}{p(x)} - 1\right) \\ &\geq -\sum_{x\in\mathcal{X}, p(x)>0} (q(x) - p(x)) = -Q\{x \mid p(x) > 0\} + 1 \geq 0. \end{aligned}$$

Falls $D(P\|Q) = 0$, so ist $\operatorname{supp}(P) = \operatorname{supp}(Q)$, denn

$$\begin{aligned} &\text{für } \subset\text{:} \quad \text{da } P \ll Q \\ &\text{für } \supset\text{:} \quad \text{wegen } Q\{x \mid p(x) > 0\} = 1 \end{aligned}$$

Es gilt sogar $P = Q$, denn es muss $\forall x \in \mathcal{X}$ gelten:

$$0 \geq \ln\left(\frac{q(x)}{p(x)}\right) - \frac{q(x)}{p(x)} + 1 \stackrel{!}{=} 0.$$

Dies gilt genau dann, wenn

$$\frac{q(x)}{p(x)} \stackrel{!}{=} 1 \text{ ist.}$$

Abbildung 5. Konvexitaet -> nur ein Beruehrungspunkt

Also haben wir bewiesen:

Lemma 4.3. *Es gelten*

$$D(P\|Q) \geq 0$$
$$D(P\|Q) = 0 \Leftrightarrow P = Q$$

Beispiel 4.4. Sei $\mathcal{X}$ endlich und Q die Gleichverteilung auf $\mathcal{X}$, $X\colon \Omega \to \mathcal{X}$ eine ZVe mit Verteilung $P_X = P$. Dann gilt:

$$\begin{aligned} 0 \leq D(P\|Q) &= -\sum_{x\in\mathcal{X}, p(x)>0} p(x) \ln\left(\frac{\frac{1}{|\mathcal{X}|}}{p(x)}\right) \\ &= -\ln\left(\frac{1}{|\mathcal{X}|}\right) \sum_{x\in\mathcal{X}, p(x)>0} p(x) + \sum_{x\in\mathcal{X}, p(x)>0} p(x)\ln(p(x)) \\ &= \ln|\mathcal{X}| - H(X) \\ \Leftrightarrow \quad H(X) &\leq \ln|\mathcal{X}|. \end{aligned}$$

Man überlege sich, dass Gleichheit genau dann gilt, wenn X gleichverteilt auf $\mathcal{X}$ ist.

Fazit:
$0 \leq H(X) \leq \ln|\mathcal{X}|$ für beliebige Zufallsvariablen X auf endlichem $\mathcal{X}$, wobei beide extremalen Werte tatsächlich angenommen werden können.

Interpretation: Linke Seite wird für $X \sim \delta_x$ angenommen, also, wenn gar kein Zufall sondern vollständige Information vorliegt. Für Gleichverteilungen liegt die schlechteste Informationslage (größtes Unsicherheit) vor.

4.2. **Entropie von Produkten und Kettenregel.**

Die Resultate in diesem Abschnitt 4.2 sollen die schon anfangs angesprochene Intuition, dass die Entropie die Unordnung bzw. Mangel an Information in einem W'Mass quantifiziert, vertiefen. Dazu ist es nützlich, Teilsysteme zu betrachten, die zu einem größerem Gesamtsystem zusammengefügt werden. Konkret betrachten wir im Folgenden Maße auf Produkträumen.
Seien $X\colon \Omega \to \mathcal{X}, Y : \tilde{\Omega} \to \mathcal{Y}$ diskrete ZVen. Sei $H(X,Y)$ Entropie des Zufallsvektors $(X,Y) : \Omega \times \tilde{\Omega} \to \mathcal{X} \times \mathcal{Y}$. Sei P die Verteilung von (X,Y) auf $\mathcal{X} \times \mathcal{Y}$ mit Gewichtsfunktion $p : \mathcal{X} \times \mathcal{Y} \to [0,1]$ und p_X und p_Y die Gewichtsfunktionen der Randverteilungen. Dann gilt

$$
\begin{aligned}
(4.1)\quad & I(X;Y) := H(X) + H(Y) - H(X,Y) \\
&= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\left(\ln(p_X(x)) + \ln(p_Y(y))\right) + \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\left(\ln(p(x,y))\right) \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\cdot \ln\left(\frac{p(x,y)}{p_X(x)p_Y(y)}\right) = D(P\|P_X\otimes P_Y) \geq 0.
\end{aligned}
$$

Also ist $H(X) + H(Y) - H(X,Y)$ die relative Entropie von dem Produkte der Marginalverteilungen zur der gemeisamen Verteilung P.

Definition 4.5. Die Größe $I(X;Y)$ heißt *gegenseitige (mutual) Information* von X und Y oder *Transinformation.*

Mit Lemma 4.3 folgt:

Lemma 4.6 (Subadditivität der Shannon-Entropie)**.**

$$
\begin{aligned}
&H(X,Y) \leq H(X) + H(Y) \\
&H(X,Y) = H(X) + H(Y) \text{ genau dann, wenn } P = P_X \otimes P_Y
\end{aligned}
$$

Diese Abschätzung entspricht der Intuition des Informationsbegriffs. Die gemeinsame Verteilung von X, Y enthält neben den Informationen zu den Marginalverteilungen zusätzlich noch Informationen zu den Abhängigkeiten zwischen X und Y. Sind X und Y unabhängig, so sind $H(X,Y)$ und $H(X) + H(Y)$ gleich.

Definition 4.7. Die *bedingte Entropie* von X bzgl. Y ist definiert als

$$H(X \mid Y) := H(X,Y) - H(Y)$$

womit man (4.1) umschreiben kann zu

$$H(X) = H(X|Y) + D(P\|P_X \otimes P_Y)$$

Setzen wir

$$p(x,y) = P(X = x, Y = y), \quad p(x \mid y) = P(X = x \mid Y = y).$$

so folgt aus der Definition mit ähnlichen Rechneschritten wie oben

$$\begin{aligned} H(X \mid Y) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\ln(p(x,y)) + \sum_{y\in\mathcal{Y}} p_Y(y)\ln(p_Y(y)) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\ln(p(x,y)) + \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\ln(p_Y(y)) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\frac{p_Y(y)}{p_Y(y)}\ln\left(\frac{p(x,y)}{p_Y(y)}\right) \qquad (4.2) \\ &= -\sum_{y\in\mathcal{Y}} p_Y(y)\sum_{x\in\mathcal{X}} p(x \mid y)\ln(p(x \mid y)) \\ &= E_Y\left[E(-\ln(p(X \mid Y) \mid Y)\right] = E[-\ln(p(X \mid Y)] \geq 0 \end{aligned}$$

wegen der Turmeigenschaft. Da

$$\begin{aligned} 0 \leq D(P_{(X,Y)}\|P_X \otimes P_Y) &= H(X) - H(X \mid Y) \\ \Leftrightarrow \quad H(X \mid Y) &\leq H(X) \qquad (4.3) \end{aligned}$$

sehen wir, dass Konditionierung die Entropie verringert.
Für beliebige diskrete ZVen X, Y, Z folgt aus den Definitionen, dass

$$\begin{aligned} H((X,Y) \mid Z) &= H((X,Y),Z) - H(Z) = \\ &= H(X,(Y,Z)) - H(Y,Z) + H(Y,Z) - H(Z) \\ &= H(X \mid (Y,Z)) + H(Y \mid Z) \end{aligned}$$

also

$$H(X,Y \mid Z) := H((X,Y) \mid Z) = H(X \mid Y,Z) + H(Y \mid Z) \text{ gilt.} \qquad (4.4)$$

Induktiv folgt die

Lemma 4.8 (Kettenregel für bedingte Entropien)**.** *Seien $n \in \mathbb{N}$ und $X_1, \ldots, X_n$ diskrete ZVen $X_i : \Omega \to \mathcal{X}_i$, so gilt:*

$$\begin{aligned} H(X_1,\ldots,X_n) = H(X_1) &+ H(X_2 \mid X_1) + H(X_3 \mid X_1, X_2) + \ldots \\ &+ H(X_{n-1} \mid X_1,\ldots,X_{n-2}) + H(X_n \mid X_1,\ldots,X_{n-1}). \end{aligned}$$

Bewisskizze: Addiere in (4.4) auf beiden Seiten $H(Z)$. □

4.3. **Han-Ungleichung.**

Satz 4.9 (Han-Ungleichung)**.**
Seien $X_1, \dots, X_n$ diskrete ZVen. Dann gilt

$$H(X_1, \dots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$$

Beweis. Sei $i = 1, \dots, n$. Nach der Definition der bedingten Entropie gilt

$$\begin{aligned} H(X_1, \dots, X_n) &= H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) + H(X_i \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \\ &\leq H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) + H(X_i \mid X_1, \dots, X_{i-1}) \text{ (Konditionierung).} \end{aligned}$$

Summiere nun über alle $i = 1, \dots, n$:

$$\begin{aligned} nH(X_1, \dots, X_n) &\leq \sum_{i=1}^{n} \Big[H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) + H(X_i \mid X_1, \dots, X_{i-1}) \Big] \\ &= \sum_{i=1}^{n} \Big[H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) + H(X_1, \dots, X_n) \Big] \text{ (Kettenregel)} \\ \Leftrightarrow \quad (n-1)H(X_1, \dots., X_n) &\leq \sum_{i=1}^{n} H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \end{aligned}$$

□

Für $n = 2$ entspricht die Han-Ungleichung der Subadditivität.

4.4. **Isoperimetrische Ungleichung auf binärem Würfel.**
Für x, x' aus dem binären Hyperwürfel $BHC := \{-1, 1\}^n$ (engl. *binary hyper cube*) definiere die *Hamming-Distanz*

$$d_H(x, x') := \sum_{i=1}^{n} \mathbb{1}_{\{x_i \neq x_i'\}}$$

und den Graphen mit Vertexmenge $V := \{-1, 1\}^n$ mit Kantenmenge $E := \big\{\{x, y\} \mid d_H(x, y) = 1\big\}$ mit Mächtigkeiten $|V| = 2^n =: N$ und $|E| = n2^{n-1}$. Die *Dichte des Graphen* ist definiert als

$$\frac{|E|}{|V|} = \frac{n}{2} = \frac{\ln_2(N)}{2} = \frac{\ln_2(|V|)}{2}.$$

Definition 4.10 (Induzierter Teilgraph)**.** Sei $G = (V, E)$ ein Graph und $A \subset V$. Der von A *induzierte Teilgraph* $G_A := (A, E_A)$ hat Vertexmenge A und Kantenmenge

$$E(A) := \{e \in E \mid \text{ beide Endpunkte von } e \text{ liegen in } A\}.$$

Wir werden zeigen, dass für beliebige $A \subset \{-1,1\}^n$ gilt:

$$\text{Dichte von } G_A \leq \frac{\ln_2 |A|}{2}$$

Gleichheit wird angenommen, falls A Hyperwürfel $\{-1,1\}^d$ mit $d \leq n$ ist, da dann obige Rechung

$$\text{Dichte von } G_A = \frac{\ln_2(2^d)}{2}$$

liefert. Idee der Dichte ist die Beschreibung von Verhältnis zwischen Oberfläche und Volumen: $\frac{|\partial A|}{|A|}$.

Hilft bei der Beschreibung Instabilität des Systems, genauer: Bei der Beantwortung der Frage: Wie vielen Kanten müssen gekappt werden, so dass das System in zwei Teilsysteme zerfällt?

Lemma 4.11 (Obere Schranke an Dichte des induzierten Teilgraphen)**.** *Seien $A \subset BHC$ und $E(A)$ die Kantenmenge des induzierten Teilgraphen G_A.*

$$\Rightarrow \quad |E(A)| \leq \frac{|A|}{2} \cdot \ln_2 |A|.$$

Beweis. Sei $X = (X_1, \ldots, X_n)$: $\Omega \to A$ gleichverteilt mit Gewichtsfunktion p.

$$\Rightarrow \quad H(X) = -\sum_{x \in A} p(x) \ln(p(x)) \qquad = -\frac{1}{|A|} \sum_{x \in A} \ln\left(\frac{1}{|A|}\right) = \ln |A|.$$

Sei $X^{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Mit (4.2) folgt

$$\Rightarrow \quad H(X) - H(X^{(i)}) = H(X_i \mid X^{(i)}) = -\sum_{x \in A} p(x) \ln(p(x_i \mid x^{(i)})).$$

Nach Definition ist $p(x) = \frac{1}{|A|} \; \forall\, x \in A$. Führe die Notation für den Vektor mit geflippten Vorzeichen an der i-ten Stelle ein $\overline{x}^{(i)} := (x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_n)$.
Nebenrechnung:

$$p(x_i \mid x^{(i)}) = P(X_i = x_i \mid X^{(i)} = x^{(i)}) = \frac{P(X = x)}{P(X^{(i)} = x^{(i)})} = \frac{1/|A|}{P(X^{(i)} = x^{(i)})}$$

$$P(X^{(i)} = x^{(i)}) = P(X^{(i)} = x^{(i)}, X_i = x_i) + P(X^{(i)} = x^{(i)}, X_i = -x_i)$$

Fallunterscheidung:

$$\begin{aligned}\overline{x}^{(i)} \notin A \quad &\Rightarrow \quad P(X^{(i)} = x^{(i)}) = P(X^{(i)} = x^{(i)}, X_i = x_i) = P(X = x)\\ \overline{x}^{(i)} \in A \quad &\Rightarrow \quad P(X^{(i)} = x^{(i)}) = 2P(X = x)\end{aligned}$$

$$\Rightarrow \ \forall\, x \in A, i = 1, \ldots, n : p(x_i \mid x^{(i)}) = \begin{cases} 1/2, & \text{falls } \overline{x}^{(i)} \in A \\ 1, & \text{sonst.} \end{cases}$$

Also:

$$H(X_i \mid X^{(i)}) = -\sum_{x \in A} p(x) \ln\left(\frac{1}{2}\right) \mathbb{1}_{\{\overline{x}^{(i)} \in A\}}$$

woraus nach Summation folgt:

$$\begin{aligned}\sum_{i=1}^{n} [H(X) - H(X^{(i)})] &= \frac{\ln(2)}{|A|} \sum_{x \in A} \sum_{i=1}^{n} \mathbb{1}_{\{\overline{x}^{(i)} \in A\}} \\ &= \frac{\ln(2)}{|A|} \cdot 2|E(A)|\end{aligned}$$

ABBILDUNG 6. Kanten in A

Aus der umgestellten Han-Ungleichung folgt:

$$\begin{aligned}2\ln(2)\frac{|E(A)|}{|A|} = \sum_{i=1}^{n} (H(X) - H(X^{(i)})) &\leq H(X) \qquad \text{Han} \\ &= \ln|A|, \qquad \text{da } X \text{ gleichverteilt.}\end{aligned}$$

$$\Rightarrow \quad |E(A)| \leq \frac{\ln|A| \cdot |A|}{2\ln(2)} = \frac{|A|}{2} \ln_2 |A|$$

□

Beispiel 4.12 (Einfluss von Kanten bzw. ZVen auf Ereignis.). Seien $X_1, \ldots, X_n \in \{-1, 1\} = \{\text{ja, nein}\}$ ZVen und $X = (X_1, \ldots, X_n)$ der gleichverteilte Zufallsvektor auf BHC.
Analog zu oben sei $\overline{X}^{(i)} := (X_1, \ldots, X_{i-1}, -X_i, X_{i+1}, \ldots, X_n)$ der Zufallsvektor mit geflippter i-ter Koordinate.

Definition 4.13. Für $A \subset \{-1,1\}^n$ ist der *Einfluss* der i-ten Variable definiert als:

$$\begin{aligned} I_i(A) &:= P(\mathbb{1}_{\{X\in A\}} \neq \mathbb{1}_{\{\overline{X}^{(i)}\in A\}}) \\ &= P(\{\omega \mid X(\omega) \in A, \overline{X}^{(i)}(\omega) \notin A\}) + P(\{\omega \mid X(\omega) \notin A, \overline{X}^{(i)}(\omega) \in A\}). \end{aligned}$$

In einer Konfiguration $\omega \in \{X \in A\}\Delta\{\overline{X}^{(i)} \in A\}$ heißt die i-te Variable X_i *pivotal* bzw. *entscheidend* für A. Der *totale Einfluss* oder die *Instabilität von* A ist

$$I(A) := \sum_{i=1}^{n} I_i(A).$$

Der *Kantenrand* von A wird definiert als:

$$\partial_E(A) := \big\{\{x,y\} \mid x \in A, y \in A^c, d_H(x,y) = 1\big\}$$

Damit gilt:

$$\begin{aligned} I(A) &= \frac{1}{2^n} \sum_{x\in\{-1,1\}^n} \sum_{i=1}^{n} \left(\mathbb{1}_{\{x\in A, \overline{x}^{(i)}\notin A\}} + \mathbb{1}_{\{x\notin A, \overline{x}^{(i)}\in A\}}\right) \\ &= \frac{1}{2^n} 2|\partial_E(A)| = \frac{|\partial_E(A)|}{2^{n-1}} \end{aligned}$$

Aus der isoperimetrischen Ungleichung in Lemma 4.11 erhalten wir:

Satz 4.14.
Für $A \subset \{-1,1\}^n$ setzen wir $P(A) = \frac{|A|}{2^n}$ (Laplace-Experiment) Dann gilt:

$$I(A) \geq 2P(A)\ln_2\left(\frac{1}{P(A)}\right).$$

Später folgt dies als Spezialfall von Theorem 5.1 in Kapitel 5.1.

Beweis. Jeder Punkt im BHC, also auch in A gehört zu n Kanten. Es gibt zwei Möglichkeiten: Beide Endpunkte sind in A, oder nur einer:

$$\begin{aligned} & n|A| = 2|E(A)| + 1 \cdot |\partial_E(A)| \\ \Rightarrow \quad & |\partial_E(A)| = n|A| - 2|E(A)| \geq [n - \ln_2(|A|)]|A| = |A|\ln_2\left(\frac{2^n}{|A|}\right). \\ \Rightarrow \quad & I(A) = \frac{2|\partial_E(A)|}{2^n} \geq 2P(A)\ln_2\left(\frac{1}{P(A)}\right). \end{aligned}$$

□

Bemerkung/Aufgabe: Sei X gleichverteilt auf $\{-1,1\}^n$ mit $X_1,\dots,X_n$ unabhängig. Dann ist $Z = f(X_1,\dots,X_n) := \mathbb{1}_{\{X\in A\}}$ eine Funktion dieser ZVen, die nur zwei Werte annimmt, also eine Bernoulli-ZV. Aus der E-S-U kann man folgern:

$$4P(A)(1-P(A)) = \mathrm{Var}(Z) \le I(A).$$

Frage: Wann ist diese Abschätzung und wann Satz 4.14 schärfer?

Anwendung/Diskussion von pivotalen Kanten in der Perkolationstheorie:
Sei $X \in \{-1,1\}^n$ gleichverteilt $\Leftrightarrow$
$X_1,\dots,X_n$ sind unabhängig, identisch verteilt mit $P(X_1 = 0) = P(X_1 = 1) = \frac{1}{2}$.
Verallgemeinerung: $P(X_1 = 1) = p, P(X_1 = -1) = 1-p$ immer noch unabhängig, identisch verteilt.
Analog zu oben:

$$I_i(A) = P(X_i \text{ pivotal für } A) \text{ und}$$
$$I(A) = \sum_{i=1}^{n} I_i(A).$$

Statt Indexmenge $\{1,\dots,n\}$ kann man die Vertexmenge eines endlichen oder unendlichen Graphen wählen.

Beispiel 4.15 (Perkolation auf $\mathbb{Z}^2$). Betrachte Graphen mit Vertexmenge $V = \mathbb{Z}^2$ und mit Kanten $e \in E$ zwischen allen Vertizes mit Euklidischen Abstand Eins.
Seien $X_e : \Omega \to \{0,1\}$, $e \in E$ unabhängig und identisch verteilt. Die Wahrschelichkeit für $\{X_0 = 1\}$ sei $p \in [0,1]$. Das induzierte Produktmaß bezeichnen wir mit $\mathbb{P}_p$
$X_e = 1$ bedeutet, dass e-te Kante Ausdünnung überlebt.
$X_e = 0$ bedeutet, dass e-te Kante entfernt wird (Perkolation).
Es entsteht nun ein Teilgraph G_ω , der von ω ab hängt, da zufällig erzeugt.

Betrachte nun ein Ereignis A, dass G_ω eine bestimmte Eigenschft erfüllt, z.B. dass der Ursprung $0 \in V$ mit allen seinen Nachbarecken durch eine aktive Kante verbunden ist.
Wie hängt die W'keit für das Eintreten von A von dem Bernoulli-Parameter p ab? Partielle Antwort liefert:

Lemma 4.16 (Formel von Russo)**.** *Sei $A \subset \Omega$ wachsendes Ereignis, das nur vom Zustand von endlich vielen Kanten abhängt (A Zylindermenge).*
Dann gilt

$$\frac{d}{dp}\mathbb{P}_p(A) = \mathbb{E}_p(|\delta A|) = \int\limits_{\Omega} |\delta A(\omega)|\mathbb{P}_p(d\omega))$$

Hierbei benutzen wie folgende

Definition 4.17. Seien $A \subset \Omega$ und $\omega \in \Omega$. Eine Kante $e \in E$ heißt *pivotal/entscheidend* für A in der Konfiguration $\omega \Leftrightarrow \mathbb{1}_A(\omega) - \mathbb{1}_A(\omega') \neq 0$
wobei

$$\omega'_f = \begin{cases} 1 - \omega_e & \text{für } f = e \\ \omega_f & \text{sonst} \end{cases} \qquad \text{Zustand der Kante } e \text{ geswitcht.}$$

Zustand der Kante e entscheidet über Eintreten oder Nichteintreten von Ereignis A in der Konfiguration ω, genauer in der partiellen Konfiguration $\omega^{\perp} = (\omega_f)_{f \in E \setminus \{e\}}$. Die Menge der pivotalen Kanten von A wird mit δA bezeichnet.

Spezialfall: A wachsend, dann ist e pivotal, falls

$$(\omega\perp \text{ ergänzt mit } \omega_e = 1) \in A$$
$$(\omega\perp \text{ ergänzt mit } \omega_e = 0) \notin A$$

Bezug zum Einfluss der Kante $e = e_i$:

$$I_i(A) = \mathbb{P}_p(e_i \text{ ist pivotal})$$
$$I(A) = \sum_{e \in E} \mathbb{E}_p(\mathbb{1}_{e \text{ ist pivotal}}) = \mathbb{E}_p(|\delta A|)$$

Beweis. Sei $\Lambda \subset E$ endlich, so dass Eintreten von A nur vom Zustand der Kanten in Λ abhängt. Betrachte verallgemeinertes unabhängiges Kantenperkolationsmodell mit $\mathbb{P}(\text{Kante } e \text{ aktiv}) = p_e \in [0,1]$
Definiere Folge/Vektor $\vec{p} = (p_e, e \in \Lambda)$
und Produktmaß $\mathbb{P}_{\vec{p}} = \bigotimes\limits_{e \in \Lambda} (p_e\delta_1 + (1 - p_e)\delta_0)$.
Im Spezialfall

$$\vec{p} = (\underbrace{p}_{\text{unabh. von } e}, e \in \Lambda) \quad \text{ist} \quad \mathbb{P}_{\vec{p}} = \mathbb{P}_p \circ \pi_{\Lambda}^{-1} \tag{4.5}$$

wobei $\pi_\Lambda \colon \Omega \to \Omega_\Lambda := \{0,1\}^\Lambda$ die Projektion auf den Produktraum zu der endlichen Kantenmenge $\Lambda \subset E$.

Wollen Kettenregel auf Komposition folgender Funktionen anwenden:

$$h : \mathbb{R} \to \mathbb{R}^\Lambda, h(p) = (p, p, \ldots, p) = (p, e \in \Lambda)$$
$$g : \mathbb{R}^\Lambda \to \mathbb{R}, g(\vec{p}) = \mathbb{P}_{\vec{p}}(A)$$

Dann gilt $\mathbb{P}_p(A) \overset{(4.5)}{=} \mathbb{P}_{h(p)}(A) = g(h(p))$
Kettenregel:

$$\begin{aligned}\frac{d}{dp}\mathbb{P}_p(A) = \frac{d}{dp}g(h(p)) &= \sum_{e\in\Lambda} \frac{\partial}{\partial p_e} g(h(p)) \ \frac{d}{dp}h_e(p) \\ &= \sum_{e\in\Lambda} \underbrace{\frac{\partial}{\partial p_e}\mathbb{P}_{\vec{p}}(A)|_{p_e=p}}_{\text{zu berechnen}} \ \underbrace{\frac{d}{dp}p}_{=1}\end{aligned}$$

Für $e \in E$ sei $\omega^\perp = \omega^{\perp e} \in \Omega_{\Lambda\setminus\{e\}} = \{0,1\}^{\Lambda\setminus\{e\}}$ und sei $\omega^{+e}, \omega^{-e} \in \Omega_\Lambda = \{0,1\}^\Lambda$ gegeben durch:

$$\omega^{+e}(f) = \begin{cases}\omega^\perp(f) & f \in \Lambda \setminus \{e\} \\ 1 & f = e\end{cases}, \qquad \omega^{-e}(f) = \begin{cases}\omega^\perp(f) & f \in \Lambda \setminus \{e\} \\ 0 & f = e\end{cases}$$

Zerlegung/Fallunterscheidungsformel/diskretes W-Modell

$$\mathbb{P}_{\vec{p}}(A) = \sum_{\omega_e\in\{0,1\}, e\in\Lambda} \mathbb{P}_{\vec{p}}(\omega)\mathbb{1}_A(\omega) = \sum_{\omega^\perp\in\Omega_{\Lambda\setminus\{e\}}} \mathbb{P}_{\vec{p}}(\omega^\perp)\big[p_e\mathbb{1}_A(\omega^{+e}) + (1-p_e)\mathbb{1}_A(\omega^{-e})\big]$$

Ableitung nach p_e:

$$\begin{aligned}\frac{\partial}{\partial p_e}\mathbb{P}_{\vec{p}}(A) &= \sum_{\omega^\perp\in\Omega_{\Lambda\setminus\{e\}}} \mathbb{P}_{\vec{p}}(\omega^\perp)\left[\mathbb{1}_A(\omega^{+e}) - \mathbb{1}_A(\omega^{-e})\right] \\ &= \sum_{\omega^\perp\in\Omega_{\Lambda\setminus\{e\}}} \mathbb{P}_{\vec{p}}(\omega^\perp)\ \mathbb{1}_{\{e\in\delta A\}}(\omega^\perp) = \mathbb{P}_{\vec{p}}(e \in \delta A)\end{aligned}$$

Hierbei haben wir benutzt:

$\mathbb{1}_A(\omega^{+e}) - \mathbb{1}_A(\omega^{-e}) \geq 0$ da A wachsend
$\mathbb{1}_A(\omega^{+e}) - \mathbb{1}_A(\omega^{-e}) = 1$ genau dann, wenn $\omega^{+e} \in A$ und $\omega^{-e} \notin A$
, d.h. e pivotal für A in $\omega^\perp$

Es folgt

$$\frac{d}{dp}\mathbb{P}_p(A) = \sum_{e\in\Lambda} \frac{\partial}{\partial p_e}\mathbb{P}_{\vec{p}}(A) \mid_{p_e=p} = \sum_{e\in\Lambda} \underbrace{\mathbb{P}_{\vec{p}}(e \in \delta A) \mid_{p_e=p}}_{=\mathbb{P}_p(e\in\delta A)}$$

$$= \mathbb{E}_p\left(\sum_{e\in\Lambda} \mathbb{1}_{\{e\in\delta A\}}\right) = \mathbb{E}_p(|\delta A|)$$

□

4.5. **Kombinatorische Entropien.**
Viele kombinatorisch Entropien sind selbsbeschränkende Funktionen $Z = f(X_1, \ldots, X_n)$ von ZV $X_1, \ldots, X_n$. In Kapitel 3.3 haben wir gesehen, dass dann die E-S-U anwendbar ist und Schranke an die Varianz liefert.
Wir betrachten zunächts ein Beispiel und nutzen die Han-Ungleichung, um z.z., dass f self-bounding ist.

Beispiel 4.18. (Vapnik-Chervonenkis-Entropie (VC-Entropie))
Sei $\mathcal{X} \neq \emptyset, \mathcal{A} \subset \mathcal{P}(\mathcal{X})$ (nicht notwendig σ-Algebra) und $x = (x_1, \ldots, x_n)$ Vektor in $\mathcal{X}^n$, oft identifiziert mit Menge $F = \{x_1, \ldots, x_n\} \subset \mathcal{X}^n$ mit $|F| \leq n$. Definiere die *Spur* (trace) von A auf F:

$$\mathrm{Tr}(x) := \{A \cap F \mid A \in \mathcal{A}\} \subset \mathcal{P}(F).$$

Dann heißt

$$T(x) := |\,\mathrm{Tr}(x)|$$

shatter coefficient (Identifikationskoeffizeint) von x. Die VC-Entropie ist

$$h(x) = \ln_2(T(x)).$$

Da

$$T(x) \leq |\mathcal{P}(F)| \leq 2^n \quad \text{gilt} \quad h(x) \leq \ln_2(2^n) = n$$

(vgl. obere Schranke an Shannon-Entropie) und

$$T(x) = 2^n \Leftrightarrow h(x) = n \Leftrightarrow \mathcal{A} \text{ identifiziert } x.$$

Lemma 4.19. *Die VC-Entropie $\mathcal{X}^n \ni x \to h(x)$ ist self-bounding.*

Beweis. Es reicht ein $g : \mathcal{X}^{n-1} \to \mathbb{R}$, zu finden, so dass $\forall i = 1, \ldots, n$

$$0 \leq h(x) - g(x^{(i)}) \leq 1, \text{ wobei } x^{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \tag{4.6}$$

$$\text{und } \sum_{i=1}^{n} (h(x) - g(x^{(i)})) \leq h(x). \tag{4.7}$$

Natürlicher Kandidat

$$g(x^{(i)}) = \ln_2(|\operatorname{Tr}(x^{(i)})) = \ln_2(|\{A \cap \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\} \mid A \in \mathcal{A}\}|).$$

Da als Menge $x^{(i)} \subset x$ gilt, folgt $T(x^{(i)}) \leq T(x)$ und

$$g(x^{(i)}) \leq h(x) \, \forall i = 1, \ldots, n$$

Andererseits hat x maximal einen Punkt mehr als $x^{(i)}$

$$\implies \sharp \text{ der Teilmengen verdoppelt sich maximal}$$
$$\implies T(x) \leq 2T(x^{(i)}) \implies h(x) \leq g(x^{(i)}) + 1$$

insgesamt also

$$0 \leq h(x) - g(x^{(i)}) \leq 1.$$

Nun ist noch die zweite Eigenschaft (4.7) z.z.:
Sei $Y = (Y_1, \ldots, Y_n) \in \{0,1\}^n$ ein zufälliger Vektor, dessen Verteilung die Gleichverteilung auf der endlichen Familie von Mengen $\operatorname{tr}(x)$ ist. (Y kann ich ja als Indikatorfunktion einer Menge auffassen.)

$$\Rightarrow \; h(x) = \ln_2 |\operatorname{Tr}(x)| = \frac{\ln |\operatorname{Tr}(x)|}{\ln(2)} = \frac{H(Y)}{\ln(2)}$$

da ja bei Gleichverteilung $H(Y) = \ln |\text{Menge}|$ gilt. Die ZVe

$$Y^{(i)} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n) \in \{0,1\}^{n-1}$$

ist Indikatorfunktion einer Menge mit Kardinalität $\leq n-1$. Egal welche Verteilung sie hat, impliziert Bsp.4.4

$$\Rightarrow \quad H(Y^{(i)}) \leq \ln |\operatorname{Tr}(x^{(i)})| = \ln T(x^{(i)}) = \ln(2) \ln_2(T(x^{(i)})) = \ln(2) g(x^{(i)}).$$

Die Han-Ungleichung liefert

$$h(x) = \frac{H(Y)}{\ln(2)} \overset{\text{Han-U.}}{\leq} \frac{1}{n-1} \sum_{i=1}^{n} \frac{H(Y^{(i)})}{\ln(2)} \leq \frac{1}{n-1} \sum_{i=1}^{n} g(x^{(i)})$$

$$\text{Umstellen } \Rightarrow \; \sum_{i=1}^{n} (h(x) - g(x^{(i)})) \leq h(x).$$

□

Zusammen mit Korollar 3.10 folgt unmittelbar.

Korollar 4.20.
Seien $X_1, \ldots, X_n \in \mathcal{X}$ unabhängige ZVen und $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$. Ist $Z = h(X)$ die VC-Entropie von $X = (X_1, \ldots, X_n)$ (bzgl. $\mathcal{A}$), gilt:

$$\mathrm{Var}(Z) \leq E(Z).$$

Beweisidee von oben kann auf andere Varianten der kombinatorischen Entropie erweitert werden:

Definition 4.21. Seien $\mathcal{X} \neq \emptyset$ Menge, $x_i \in \mathcal{X}_i$ für $i = 1, \ldots, n$, also $x = (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \mathcal{X}_n$ und $\mathcal{Y} \neq \emptyset$ eine (potentiell andere) Menge. Sei

$$\mathrm{Tr} : \mathcal{X}_1 \times \ldots \mathcal{X}_n \to \mathcal{P}(\mathcal{Y}^n) \text{ also } \mathrm{Tr}(x) \subset \mathcal{Y}^n.$$

Für $i = 1, \ldots, n$ und $x^{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ definiere die Projektionsmenge

$$\begin{aligned} \mathrm{Tr}(x^{(i)}) = \{&y^{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) \in \mathcal{Y}^{n-1} \\ &\exists\, y_i \in \mathcal{Y} : \text{ so dass } (y_1, \ldots, y_i, \ldots, y_n) \in \mathrm{Tr}(x)\} \end{aligned} \tag{4.8}$$

Für $b > 1$ setze $h(x) = \ln_b |\,\mathrm{Tr}(x)|$. Solche Funktionen heißen *kombinatorische Entropie*.

Satz 4.22.
Sei $\mathcal{X}^n \ni x \mapsto h(x) = \ln_b |\,\mathrm{Tr}(x)|$ eine kombinatorische Entropie, so dass für alle $x \in \mathcal{X}^n$ und $i = 1, \ldots, n$ gilt:

$$h(x) - \ln_b |\,\mathrm{Tr}(x^{(i)})| \leq 1. \tag{4.9}$$

Dann ist h self-bounding und für jeden Vektor $X = (X_1, \ldots, X_n)$ von unabhängigen ZVen in $\mathcal{X}$ gilt

$$\mathrm{Var}(h(X)) \leq E(h(X)).$$

Beweis. Übung. □

4.6. **Han-Ungleichung für relative Entropien.**
Sei $\mathcal{X} \neq \emptyset$ abzählbar, P, Q W'maße auf $\mathcal{X}^n = \mathcal{X}^{\{1,\dots,n\}} := \{f\colon \{1,\dots,n\} \to \mathcal{X}\}$, wobei $P = P_1 \otimes \dots \otimes P_n$ ein Produktmaß mit Randmaßen P_i auf $\mathcal{X}$. Sei $Q^{(i)}$ die Randverteilung von Q auf $\mathcal{X}^{\{1,\dots,i-1,i+1,\dots,n\}}$.
Folgende Ungleichung spielt für die Subadditivität der Entropie, die man wiederum nutzen kann, um exponentielle Konzentrationsungleichungen herzuleiten.

Satz 4.23. *(Han-Ungleichung für relative Entropien)*
Es gelten:

$$D(Q\|P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q(i)\|P^{(i)}) \text{ oder äquivalent}$$

$$D(Q\|P) \leq \sum_{i=1}^{n} (D(Q\|P) - D(Q^{(i)}\|P^{(i)})).$$

Beweis. Wie gehabt:

$$\begin{aligned} \text{Zu} \quad & x = (x_1, \dots, x_n) \\ \text{setze } & x^{(i)} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n). \end{aligned}$$

Bezeichne mit p, q die Gewichtsfunktionen zu P und Q. Entsprechende Rand-Gewichtsfunktionen $q^{(i)}$ und $p^{(i)}$ sind definiert durch

$$\begin{aligned} q^{(i)}(x^{(i)}) &= \sum_{y \in \mathcal{X}} q(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n) \\ p^{(i)}(x^{(i)}) &= \sum_{y \in \mathcal{X}} p(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n) = \prod_{j=1,\dots,n; j \neq i} p(x_j). \end{aligned}$$

bemäßder Produktannahme. Dann ist $Q^{(i)}$ Randverteilung zur Gewichtsfunktion $q^{(i)}$. Es gelten:

(4.10)

$$H(Q) = \sum_{x \in \mathcal{X}^n} q(x) \ln(q(x)) \overset{\text{Han-U.}}{\geq} \frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)} \in \mathcal{X}^{n-1}} q^{(i)}(x^{(i)}) \ln(q^{(i)}(x^{(i)}))$$

(4.11)

$$D(Q\|P) = \sum_{x \in \mathcal{X}^n} q(x) \ln\left(\frac{q(x)}{p(x)}\right) = \sum_{x \in \mathcal{X}^n} q(x) \ln(q(x)) - \sum_{x \in \mathcal{X}^n} q(x) \ln(p(x))$$

und analog:

(4.12)
$$D(Q^{(i)}\|P^{(i)}) = \sum_{x^{(i)}\in\mathcal{X}^{n-1}} q^{(i)}(x^{(i)}) \left(\ln(q^{(i)}(x^{(i)}))) - \ln(p^{(i)}(x^{(i)}))\right).$$

Nach dem Einsetzen von (4.10) und (4.12) in die Formel (4.11) für $D(Q\|P)$ müssen wir nur noch den gemischten $q-p$-Term verstehen. Dazu nutzen wir, dass wegend der Produktstruktur von P gilt: $p(x) = p^{(i)}(x^{(i)}) \cdot p_i(x_i)$.

$$\begin{aligned} n \sum_{x\in\mathcal{X}^n} q(x)\ln(p(x)) &= n \sum_{x\in\mathcal{X}^n} q(x) \left(\ln(p^{(i)}(x^{(i)})) + \ln(p_i(x_i))\right) \\ &= \sum_{i=1}^{n} \sum_{x\in\mathcal{X}^n} q(x)\ln(p^{(i)}(x^{(i)})) + \sum_{x\in\mathcal{X}^n} q(x) \sum_{i=1}^{n} \ln(p_i(x_i)) \\ &= \sum_{i=1}^{n} \sum_{x\in\mathcal{X}^n} q(x)\ln(p^{(i)}(x^{(i)})) + \sum_{x\in\mathcal{X}^n} q(x)\ln(p(x)) \end{aligned}$$

Äquivalent umstellen:

$$\begin{aligned} \Leftrightarrow \quad (n-1) \sum_{x\in\mathcal{X}^n} q(x)\ln(p(x)) &= \sum_{i=1}^{n} \sum_{x\in\mathcal{X}^n} q(x)\ln(p^{(i)}(x^{(i)})) \\ &= \sum_{i=1}^{n} \sum_{x^{(i)}\in\mathcal{X}^{n-1}} q^{(i)}(x^{(i)})\ln(p^{(i)}(x^{(i)})), \end{aligned}$$

da nach der Definition von $q^{(i)}$

$$\begin{aligned} \sum_{x^{(i)}\in\mathcal{X}^{n-1}} q^{(i)}(x^{(i)}) f_i(x^{(i)}) &= \sum_{x^{(i)}\in\mathcal{X}^{n-1}} \sum_{y\in\mathcal{X}} q(x_1,\ldots,x_{i-1},y,x_{i+1},\ldots,x_n) f_i(x^{(i)}) \\ &= \sum_{x\in\mathcal{X}^n} q(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) f_i(x^{(i)}) \end{aligned}$$

für beliebige $f_i\colon x^{(i)} \mapsto f_i(x^{(i)}) \in \mathbb{R}$ gilt. □

4.7. **Sub-Additivität der Entropie.**
Im folgenden sei $\phi\colon [0,\infty) \to \mathbb{R}, \phi(x) = x\ln(x)$ für $x > 0$ und $\phi(x) = 0$ für $x = 0$ gemäss der Konvention $0 \cdot \ln 0 = 0$. Dann ist ϕ strikt konvex und $\phi \geq -\frac{1}{e}$ auf $(0,\infty)$, denn

$$\begin{aligned} \phi'(x) &= \frac{x}{x} + \ln x = 1 + \ln x \\ \phi''(x) &= \frac{1}{x} > 0 \end{aligned}$$

sowie

$$0 = \phi'(x) \Leftrightarrow \ln x = -1 \Leftrightarrow x = \frac{1}{e}$$

mit Wert

$$\phi(\frac{1}{e}) = \frac{1}{e} \cdot \ln(\frac{1}{e}) = -\frac{1}{e}$$

Wegen der strikten Konvexität ist dies ist das einzige und isolierte Minimum von ϕ.

Definition 4.24 (Entropie)**.** Sei $Z \geq 0, Z \in \mathcal{L}^1(\Omega, P)$. Dann nennen wir

$$\mathrm{Ent}(Z) := E(\phi(Z)) - \phi(E(Z))$$

die *Entropie* von Z. Die Jensen-Ungleichung

$$E(\phi(Z)) \geq \phi(E(Z))$$

impliziert $\mathrm{Ent}(Z) \in [0, \infty]$.

Dies ist ein zu der Shannon-Entropie verwandnter, aber distinkter Begriff. Die Efron-Stein-Ungleichung besagt:

$$\begin{aligned} E(Z^2) - E(Z)^2 = \mathrm{Var}(Z) &\leq \sum_{i=1}^{n} E((Z - E^{(i)}(Z))^2) \\ &= \sum_{i=1}^{n} E(E^{(i)}(Z - E^{(i)}(Z))^2) = \sum_{i=1}^{n} E(E^{(i)}(Z^2) - E^{(i)}(Z)^2). \end{aligned}$$

Wir zeigen nun eine analoge Zerlegungsabschätzung für ϕ.

Satz 4.25. *(Sub-Additivität der Entropie)*
Seien $X_1, \ldots, X_n$ unabhängige ZVen mit Werten in abzählbarer Menge $\mathcal{X}$, $f\colon \mathcal{X}^n \to [0, \infty)$, und $Z := f(X_1, \ldots, X_n)$. Dann gilt:

$$E(\phi(Z)) - \phi(E(Z)) \leq \sum_{i=1}^{n} E(E^{(i)}(\phi(Z)) - \phi(E^{(i)}(Z))),$$

Mit der Notation $\mathrm{Ent}^{(i)}(Z) := E^{(i)}(\phi(Z)) - \phi(E^{(i)}(Z))$ wird das zu

$$\mathrm{Ent}(Z) \leq E\Big(\sum_{=1}^{n} \mathrm{Ent}^{(i)}(Z)\Big)$$

Beweis. Sei o.E. $E(Z) = 1$ (sonst normierte Version betrachten, vgl. Übung). Bezeichne P das W'maß des Vektors $(X_1, \ldots, X_n)$ mit Gewichtsfunktion

$$x = (x_1, \ldots, x_n) \mapsto p(x) = \prod_{i=1}^{n} p_i(x_i).$$

Definiere das W'maß Q auf $\mathcal{X}^n$ durch $q(x) = f(x)p(x)\,\forall x \in \mathcal{X}^n$ (wohldefiniertes W'maß, da $E(Z) = 1$, sonst zusätzliche Konstanten). Dann gilt:

$$\begin{aligned} E(\phi(Z)) - \underbrace{\phi(E(Z))}_{=0} &= E(Z\ln(Z)) = \sum_{x\in\mathcal{X}^n} q(x)\ln(f(x)) \\ &= \sum_{x\in\mathcal{X}^n} q(x)\ln\left(\frac{q(x)}{p(x)}\right) = D(Q\|P) \\ &\overset{\text{Han-U.}}{\leq} \sum_{i=1}^{n}\Big(D(Q\|P) - D(Q^{(i)}\|P^{(i)})\Big) \\ &= \sum_{i=1}^{n}\left(\underbrace{E(E^{(i)}(\phi(Z)))}_{Turmeig.} - \underbrace{D(Q^{(i)}\|P^{(i)})}_{\overset{\text{z.z.}}{=} E\phi(E^{(i)}(Z))}\right). \end{aligned}$$

Dazu benutze, dass wg. Unabh.

$$p^{(i)}(x^{(i)}) = \prod_{j=1,\ldots,n;j\neq i} p_j(x_j).$$

Analog zu früherer Rechnung

$$\begin{aligned} q^{(i)}(x^{(i)}) &= \sum_{y\in\mathcal{X}} q(x_1,\ldots,x_{i-1},y,x_{i+1},\ldots,x_n) \\ &= \sum_{y\in\mathcal{X}} f(x_1,\ldots,y,\ldots,x_n)p_i(y)\prod_{j=1,\ldots,n;j\neq i} p_j(x_j) \\ &= E^{(i)}(Z)\,p^{(i)}(x^{(i)}). \end{aligned}$$

Einsetzen in K-L-D ergibt:

$$\begin{aligned} D(Q^{(i)}\|P^{(i)}) &= \sum_{x^{(i)}\in\mathcal{X}^{n-1}} q^{(i)}(x^{(i)})\ln\left(\frac{q^{(i)}(x^{(i)})}{p^{(i)}(x^{(i)})}\right) \\ &= \sum_{x^{(i)}\in\mathcal{X}^{n-1}} p^{(i)}(x^{(i)})E^{(i)}(Z)\ln(E^{(i)}(Z)) \\ &= \sum_{x_i\in\mathcal{X}} p_i(x)\left(\sum_{x^{(i)}\in\mathcal{X}^{n-1}} p^{(i)}(x^{(i)})E^{(i)}(Z)\ln(E^{(i)}(Z))\right) \\ &= \sum_{x\in\mathcal{X}^n} p(x)\phi(E^{(i)}(Z)) = E(\phi(E^{(i)}(Z))). \end{aligned}$$

□

Nun betrachten wir eine Verallgemeinerung von Theorem 4.14 für $p \neq \frac{1}{2}$. Dazu sei $p \in [0,1]$ und $X = (X_1, \ldots, X_n)$, wobei X_i unabhängige, identisch verteilte Bernnoulli-ZVen mit $P(X_i = 1) = p = 1 - P(X_i = -1)$. Ähnlich wie bei dem Perkolationsbeispiel setzen wir für $i \in \{1, \ldots, n\}$

$$X_i^+ := (X_1, \ldots, X_{i-1}, 1, X_{i+1}, \ldots, X_n) \text{ und}$$
$$X_i^- := (X_1, \ldots, X_{i-1}, -1, X_{i+1}, \ldots, X_n).$$

Sei $A \subset \{-1,1\}^n$. Der positive bzw. negative Einfluss der i-ten Koordinate auf A ist

$$I_i^+(A) = P(X_i^+ \in A \text{ und } X_i^- \notin A)$$
$$I_i^-(A) = P(X_i^+ \notin A \text{ und } X_i^- \in A)$$

Dann ist der (im Fall $p = 1/2$ bereits bekannte) Einfluss der i-ten Koordinate auf A:

$$I_i(A) := P(\mathbb{1}_A(X) \neq \mathbb{1}_A(\overline{X}^{(i)}) = I_i^+(A) + I_i^-(A)$$

Der totale positive bzw. negative Einfluss ist:

$$I^+(A) = \sum_{i=1}^n I_i^+(A), \quad I^-(A) = \sum_{i=1}^n I_i^-(A).$$

Satz 4.26.
Seien $A \subset \{-1,1\}^n$, $p \in (0,1)$, X und $I_i^+(A), I_i^-(A)$ wie oben. Setze $P(A) := P(X \in A)$. Dann ist

$$P(A) \ln\left(\frac{1}{P(A)}\right) \leq I^+(A) p \ln\left(\frac{1}{p}\right) + I^-(A)(1-p)\ln\left(\frac{1}{1-p}\right).$$

Für $p = \frac{1}{2}$ liefert dies Theorem 4.14.
Betrachte folgenden Spezialfall

Definition 4.27. Teilmenge A von $\{-1,1\}^n$ heißt *wachsend* :$\Leftrightarrow$

$\mathbb{1}_{\{x \in A\}} \geq \mathbb{1}_{\{y \in A\}} \forall x = (x_1, \ldots, x_n) \forall y = (y_1, \ldots, y_n)$ mit $x_i \geq y_i \forall i = 1, \ldots, n$.
In anderen Worten: $x \in A \Rightarrow (x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n) \in A$.

Viele interessante (und vernünftige) Ereignisse sind wachsend.

Korollar 4.28.
Ist $A \subset \{-1,1\}^n$ wachsend, so ist $I^-(A) = 0$ und wir erhalten:

$$I(A) = I^+(A) \geq \frac{P(A)\ln\left(\frac{1}{P(A)}\right)}{p\ln\left(\frac{1}{p}\right)} = \frac{P(A)\ln(P(A))}{p\ln(p)} = \frac{P(A)}{p}\ln_{1/p}(P(A))$$

Beweis. (von Satz 4.26) Sei $Z = \mathbb{1}_{X \in A} \in \{0,1\}$. Mit Satz 4.25 folgt:

$$\underbrace{E(\phi(Z))}_{=0} - \phi(E(Z)) \leq \sum_{i=1}^{n} E(\underbrace{E^{(i)}(\phi(Z))}_{=0} - \phi(E^{(i)}(Z))), \tag{4.13}$$

da $\phi(0) = \phi(1) = 0$. Wegen $\phi(E^{(i)}(Z)) = \phi(p\mathbb{1}_A(X_i^+) + (1-p)\mathbb{1}_A(X_i^-))$ gilt zudem:

$$\phi(E^{(i)}(Z)) = \begin{cases} p\ln(p), & \text{wenn } X_i^+ \in A \text{ und } X_i^- \notin A \\ (1-p)\ln(1-p), & \text{wenn } X_i^- \in A \text{ und } X_i^+ \notin A \\ 0, & \text{wenn } i \text{ nicht pivotal für } A \text{ ist} \end{cases}$$

Die rechte Seite von (4.13) ergibt:

$$\begin{aligned} &-\sum_{i=1}^{n} E\left(p\ln(p)\mathbb{1}_A(X_i^+)\cdot\mathbb{1}_{A^c}(X_i^-) + (1-p)\ln(1-p)\mathbb{1}_A(X_i^-)\cdot\mathbb{1}_{A^c}(X_i^+)\right) \\ &= -\sum_{i=1}^{n}\left(p\ln(p)I_i^+(A) + (1-p)\ln(1-p)I_i^-(A)\right) \\ &= p\ln\left(\frac{1}{p}\right)I^+(A) + (1-p)\ln\left(\frac{1}{1-p}\right)I^-(A). \end{aligned}$$

Die linke Seite von (4.13) ergibt:

$$-\phi(E(Z)) = -\phi(P(A)) = -P(A)\ln(P(A)) = P(A)\ln\left(\frac{1}{P(A)}\right).$$

□

4.8. **Entropie für allgemeine Zufallsvariablen.**
Die Shannon-Entropie haben wir in dem speziellen Setting von diskreten ZV eingeführt. Nun wenden wir uns einem allgemeineren Entropiebegriff zu. Einige Eigenschaften und Bezüge übertragen sich, auch neue Phänomene können studiert werden.
Caveat: Auch wenn sie viele Eingeschaften teilen, werden *unterschiedliche* Funktionen als Entropie bezeichnet.
Im Folgenden sei wieder

$$\phi\colon [0,\infty) \to \mathbb{R}, \quad \phi(x) = \begin{cases} x\ln(x), & x > 0 \\ 0, & x = 0. \end{cases}$$

Erinnerung: Kurvendiskussion liefert: ϕ ist konvex und besitzt globales Minimum an der Stelle $x = e^{-1}$ mit globalem Minimum $\phi(e^{-1}) = -e^{-1}$.

Seien $(\Omega, \mathcal{A}, \mu)$ Maßraum (μ nicht zwingend normiert), $X, Y : \Omega \to [0,\infty)$ messbare Abbildungen, so dass die Maße $P, Q : \mathcal{A} \to [0,1]$ definiert durch

$$P(A) := \int_A X(\omega)\mu(d\omega)$$
$$Q(A) := \int_A Y(\omega)\mu(d\omega),$$

W'maße sind. Offensichtlich $P, Q \ll \mu$. Setze

$$W = X\ln\left(\frac{X}{Y}\right) : \Omega \to (-\infty, \infty]$$

Lemma 4.29. *Sei $P(\{\omega \in \Omega \mid Y(\omega) = 0\}) = 0$. Dann gelten*

(a) *$\int_\Omega W(\omega)\mu(d\omega)$ ist wohldefiniert in $[0,\infty]$.*
(b) *Gilt $\int_\Omega W(\omega)\mu(d\omega) = 0$, so ist $P = Q$.*

Dies motiviert folgende

Definition 4.30. Seien $(\Omega, \mathcal{A}, \mu), X, Y, P, Q$ wie oben. Dann heißt

$$D(P\|Q) := \begin{cases} E_P(\ln(\frac{X}{Y})) = \int_\Omega X\ln(\frac{X}{Y})d\mu, & P(Y=0) = 0 \\ \infty, & \text{sonst} \end{cases}$$

die *relative Entropie* von Q nach P oder *Kulback-Leibler-Divergenz*.

Beweis. (des Lemmas) Sei $f\colon \Omega \to [0,\infty)$

$$f(\omega) := \begin{cases} \frac{X(\omega)}{Y(\omega)}, & \text{falls } Y(\omega) > 0 \\ 1, & \text{sonst} \end{cases}$$

Betrachte Maß $\tilde{P} : \mathcal{A} \to [0,1]$ mit Dichte $f \cdot Y$ bzgl. μ.

$$\tilde{P}(A) = \int \mathbb{1}_A f Y d\mu = \int_{\{Y>0\}} \mathbb{1}_A f Y d\mu + \int_{\{Y=0\}} \mathbb{1}_A f \cdot 0 \, d\mu$$

$$= P(\{Y > 0\} \cap A) = P(A), \text{ da } P(Y = 0) = 0. \tag{4.14}$$

Das von fY induzierte Maß $\tilde{P}$ ist also identisch mit P. O.E. (durch Anpassung auf einer μ-Nullmenge) ist $X = fY$. Insbesondere gilt

$$P = f(Y\mu) = fQ \ll Q.$$

Setze

$$\psi \colon [0,\infty) \to [0,\infty), \psi(s) = \begin{cases} 1 - s + s\ln(s), & s > 0 \\ 1, & s = 0 \end{cases}$$

Dann gilt für $s > 0$

$$\psi'(s) = -1 + \ln(s) + \frac{s}{s} = \ln(s) = \begin{cases} < 0 & s \in (0,1) \\ = 0 & s = 1 \\ > 0 & s > 1 \end{cases}$$

stetig fortsetzbar zu $\psi'(0) = -\infty$. Insbesondere gilt $\psi(s) \geq \psi(1) = 0$ für alle $s \geq 0$. Das heißt $\psi \geq 0$ überall und ψ ist strikt konvex. Damit ist auch $(\psi \circ f)(\omega) \geq 0$ auf ganz Ω. Also ist

$$\int_\Omega \psi(f(\omega)) Y(\omega) \mu(d\omega) = E_Q(\psi \circ f)$$

wohldefiniert als Element von $[0,\infty]$. Wegen

$$E_Q(1 - f) = E_Q[(1 - f)\mathbb{1}_{\{Y>0\}}] = E_Q(1) - E_Q\left(\frac{X}{Y}\right) = 1 - E_P(1) = 0$$

ist

$$E_Q(f\ln(f)) = E_Q(\psi \circ f) - E_Q(1 - f) = E_Q(\psi \circ f) \in [0,\infty]$$

Da für $P(Y > 0) = 1$

$$E_Q(f\ln(f)) = \int_\Omega \mathbb{1}_{\{Y>0\}} \frac{X}{Y} \ln\left(\frac{X}{Y}\right) Y d\mu = \int_\Omega \mathbb{1}_{\{Y>0\}} \ln\left(\frac{X}{Y}\right) X d\mu$$

$$= E_P\left(\ln\left(\frac{X}{Y}\right)\right) = D(P\|Q)$$

ist die relative Entropie wohldefiniert.
Zu (b): gilt $D(P\|Q) = 0$, so nach obiger Berechnung auch:

$$E_Q(\psi \circ f) = E_Q(f\ln(f)) = 0$$

Da Integrand $\psi \circ f \geq 0$, folgt

$$\psi \circ f = 0 \quad Q\text{-f.s.}$$

Da ψ strikt konvex mit einziger Nullstelle $s = 1$, folgt

$$f \equiv 1 \quad Q\text{-f.s.}$$

Da $P \ll Q$, impliziert $Q(A) = 0$ auch $0 = P(A)$, womit

$$f \equiv 1 \quad P\text{-f.s.}$$

Analog zur Rechnung (4.14) zeigt man $P = Q$. □

Bemerkung 4.31 (Dominierendes Mass). Das Maß μ muss nicht endlich sein. Typischerweise verlangt man σ-Endlichkeit, was z.B. das Lebesgue-Maß auf $\mathbb{R}$ oder das Zählmaß auf $\mathbb{N}$ erfüllen. μ heißt *dominierendes Maß*. Dieser Begriff spielt bei *regulären Modellen* in der Statistik ein zentrale Rolle, die relative Entropie beim Beweis der Konsistenz von Maximum-Likelihood-Schätzern.

- Spezialfall für diskrete Verteilungen: $P = X\mu, Q = Y\mu$, μ Zählmaß, mit Gewichtsfunktionen p, q ($\rightarrow$ ursprüngliche Form der Kulback-Leibler-Divergenz). Falls $P(Y = 0) = 0$ gilt:

$$\begin{aligned} D(P\|Q) &= \int_\Omega X \ln\left(\frac{X}{Y}\right) d\mu \\ &= \sum_{\omega\in\Omega} X(\omega) \ln\left(\frac{X(\omega)}{Y(\omega)}\right) \\ &= \sum_{\omega\in\Omega} p(\omega) \ln\left(\frac{p(\omega)}{q(\omega)}\right). \end{aligned}$$

- Spezialfall für reellvertige Verteilungen: $P = Xdx, Q = Ydx$, μ Lebesguemaß auf $\mathbb{R}$, mit Dichtefunktionen X, Y.

Bemerkung 4.32 (Diskussion/Übung).
1. Falls ein zweites Maß $\tilde{\mu}$ existiert, so dass $P = (\tilde{X}\tilde{\mu}), Q = (\tilde{Y}\tilde{\mu})$, so ist $D(P\|Q)$ definiert unabhängig von der Wahl von $\tilde{\mu}$.
2. Beachte ähnliche Struktur zur Varianz: Ist $\psi(x) = x^2$ so ist

$$E(\psi(Y)) - \psi(E(Y)) = E(Y^2) - E(Y)^2.$$

Die Entropie einer ZV definieren wir so:

Definition 4.33. Ist $Y \geq 0$ eine ZVe auf dem W'raum $(\Omega, \mathcal{A}, P)$, so setzen wir wie gehabt

$$\mathrm{Ent}(Y) = E(\phi(Y)) - \phi(E(Y))$$

Dies eine wohldefinierte Zahl in $[0,\infty]$ wegen der schon diskutierten Konvexitäteigenschaften von ϕ.

Bemerkung 4.34. Wir könnern wir als Spezialfall von Definition 4.30 betrachten. Denn gilt zusätzlich $\|Y\| = E(Y) = 1$, so kann man umformen

$$\begin{aligned}\mathrm{Ent}(Y) &= E(Y\ln Y) - E(Z)\ln E(Y)\\ &= E(Y\ln Y)\\ &= \int_\Omega (Y\ln Y)\,dP\\ &= \int_\Omega (Y\ln\frac{Y}{1})\,dP\\ &= D(YP\|P)\end{aligned}$$

Also ist die Entropie von Y bzgl des W'maßes P die K-L-D von P nach YP. In diesem Kontext ist P selbst das dominierende Mass μ.

4.9. **Dualität und Variationsformel.**
Aus der Stochastikvorlesung und auch aus Proposition 3.15 wissen wir, dass man $E(X)$ und $\mathrm{Var}(X)$ durch Optimierungsprobleme charakterisieren kann. Ähnliches gilt für den Entropiebegriff, die KEF oder die Kulback-Leibler-Divergenz:

Satz 4.35. *(Dualitätsformel für Entropie)*
Sei $0 \not\equiv Y \geq 0$ *ZVe auf* $(\Omega,\mathcal{A},P)$*, so dass* $E(|\phi(Y)|) < \infty$*. Dann gilt:*

(I) $\mathrm{Ent}(Y) = \max_{U\in\mathcal{U}} E(UY)$

wobei $\mathcal{U} := \{U\colon \Omega\to\overline{\mathbb{R}}$ *messbar* $\mid E(e^U) = 1\}$

(II) *Erfüllt ZVe* U *die Ungleichung*

$E(UY) \leq \mathrm{Ent}(Y)$

für alle ZVen $Y\colon\Omega\to[0,\infty)$ *mit* $\phi(Y)\in\mathcal{L}^1(P), E(Y) = 1,$

dann gilt: $E(e^U)\leq 1.$

Hier sind Erwartungswert $E(\cdot) = E_P(\cdot)$ und Entropie $\mathrm{Ent}(\cdot) = \mathrm{Ent}_P(\cdot)$ stets bzgl. des Masses P gemeint.

Bemerkung 4.36 (Zur rechten Seite von (I) in Theorem 4.35). Ist die rechte Seite wohldefiniert? Die Berechnung der Legendre-Fenchel-Duale/Konjugierten ergibt:

$$\sup_{x>0}(ux - \phi(x)) = e^{u-1}$$

für alle $u \in \mathbb{R}$. Somit gilt für ZVen U, Y:

$$UY \leq \phi(Y) + \frac{1}{e}e^U \text{ f.s.}$$

da $Y \geq 0$ und somit

$$U\mathbb{1}_{\{U\geq 0\}}Y \leq \phi(Y)\mathbb{1}_{\{U\geq 0\}} + \frac{1}{e}e^U\mathbb{1}_{\{U\geq 0\}} \leq |\phi(Y)| + \frac{1}{e}e^U \text{ f.s.},$$

da $e^U \geq 0$, also

$$E(U_+Y) \leq E(|\phi(Y)|) + \frac{1}{e}E(e^U) < \infty$$

für $U \in \mathcal{U}$. Also ist $U_+Y \in \mathcal{L}^1(P)$ und wir können $E(UY)$ defineren durch:

$$E(UY) := \underbrace{E(U_+Y)}_{\in\mathbb{R}} - \underbrace{E(U_-Y)}_{\in[0,+\infty]} \in [-\infty, \infty).$$

Bemerkung 4.37 (Andere Darstellung der Dualitätsformel (I) in Theorem 4.35).

$$\begin{aligned}\operatorname{Ent}(Y) &= \sup_T E\left[Y\left(\ln(T) - \ln E(T)\right)\right] \\ &= \sup_T E\left[Y\left(\ln(T/E(T))\right)\right],\end{aligned}$$

wobei das Supremum über alle ZVen $0 \leq T \in \mathcal{L}^1(P), T \not\equiv 0$ läuft. (Richtige Interpretation: Übung)

Beweis. Teil (I): Sei U ZVe so ausgewählt, dass $E(e^U) = \int_\Omega e^U dP = 1$, also dass $e^U P$ ein W'maß ist. Berechne Entropie bzgl. W'maßes $e^U P$ und beachte, dass jede Entropie nichtnegativ ist:

$$\begin{aligned}0 &\leq \operatorname{Ent}_{e^U P}(Ye^{-U}) \\ &= \int Ye^{-U}\ln(Ye^{-U})d(e^U P) - \left(\int Ye^{-U}d(e^U P)\right)\ln\left(\int Ye^{-U}d(e^U P)\right) \\ &= \int Y\ln(Ye^{-U})dP - \left(\int YdP\right)\ln\left(\int YdP\right) \\ &= \int \left(Y\ln(Y) - YU\right)dP - \phi(E_P(Y)) \\ &= E_P(\phi(Y)) - \phi(E_P(Y)) - E_P(UY).\end{aligned}$$

Umstellen liefert

$$\operatorname{Ent}(Y) \geq E(UY)$$

Gleichheit wird angenommen für $e^U = Y(E(Y))^{-1}$.
Letztere Funktion ist folgendermasen (wohl)definiert:

Beachte: $Y \geq 0$ und f.s. nicht konstant Null, also $E(Y) > 0$ und $\frac{Y}{E(Y)}$ existiert. Setze

$$U(\omega) = \begin{cases} \ln\left(\frac{Y(\omega)}{E(Y)}\right), & Y(\omega) > 0 \\ -\infty, & Y(\omega) = 0 \end{cases}$$

Einsetzen ergibt:

$$\begin{aligned} E(YU) =& E\left(Y \ln\left(\frac{Y}{E(Y)}\right) \mathbb{1}_{\{Y>0\}}\right) - E\left(Y \cdot \infty \cdot \mathbb{1}_{\{Y=0\}}\right) \\ =& E\left(Y \ln\left(\frac{Y}{E(Y)}\right)\right) = E(Y \ln(Y)) - E(Y \ln(E(Y))) \\ =& E(Y \ln(Y)) - \ln(E(Y))E(Y) \\ =& E(\phi(Y)) - \phi(E(Y)) = \mathrm{Ent}(Y) \end{aligned}$$

Teil (II): Sei $E(UY) \leq \mathrm{Ent}(Y)$ für alle ZVen $Y \colon \Omega \to [0, \infty)$ mit $\phi(Y) \in \mathcal{L}^1(P), E(Y) = 1$. Ist $E(e^U) = 0 \leq 1$, so ist ja die Behauptung erfüllt. Anderfalls muss $E(e^U) > 0$ gelten. Dies ermöglicht es uns, eine optimierende Folge von ZV zu definieren.

Welche Eingeschaften soll sie erfüllen? Wir wollen folgende intuitive Idee verfolgen:

Zu $n \in \mathbb{N}$ sei

$$c_n := E(e^{U \wedge n}) := E(e^{\min(U,n)}) \leq e^n$$

Aus dem Satz über monotone Konvergenz folgt $c_n \to 2\delta := E(e^U) > 0$ für $n \to \infty$. Insbesondere existiert ein $n_0 \in \mathbb{N}$, so dass für alle $n \geq n_0$ gilt $c_n \geq \delta > 0$ und somit

$$Y_n := \frac{1}{c_n} e^{U \wedge n} \quad \text{für } n \geq n_0$$

wohldefiniert und normiert in $\mathcal{L}^1(P)$. Sobald $\phi(Y_n) \in \mathcal{L}^1$ gewährleistet ist (am Ende des Beweises) dürfen wir $E(UY_n) \leq \mathrm{Ent}(Y_n)$ schliessen und weiterverwenden.
Forme r.S. um

$$\mathrm{Ent}(Y_n) = E\left(Y_n(\ln Y_n)\right) = \frac{1}{c_n} E\left(e^{U \wedge n}\left(\ln \frac{e^{U \wedge n}}{c_n}\right)\right)$$

so dass unsere Annahme ergibt:

$$\begin{aligned} E\left[Ue^{U \wedge n}\right] &\leq E\left[e^{U \wedge n}\left(U \wedge n - \ln c_n\right)\right] \\ &= E\left[\underbrace{e^{U \wedge n}}_{\geq 0}\underbrace{(U \wedge n)}_{\leq U}\right] - E\left[e^{U \wedge n}\right] \ln c_n \\ &\leq E\left[Ue^{U \wedge n}\right] - c_n \ln c_n \end{aligned}$$

Kürzen und umstellen ergibt:

$$\underbrace{c_n}_{>0} \ln c_n \leq 0$$

woraus $\ln c_n \leq 0$, somit $c_n \leq 1$ folgt. Mit dem Satz über monotone Konvergenz schliessen wir

$$E\left[e^U\right] = \lim_{n \to \infty} c_n \leq 1$$

wie gewünscht.
Es bleibt noch die Schranke $\phi(Y_n) \in \mathcal{L}^1$ nachzuholen. Dabei nutzt uns Hilfsaussage

$$x > 0 \Longrightarrow\ x > \ln x \Longrightarrow\ -xe^{-x} > 1,$$

für die beidseitige pktw. Abschätzung

$$-1 < (U \wedge n)e^{U \wedge n} \leq ne^n$$

und diese für die pktw. Abschätzung des skalierten Integranden

$$\begin{aligned}0 &\leq c_n|\phi(Y_n)| \\ &= |(U \wedge n)e^{U\wedge n}(U \wedge n) - e^{U\wedge n}\ln(c_n)| \\ &\leq \max(ne^n, |-1|) + e^n \max(n, |\ln\delta|) \leq e^n(2n + |\ln\delta|),\end{aligned}$$

der beschränkt und daher P-intbar ist. □

Mit dem letzten Satz lässt sich eine Brücke zur KEF schlagen, die eine zentrale Rolle bei den Cramér-Chernoff-Schranke in Kapitel 2 spielte.

Korollar 4.38.
Sei $Z\colon (\Omega, \mathcal{A}, P) \to \mathbb{R}$ integrierbare ZVe. Dann gilt

$$\forall\lambda \in \mathbb{R}:\ \psi_{Z-E(Z)}(\lambda) = \sup_{Q \ll P, Q\ W'maß} \left[\lambda(E_Q(Z) - E(Z)) - D(Q\|P)\right],$$

wobei das Supremum über alle bezüglich P absolut-stetigen W'maße Q gebildet wird. Dabei ist weiterhin $E(Z) = E_P(Z)$ der Erwartungswert bezüglich P

Beweis. Kürze ab: $\psi(\lambda) = \psi_{Z-E(Z)}(\lambda)$. Da $Q \ll P$, existiert nach dem Satz von Radon-Nikodym eine Dichte $Y \geq 0$ mit $Q = Y \cdot P$:

$$E(Y) = \int_\Omega Y dP = \int_\Omega dQ = 1,$$

da Q W'maß. Somit Y normiert. Bemerkung 4.34 impliziert für endliche $D(Q\|P)$:

$$E(\phi(Y)) = \mathrm{Ent}(Y) = D(Q\|P) < \infty$$

Setze $U := \lambda(Z - E(Z)) - \psi(\lambda)$. $\implies Ee^U = 1$ nach Defn. der KEF. Dann gilt im Fall $D(Q\|P) < \infty$:

$$\begin{aligned}D(Q\|P) = \mathrm{Ent}(Y) &\overset{4.35(I)}{\geq} E(YU) = E(Y\lambda(Z - E(Z))) - E(Y\psi(\lambda)) \\ &= \lambda(E_Q(Z) - \underbrace{E(Y)}_{=1}E(Z)) - \psi(\lambda)\underbrace{E(Y)}_{=1} \\ \Leftrightarrow \quad \psi(\lambda) &\geq \lambda(E_Q(Z) - E(Z)) - D(Q\|P).\end{aligned}$$

Dies gilt für alle normierten Maße $Q \ll P$ mit $D(Q\|P) < \infty$ (die anderen sind für das Supremum uninteressant), also

$$\psi_{Z-E(Z)}(\lambda) \geq \sup_{Q \ll P, Q\ \mathrm{W'maß}} (\lambda(E_Q(Z) - E(Z)) - D(Q\|P))$$

Nun wird die komplementäre Ungleichung bewiesen. Setze diesmal:

$$U := \lambda(Z - E(Z)) - \sup_{R \ll P, R \text{ W'maß}} \left(\lambda(E_R(Z) - E(Z)) - D(R\|P)\right).$$

Sei $Y \geq 0$ ZVe mit $E(Y) = 1$ und $\phi(Y) \in \mathcal{L}^1$ (Testfunktion) und $Q = YP$. Dann gilt:

$$\begin{aligned} E(YU) &= \lambda E[Y(Z - E(Z))] - E\left[Y \sup_{R \ll P, R \text{ W'maß}} \left(\lambda(E_R(Z) - E(Z)) - D(R\|P)\right)\right] \\ &\leq \lambda(E_Q(Z) - E(Z)) - \lambda\left(E_Q(Z) - E(Z)\right) + D(Q\|P) \\ &= D(Q\|P) = \mathrm{Ent}(Y), \end{aligned}$$

da Y normiert. Hier haben wir also im Supreprum konkret $R = Q \ll P$ eingesetzt.
Aus Teil (II) von Theorem 4.35 (Rückrichtung) folgern wir nun $E(e^U) \leq 1$ und setzen ein

$$\begin{aligned} &\Rightarrow \; E\left(e^{\lambda(Z-E(Z))} \exp\left(-\sup_{R \ll P, R \text{ W'maß}} \left(\lambda(E_R(Z) - E(Z) - D(R\|P))\right)\right)\right) \leq 1 \\ &\Rightarrow \; E\left(e^{\lambda(Z-E(Z))}\right) \leq \exp\left(\sup_{R \ll P, R \text{ W'maß}} \left(\lambda(E_R(Z) - E(Z) - D(R\|P))\right)\right) \\ &\Rightarrow \; \psi(\lambda) \leq \sup_{R \ll P, R \text{ W'maß}} \left(\lambda(E_R(Z) - E(Z) - D(R\|P)\right). \end{aligned}$$

□

Durch eine ‚Umstellung' erhalten wir eine Dualitaetsformel für die K-L-D.

Korollar 4.39. *Seien P, Q zwei W'Maße auf $(\Omega, \mathcal{A})$. Dann gilt:*

$$D(Q\|P) = \sup_Z \left[E_Q Z - \ln E\left(e^Z\right)\right], \tag{4.15}$$

wobei das Supremun ueber alle ZV $Z \colon \Omega \to \mathbb{R}$ mit $E\left(e^Z\right) < \infty$ gebildet wird.

Kurz: Bei fixem P ist $D(Q\|P)$ das konvexe Dual zu Abbildung $Z \to E\left(e^Z\right)$.

Beweis. Unterscheiden zwei Faelle: Gilt $Q \ll P$, so gibt es Radon-Nikodym-Dichte $Y := \frac{dQ}{dP}$. Diese ist normiert $E(Y) = \int Y dP = \int 1 dQ = 1$. Bemerkungen 4.34 und 4.37 geben:

$$\begin{aligned} D(Q\|P) &= \sup_{0\le T\in\mathcal{L}^1} E\left[Y(\ln T - \ln(ET))\right] \\ &= \sup_{0\le T\in\mathcal{L}^1} \left[E_Q(\ln T - (EY)\ln(ET)\right] \\ &= \sup_{0\le T\in\mathcal{L}^1} \left[E_Q(\ln T - \ln(ET)\right] \end{aligned}$$

wg. Normierung von Y. Nun führen wir Substitution $Z = \ln T$, also $e^Z = T$ ein, so dass

$$E(e^Z) = E(T) < \infty$$

Also

$$D(Q\|P) = \sup_{Z \text{ ZV mit } e^Z\in\mathcal{L}^1} \left[E_Q(Z - \ln(Ee^Z)\right]$$

Gilt $Q \ll P$ nicht, so gibt es ein Ereignis A mit $Q(A) > 0$ aber $P(A) = 0$ und nach Defn. gilt $D(Q\|P) = \infty$. Andererseits ergibt Einsetzten von $Z_n := n\mathbb{1}_A$ in r.S. (4.15)

$$\begin{aligned} & nE_Q(\mathbb{1}_A) - \ln E\left[e^{n\mathbb{1}_A}\right] \\ =& nQ(A) - \ln\left[\underbrace{P(A)}_{=0} e^n + P(A^c)\cdot 1\right] \\ =& nQ(A) - \ln 1 = nQ(A) \to \infty \text{ falls } n\to\infty. \end{aligned}$$

Demnach sind beide Seiten in (4.15) $+\infty$. □

Bemerkung 4.40 (Verallgemeinerung von Bekanntem:). der Erwartungswert minimiert den mittleren quadratischen Abstand einer deterministishen Zahl zum zufaelligen Wert einer ZV. Dies ist ein Speziallfall von allgemeinerem Prinzip:

Satz 4.41. *Sei $I \subset \mathbb{R}$ offenes Intervall und $f\colon I \to \mathbb{R}$ konvex und diffbar. Sei $X\colon \Omega \to I$ eine ZV auf $(\Omega, \mathcal{A}, P)$. Dann*

$$E\left[f(X) - f(EX)\right] = \inf_{a\in I} E\left[f(X) - f(a) - f'(a)(X-a)\right]$$

Beweis. Buch von Boucheron, Lugosi und Massart/Praesenzuebung □

Bemerkung 4.42 (Bregman-Divergenz). Für solches f wird für $x, y \in I$

$$g_x(y) := \left[f(y) - f(x) - f'(x)(y-x)\right]$$

die *Bregman-Divergenz* von x nach y genannt. Für konvexe f ist sie nicht-negativ.

Wir fuehren den erwaehnten, bekannten Fall des EW auf den Satz zurueck: Sei $f(x) = x^2$, dann $f'(x) = x2$, $f'(a) = 2a$. Betrachte r.S.

$f(x)-f(a)-f'(a)(x-a) = x^2-a^2-2a(x-a) = x^2-a^2-2ax+2a^2 = (x-a)^2$

Nun l.S.

$$\begin{aligned} E\left[f(X) - f(EX)\right] &= E\left[X^2 - (EX)^2\right] \\ &= E\left[X^2\right] - (EX)^2 \\ &= E\left[X^2\right] - 2(EX)^2 + (EX)^2 \\ &= E\left[X^2 - 2X(EX) + (EX)^2\right] \end{aligned}$$

Einsetzten von $f(x) = \phi(x) = x\ln(x)$ gibt uns ein Resultat für die Entropie:

Korollar 4.43. *Sei $Y\colon \Omega \to [0,\infty)$ ZV, s.d. $E[\phi(Y)] < \infty$. Da ϕ konvex und diffbar ist:*

$$\mathrm{Ent}(Y) = \inf_{u>0} E\left[Y(\ln Y - \ln u) - (Y-u)\right]$$

4.10. **Transportkosten-Abschätzung.**
Das Transportkosten-Problem zwischen zwei W'maßen P und Q kann man auf folgende Weise veranschaulichen:

ABBILDUNG 7. Dynamisches Bild für das Transportkosten-Problem

Ansatz der Analysis von partiellen Differentialgleichungen: Formuliere Zeitabhängige Differentialgleichung im Zeitintervall $[0, 1]$. Die Lösung soll als Anfangswert die Dichtefunktion von P und als Endwert die Dichtefunktion von Q haben.
Ansatz der Theorie stochastischer Prozesse: Konstruiere interpolierenden stochastischen Prozess im Zeitintervall $[0, 1]$. Die Verteilung des Prozesses zum Ztpkt. 0 bzw. 1 soll P bzw. Q sein.

ABBILDUNG 8. Statisches Bild für das Transportkosten-Problem und die Kopplung

Bemerkung 4.44 (Masstheoretischer Rahmen). Sei (M, d) ein metrischer Raum, μ, ν Borel-Masse auf (M, d) sowie

$$K(\mu,\nu) = \{\kappa \text{ Mass auf } (M \times M, \mathcal{B} \otimes \mathcal{B}) \mid \text{Randmaß von } \kappa \text{ auf 1. Koordinate ist } \mu, \text{ Randmaß von } \kappa \text{ auf 2. Koordinate ist } \nu \}$$

Sei $p \geq 1$. Im folgenden nehmen wir folgende Momentenbedingung an: Es existiert ein $x_o \in M$ so, dass beide Masse $\rho \in \{\mu, \nu\}$

$$\int_M d(x, x_o)^p \rho(dx) < \infty$$

erfüllen. Dann definiere die Wasserstein-Metrik

$$W_p(\mu, \nu) = \inf_{\kappa \in K(\mu,\nu)} \left(\int_{M \times M} d(x, y)^p \kappa(dx, dy) \right)^{1/p}$$

was man im Fall, dass μ, ν Verteilungen der M-wertigen ZV X und Y, also insbes. W'maße sind, schreiben kann als

$$= \inf_{\kappa \in K(\mu,\nu)} [(\mathbb{E}_\kappa(d(X, Y)^p)]^{1/p}$$

In den Anwendungen sind die Fälle $p = 1$ und $p = 2$ besonders interessant. Im Fall $p = 1$ und zweier Borel-W'maße $\mu = Q$, $\nu = P$ auf einem metrischen Raum $M = \Omega$ haben wir die duale Kantorovich-Charakterisierung

$$W_1(Q, P) = \sup_Z (E_Q(Z) - E_P(Z))$$

wobei das Supremum über alle Funktionen $Z \colon M \to \mathbb{R}$, die 1-*Lipschitz* sind, d.h. die Bedingung

$$\forall \omega, \tilde{\omega} \in M : \ |Z(\omega) - Z(\tilde{\omega})| \leq d(\omega, \tilde{\omega})$$

erfüllen, genommen wird.
Bisher haben wir nur die Definition der relevanten Größe gegeben. Kann man effiziente obere Schranken angeben?
Anwendungen:

- Sub-Gausssche Konzentrationsungleichungen.
- Kontraktionseigenschaften von Evolutionsgleichungen (Lösungen driften unter der zeitlichen Entwicklung nicht auseinander.)

Bei dem folgenden Lemma sollte man bei der Funktion ξ an eine kumulantenerzeugenden Funktion denken. In der Tat, im Korollar speziallisieren wir uns auf die KEF einer normalverteilten ZV.

Lemma 4.45. *Sei $Z \colon (\Omega, P) \to \mathbb{R}$ intbare ZV, $0 < b \leq \infty$, $\xi \colon [0, b) \to \mathbb{R}$, $\xi \in C^1$, konvex, $\xi(0) = 0 = \xi'(0)$. Setze für $x \geq 0$:*

$$\xi^*(x) := \sup_{\lambda \in (0,b)} (\lambda x - \xi(\lambda))$$

und für $t \geq 0$

$$(\xi^*)^{-1}(t) := \inf\{x \geq 0 \mid \xi^*(x) > t\}$$

Dann sind äquivalent

(i) $\forall \lambda \in (0,b): \quad \psi_{Z-EZ} := \ln E\left(e^{\lambda(Z-EZ)}\right) \leq \xi(\lambda)$

(ii) für alle W'maße $Q \ll P$ *mit* $D(Q\|P) < \infty$ *gilt:*

$$E_Q(Z) - E_P(Z) \leq (\xi^*)^{-1}\left(D(Q\|P)\right)$$

Spezialfall Sub-Gaußsche Zufallsvariablen:

Korollar 4.46. *Sei* $\nu > 0$ *fest. Dann sind äquivalent:*

(i)

$$\forall \lambda > 0: \quad \psi_{Z-EZ}(\lambda) \leq \frac{\nu\lambda^2}{2}$$

(ii) für alle W'maße $Q \ll P$ *mit* $D(Q\|P) < \infty$ *gilt:*

$$E_Q(Z) - E_P(Z) \leq \sqrt{2\nu\left(D(Q\|P)\right)}$$

Beweis vom Korollar. Ist $\xi(\lambda) = \frac{\nu\lambda^2}{2}$, so $(\xi^*)^{-1}(t) = \sqrt{2\nu t}$. Einsetzen ergibt die Behauptung. □

Beweis vom Lemma. Sei (i) wahr.
Die Aussage ist nach Korollar 4.38 äquivalent zu

$$\forall \lambda \in (0,b): \sup_{Q \ll P, Q \text{ W'maß}} [\lambda(E_Q(Z) - E(Z)) - D(Q\|P)] \leq \xi(\lambda)$$

und dies wiederum äquivalent zu

$$\forall \text{ W'maße } Q \ll P: \; E_Q(Z) - E(Z) \leq \inf_{\lambda \in (0,b)} \left[\frac{1}{\lambda}(\xi(\lambda) + D(Q\|P))\right]$$

Andererseits folgt aus der Charakterisierung der Pseudo-Inversen in Lemma 2.14

$$(\xi^*)^{-1}\left(D(Q\|P)\right) = \inf_{\lambda \in (0,b)} \frac{\xi(\lambda) + D(Q\|P)}{\lambda}$$

falls $D(Q\|P)) < \infty$. Hierbei nutzen wir die an ξ geforderten Eigenschaften. Letzeres ergibt gerade Aussage (ii). □

Definition 4.47. Sei $\Omega = M$ ein metrischer Raum mit W'maß P auf der Borel-σ-Algebra und $\nu > 0$. Man sagt, dass P eine *Transportkostenabschätzung* vom Typ $T_p(\nu)$ erfüllt, falls

$$\sup_{Q \text{ W'maß}} (W_p(Q,P)) \leq \sqrt{2\nu\left(D(Q\|P)\right)}$$

Im Fall $p = 1$ ist dies ist ´aquivalent zu

$$\sup_{Q \text{ W'maß } Z\colon M \to \mathbb{R}1-\text{Lipschitz}} \sup (E_Q(Z) - E_P(Z)) \leq \sqrt{2\nu\,(D(Q\|P))}$$

Solche Ungleichungen stellen ein interessante Verbindung zwischen Statistik und Stochastik auf der einen und Differentialgleichungen und Differentialgeometrie auf der anderen Seite her.
Nun folgt einen Anwendung der Transportkosten-Ungleichung:

4.11. **Pinsker-Ungleichung.**
Seien P, Q W'maße auf $(\Omega, \mathcal{A})$. Stellen nun eine Beziehung her zwischen

- relativer Entropie von P nach Q (keine echte Metrik, da asymmetrisch)
  und
- totaler Variationsnorm von $P - Q$ (echte Metrik)

$$V(P, Q) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$$

Besitzen sowohl P als auch Q ein Radon-Nikodym-Dichte bzgl. eines gemeinsamen dominierenden σ-endlichen Masses μ auf Ω, so gilt

$$V(P, Q) = P(A_+) - Q(A_+) = \frac{1}{2} \int_\Omega |p(x) - q(x)|\, \mu(dx)$$

wobei $p := \frac{dP}{d\mu}$ und $q := \frac{dQ}{d\mu}$ die Dichten bzgl. dominierenden Masses sind und $A_+ := \{x \in \Omega \mid p(x) \geq q(x)\}$ ist. Also

$$V(P, Q) = \frac{1}{2} \|p - q\|_{L^1(\Omega,\mu)}$$

Setzen wir $A_- = \{x \in \Omega \mid p(x) \leq q(x)\}$ so folgt:

$$Q(A_-) - P(A_-) = 1 - Q(A_+) - (1 - P(A_+)) = P(A_+) - Q(A_+)$$

was (nochmal) die Symmetrie der Funktion $V(\cdot, \cdot)$ bestätigt.

Bemerkung 4.48 (Aufgabe). Sei

$$K(P, Q) = \{\kappa \text{ W'mass auf } \mathcal{A} \otimes \mathcal{A} \mid \text{ Randmaß von } \kappa \text{ auf 1. Koordinate ist } P, \text{ Randmaß von } \kappa \text{ auf 2. Koordinate ist } Q\}$$

so gilt

$$V(P, Q) = \min_{\kappa \in K(P,Q)} \kappa(X \neq Y)$$

falls X, Y ZV mit $X \sim P$ und $Y \sim Q$.

Die folgende Ungleichung von Csiszar, Kullback und Pinsker ist sozusagen der 'Urahn aller Transportkostenungleichungen'.

Satz 4.49 (Pinsker Ungleichung)**.** *Seien P, Q W'maße auf $(\Omega, \mathcal{A})$ mit $Q \ll P$. Dann gilt*

$$V(P,Q)^2 \leq \frac{1}{2} D(Q\|P)$$

Damit wird die qualitative Aussage

$$D(Q\|P) = 0 \Longrightarrow\ Q = P$$

zu der quantitativen

$$D(Q\|P) \leq \epsilon \Longrightarrow\ V(P,Q) \leq \sqrt{\frac{\epsilon}{2}}$$

Beweis. Sei $Y\colon \Omega \to \mathbb{R}$ die Dichte $Y = \frac{dQ}{dP}$,

$$A_- = \{x \in \Omega \mid Y(x) \geq 1\}$$

das , fast sichere Komplement' der optimierende Menge A_+ in der Definition von $V(P,Q)$ und $Z = \mathbb{1}_{A_-}$. Dann folgt

$$V(P,Q) = V(Q,P) = Q(A_-) - P(A_-) = E_Q(Z) - E_P(Z)$$

Nun wollen wir das Hoeffding-Lemma 2.11 anwenden. Dazu stellen wir fest: Für $\tilde{Z} := Z - E(Z) = Z - P(A_-)$ gilt $\tilde{Z} \in [-P(A_-), 1 - P(A_-)] = [a,b]$ mit Intervallbreite $b - a = 1$. Insbesondere ist $\tilde{Z}$ sub-gaußsch, genauer $\tilde{Z} \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right) = \mathcal{G}\left(\frac{1}{4}\right)$, also

$$\psi_{\tilde{Z}}(\lambda) \leq \frac{1}{4}\frac{\lambda^2}{2} = \frac{\lambda^2}{8}$$

Da $Q \ll P$, ist nun Lemma 4.45 anwendbar und wir erhalten

$$V(P,Q) = E_Q(Z) - E_P(Z) \leq \sqrt{\frac{1}{2} D(Q\|P)}$$

□

Anwendung der Pinsker-Ungleichung in der Statistik: *Untere* Schranke für Fehler bei gewissen Hypothesentest.
Im Kontext der Cramér-Chernoff-Schranke in Kapitel 2 ist die Relation

$$\forall\, c > 0 : \quad e^{-cD(Q\|P)} \leq e^{-cV(Q,P)^2} \tag{4.16}$$

wegen der Variationsdarstellung der KEF in Korollar 4.38 nüzlich. Ähnlich kann Relation (4.16) beim *Large Deviations Principle* angewendet werden.

4.12. **Birgé-Ungleichung.** ist eine Verschärfung der Pinsker-Ungleichung. Betrachte zwei Bernoulli-Verteilungen P, Q auf $\mathcal{P}(\{0,1\})$ mit Erfolgsparameter p bzw. $q \in [0,1]$. Dann ist deren relative Entropie gegeben durch

$$D(Q\|P) = h_p(q) = h(q,p) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} \tag{4.17}$$

Rechne nach und vgl. Definition 2.2!

Satz 4.50. *Seien $Q \ll P$ zwei W'maße auf $(\Omega, \mathcal{A})$. Dann*

$$\sup_{A \in \mathcal{A}} h\left(Q(A), P(A)\right) \leq D(Q\|P) \tag{4.18}$$

Beweis. Für $p \in [0,1]$ ist

$$\mathbb{R} \ni \psi_p(\lambda) := \ln\Big(p(e^\lambda - 1) + 1\Big)$$

die KEF der Bernoulli-Verteilung $\mathcal{B}_{1,p}$. Korollar 4.39 besagt

$$D(Q\|P) = \sup_{Z\colon \Omega \to \mathbb{R}, E(e^Z) < \infty} \left[E_Q Z - \ln E_P\left(e^Z\right)\right]$$

w'ahle $\lambda \geq 0, A \in \mathcal{A}, Z = \lambda \mathbb{1}_A \implies$ Bernoulli-verteilt

$$\geq \left[E_Q(\lambda \mathbb{1}_A) - \ln E\left(e^{\lambda \mathbb{1}_A}\right)\right].$$

Da $\lambda \geq 0$, folgt:

$$D(Q\|P) \geq \sup_{\lambda \geq 0} \left[\lambda Q(A) - \psi_{P(A)}(\lambda)\right].$$

mit der eben eingeführten KEF der Bernoulli-Verteilung. Deren Cramér-Transformierte bzw. Fenchel-Legendre-Duale haben wir in §2.2.3 berechnet und gezeigt:

$$\forall\, q \in [0,1]: \quad h(q,p) = \sup_{\lambda > 0} \left[\lambda q - \psi_p(\lambda)\right].$$

Wegen Stetigkeit bei $\lambda = 0$ folgt insgesamt:

$$D(Q\|P) \geq h(Q(A), P(A))$$

□

(Eine alternative Herleitung basiert auf dem *data processing lemma*).

Rechenaufgabe:

$$2(q-p)^2 \leq q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} =: h(q,p)$$

Damit ist die Schranke (4.18) stärker als Pinsker-Ungleichung.
Nun eine multivariate Erweiterung:

Satz 4.51 (Birgé-Ungleichung)**.** *Seien $P_0, P_1, \ldots, P_N$ W'mase auf $(\Omega, \mathcal{A})$ und $A_0, A_1, \ldots, A_N \in \mathcal{A}$ disjunkte Teilmengen. Gilt*

$$a := \min\{P_0(A_0), P_1(A_1), \ldots, P_N(A_N)\} \geq \frac{1}{N+1} \tag{4.19}$$

so folgt

$$h\left(a, \frac{1-a}{N}\right) \leq \frac{1}{N}\sum_{j=1}^{N} D(P_j \| P_0) \tag{4.20}$$

Bedingung (4.19) kann im Kontext des *neidfreien Teilens* interpretiert werden: $N+1$ Personen sollen sich Ω teilen. Das W'maß P_j beschreibt die Präferenz/Gewichtung der j-ten Person. Also bedeutet $a \geq \frac{1}{N+1}$, dass keine Person denkt, dass das ihr zugeteilte Stück A_j unterproportional ist. (Die Stücke sind natürlich disjunkt.)
Die Aussage impliziert nun, dass dann die 'Schwankung in der Zufriedenheit' beschräkt ist durch die Schwankungen in den Präferenzmaßen. (Dazu betrachte insbesondere den Fall, dass alle gleich P_j sind.

Beweis. Aus der variationellen Darstelleung der K-L-D in Korollar 4.39 folgt für jedes $j = 1, \ldots, N$ und $\lambda \geq 0$

$$D(P_j \| P_0) = \sup_{Z\colon \Omega \to \mathbb{R}, E(e^Z) < \infty} \left[E_{P_j} Z - \ln E_{P_0}\left(e^Z\right)\right]$$

treffe spezielle Wahl $Z = \lambda \mathbb{1}_{A_j}$

$$\geq \left[E_{P_j}(\lambda \mathbb{1}_{A_j}) - \ln E_{P_0}\left(e^{\lambda \mathbb{1}_{A_j}}\right)\right].$$

Disjunktheit impliziert $A_1 \cup \ldots \cup A_N \subset A_0^c$, also

$$\sum_{j=1}^{N} P_0(A_j) \leq 1 - P_0(A_0) \leq 1 - a$$

Die beiden letzten Ungleichungen, sowie die Jensen-Ungleichung und die Konkavität des ln nutzen wir nun in

$$\begin{aligned}\frac{1}{N}\sum_{j=1}^{N} D(P_j\|P_0) &\geq \frac{1}{N}\sum_{j=1}^{N}\left[\lambda P_j(A_j) - \ln E_{P_0}\left(e^{\lambda \mathbb{1}_{A_j}}\right)\right] \\ &\geq \frac{1}{N}\sum_{j=1}^{N}\lambda a - \frac{1}{N}\sum_{j=1}^{N}\left[\ln\left(P_0(A_j)(e^{\lambda}-1)+1\right)\right] \\ &\geq \lambda a - \ln\left[\frac{1}{N}\sum_{j=1}^{N}\left(P_0(A_j)(e^{\lambda}-1)+1\right)\right] \\ &\geq \lambda a - \ln\left[(e^{\lambda}-1)\frac{1-a}{N}+1\right]\end{aligned}$$

Diese Ungleichung gilt für alle $\lambda \geq 0$, also ist auch

$$\sup_{\lambda>0}\left[\lambda a - \ln\left((e^{\lambda}-1)\frac{1-a}{N}+1\right)\right] = \sup_{\lambda>0}\left[\lambda a - \psi_{\frac{1-a}{N}}(\lambda)\right] = h\left(a, \frac{1-a}{N}\right)$$

eine untere Schranke wie behauptet. ($\psi_{\frac{1-a}{N}}(\lambda)$ steht für die KEF der Bernoulli-Verteilung $\mathcal{B}_{\frac{1-a}{N}}$.) □

Wir diskutieren noch einige Ungleichungen in diesem Kontext, deren Beweise als Aufgaben überlassen werden. Folgendes Korollar findet bei unteren Schranken an Schätzfehler Anwendung in der Statistik.

Korollar 4.52. *Seien $P_0, P_1, \ldots, P_N$ W'maße auf $(\Omega, \mathcal{A})$, $A_0, A_1, \ldots, A_N \in \mathcal{A}$ disjunkte Teilmengen, und $a := \min\{P_0(A_0), P_1(A_1), \ldots, P_N(A_N)\}$. Dann gilt*

$$a \leq \max\left\{\frac{2e}{2e+1}, \frac{1}{N\ln(1+N)}\sum_{j=1}^{N} D(P_i\|P_0)\right\} \tag{4.21}$$

Ungleichung ist nicht-trivial, da $\frac{2e}{2e+1} < 1$.
Eine verfeinerte Variante der Birgé-Ungleichung impliziert auch formal die Pinsker-Ungleichung:

Satz 4.53 (Verfeinerte Birgé-Ungleichung)**.** *Seien $P_0, P_1, \ldots, P_N$ W'mase auf $(\Omega, \mathcal{A})$, $A_0, A_1, \ldots, A_N \in \mathcal{A}$ disjunkte Teilmengen, $a_0 := P_0(A_0)$ und $a := \min\{P_1(A_1), \ldots, P_N(A_N)\}$. Gilt*

$$a \geq \frac{1-a_0}{N} \tag{4.22}$$

so folgt

$$h\left(a, \frac{1-a_0}{N}\right) \leq \frac{1}{N}\sum_{j=1}^{N} D(P_i \| P_0) \tag{4.23}$$

Aufgabe: leite daraus die Pinsker-Ungleichung her!
Auch die Fano-Ungleichungen aus der Informationstheorie lässt sich aus der Birge-Ungleichung herleiten (Aufgabe).

Lemma 4.54. *Sei* $h_0(p) = p\ln p + (1-p)\ln(1-p)$, $N \geq 2$, $\mathcal{X} = \{x_1, \ldots, x_N\}$ *und* $\mathcal{Y} = \{y_1, \ldots, y_N\}$.

(1) Sei $X\colon \Omega \to \mathcal{X}$ *ein ZV und* $x_0 \in \mathcal{X}$ *mit* $P(X = x_0) = p \in (0,1)$. *Dann gilt für die Shannon-Entropie*

$$H(X) \leq -h_0(p) + (1-p)\ln(N-1)$$

(2) Sei $Y\colon \Omega \to \mathcal{Y}$. *Gilt* $\sum_{j=1}^{N} P(X = x_j, Y = y_j) = p \in (0,1)$. *Dann gilt für die bedingte Shannon-Entropie*

$$H(X|Y) \leq -h_0(p) + (1-p)\ln(N-1)$$

4.13. **Subadditivität der Entropie für allgemeine Zufallsvariablen.** Im Kontext diskreter ZV haben wir die Subadditivität der Entropie aus der Han-Ungleichung hergeleitet. Nun leiten wir sie im allgemeinen Fall aus der Dualitätsformel 4.35 und einer ähnlichen Zerlegung, wie sie uns schon bei der E-S-U begegnet ist, her. Wie gehabt sei $\phi\colon [0,\infty) \to \mathbb{R}, \phi(x) = x\ln(x)$ für $x > 0$ und $\phi(x) = 0$ für $x = 0$.

Satz 4.55. *Seien* $X_1, \ldots, X_n$ *unabh. ZV und* $Y = f(X_1, \ldots, X_n) \geq 0$ *messbare Funktion der* $X_1, \ldots, X_n$, *so dass* $\phi(Y) \in \mathcal{L}^1(P)$. *Für* $i \in \{1, \ldots, n\}$ *bezeichne wie gehabt*

$$\begin{aligned} X^{(i)} &:= (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ E^{(i)}(Y) &:= E\left(Y \mid X^{(i)}\right) \\ \mathrm{Ent}^{(i)}(Y) &:= E^{(i)}(\phi(Y)) - \phi(E^{(i)}(Y)) \end{aligned}$$

Dann gilt

$$\mathrm{Ent}(Y) \leq \sum_{i=1}^{n} E\left[\mathrm{Ent}^{(i)}(Y)\right]$$

Beweis. Wir brauchen noch

$$E_i(Y) := E(Y \mid X_1, \ldots, X_i) \text{ für } i = 1, \ldots, n,$$

und

$$E_0(Y) := E(Y).$$

Auf dem Raum der $\sigma(X_1, \ldots, X_i)$-messbaren und intbaren ZV ist E_i die Identität. Teleskopieren ergibt

$$Y[\ln Y - \ln(EY)] = \sum_{i=1}^{N} Y[\ln(E_i Y) - \ln(E_{i-1} Y)]$$

Dualitätsformel aus Bemerkung 4.37 gibt

$$\begin{aligned} \mathrm{Ent}^{(i)}(Y) &= \sup_{0 \le T \in \mathcal{L}} E^{(i)} \left[Y \left(\ln(T) - \ln E^{(i)}(T) \right) \right] \\ &\ge E^{(i)} \left[Y \left(\ln(E_i Y) - \ln E^{(i)}(E_i Y) \right) \right] \end{aligned}$$

Da $X_1, \ldots, X_n$ unabh., folgt $E^{(i)}(E_i(Y)) = E_{i-1}(Y)$ und die Turmeigenschaft gibt

$$\begin{aligned} E\left[Y[\ln Y - \ln(EY)]\right] &= \sum_{i=1}^{N} E\left[Y[\ln(E_i Y) - \ln(E_{i-1} Y)]\right] \\ &= \sum_{i=1}^{N} E\ E^{(i)}\left[Y[\ln(E_i Y) - \ln(E_{i-1} Y)]\right] \\ &\le \sum_{i=1}^{N} E\ \left[\mathrm{Ent}^{(i)}(Y)\right] \end{aligned}$$

□

Bemerkung 4.56. Für ZV $X_1, \ldots, X_n \colon (\Omega, P) \to \mathcal{X}$ mit endlichem $|\mathcal{X}|$ sind die Han-Ungleichung

$$H(X_1, \ldots, X_n) \le \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

und die Subadditivität

$$\mathrm{Ent}(Y) \le \sum_{i=1}^{n} E\left[\mathrm{Ent}^{(i)}(Y)\right]$$

sogar äquivalent. (Hausaufgabe)

4.14. **Brunn-Minkowski-Ungleichung.** hat einen maßtheoretisch-geometrischen Flavour. Sie spielt aber nicht nur in der Analysis, sondern auch der Informationstheorie, Geometrie und Algebra einen wichtige Rolle. Die Isoperimetrische Ungleichung im $\mathbb{R}^d$ kann mit Hilfe der Brunn-Minkowsky-Ungleichung beweisen werden, ebenso wie folgendes prominente Resultat über die Konzentration aus hochdimanesionalen Sphären:

Satz 4.57 (Concentration of measure phenomenon on the sphere)**.** *Sei $\mathbb{S} \subset \mathbb{R}^n$ die Einheitssphäre, $X \subset \mathbb{S}$ messbar, d der Euklidische Abstand in $\mathbb{R}^n$, $t > 0$ und $X_t := \{x \in S \mid d(x, X) \leq t\}$. Bezeichne σ das Oberflächenmaß auf $\mathbb{S}$. Dann gilt*

$$\forall\ t \in (0,1]: \quad \frac{\sigma(X_t)}{\sigma(\mathbb{S})} \geq 1 - \frac{\sigma(\mathbb{S})}{\sigma(X)} \exp\left(-\frac{nt^2}{4}\right)$$

Für zwei Mengen $A, B \subset \mathbb{R}^d$ sei die Minkowsky-Summe definiert durch

$$A + B := \{x + y \mid x \in A, y \in B\}$$

und für $t > 0$ sei

$$tA := \{tx \mid x \in A\}$$

Für meßbares $A \subset \mathbb{R}^d$ sei $\mathrm{vol}(A)$ das Lebesgue-Maß von A.

Satz 4.58 (Brunn-Minkowsky Ungleichung)**.** *Seien $A, B \subset \mathbb{R}^d$ nichtleere, kompakte Mengen und $t \in [0,1]$. Dann gilt*

$$\mathrm{vol}\left((1-t)A + tB\right)^{1/d} \geq (1-t)\mathrm{vol}\left(A\right)^{1/d} + t\mathrm{vol}\left(B\right)^{1/d}$$

Man kann die Voraussetzung ‚A kompakt' durch ‚A meßbar' ersetzen, muß dann aber fordert, dass $A+B$ ebenfalls meßbar ist. I.a. ist dies nämlich nicht automatisch der Fall, wobei es noch mal einen Unterschied macht, ob man über Lebesgue- oder Borel-Meßbarkeit spricht.
Warum ist unter der Kompaktheitsvoraussetzung $A + B$ meßbar?
Die Umgekehrte Ungleichung

$$\mathrm{vol}\left((1-t)A + tB\right)^{1/d} \leq (1-t)\mathrm{vol}\left(A\right)^{1/d} + t\mathrm{vol}\left(B\right)^{1/d}$$

gilt i.a. **nicht**, allerdings gibt es eine modifizierte Variante von V. Milman.

Motivation: Beweis im Fall $d = 1$. Wir beweisen zuerst Subadditivität und dann Homogenität.

Für $c, d \in \mathbb{R}$ und $A, B \subset \mathbb{R}$ kompakt gilt:

$$\begin{aligned} \operatorname{vol}(A + c) &= \operatorname{vol}(A) \\ \operatorname{vol}(B + d) &= \operatorname{vol}(B) \\ \operatorname{vol}(A + c + B + d) &= \{(x + y) + c + d \mid x \in A, y \in B\} = \operatorname{vol}(A + B) \end{aligned}$$

Wir wählen

$$c := -\max A \quad \& \quad d := -\min B$$

ABBILDUNG 9. $\tilde{A}$ und $\tilde{B}$ berühren sich in 0.

Dann folgen

$$\tilde{A} := A + c \subset (-\infty, 0], \quad \tilde{B} := B + d \subset [0, \infty), \quad \tilde{A} \cap \tilde{B} = \{0\}$$

Aus der letzten Aussage folgt wiederum

$$\tilde{A} \subset \tilde{A} + \tilde{B} \quad \& \quad \tilde{B} \subset \tilde{A} + \tilde{B}$$

und damit

$$\tilde{A} \cup \tilde{B} \subset \tilde{A} + \tilde{B}$$

Der Schnitt $\tilde{A} \cap \tilde{B} = \{0\}$ ist eine Nullmenge, daher

$$\operatorname{vol}\left(\tilde{A}\right) + \operatorname{vol}\left(\tilde{B}\right) = \operatorname{vol}\left(\tilde{A} \cup \tilde{B}\right) \leq \operatorname{vol}\left(\tilde{A} + \tilde{B}\right),$$

womit die Subadditivität bewiesen wäre.
Ist $t > 0$, so $\operatorname{vol}(tA) = t\operatorname{vol}(A)$. Mit obiger Aussage folgt insgesamt:

$$\operatorname{vol}\left((1-t)A + tB\right) \geq \operatorname{vol}\left((1-t)A\right) + \operatorname{vol}\left(tB\right) = (1-t)\operatorname{vol}\left(A\right) + t\operatorname{vol}\left(B\right)$$

□

Im mehrdimensionalen Fall kann man die Brunn-Minkowsky-Ungleichung beweisen mit Hilfe

Satz 4.59 (Prékopa-Leinder-Ungleichung)**.** *Seien* $t \in (0,1)$*,* $f, g, h \colon \mathbb{R}^d \to [0, \infty)$ *messbar, so dass*

$$\forall\, x, y \in \mathbb{R}^d : \quad h\left((1-t)x + ty\right) \geq f(x)^{1-t} \cdot g(y)^t$$

Dann

$$\int_{\mathbb{R}^d} h(x)\ dx \geq \left(\int_{\mathbb{R}^d} f(x)\ dx\right)^{1-t} \cdot \left(\int_{\mathbb{R}^d} g(y)\ dy\right)^t$$

Beweis. Mit Induktion nach Dimension d. □

Interessierte Leser können die vollständigen Beweise in dem Scan des handgeschriebenen Manuskripts nachlesen.

5. Logarithmische Sobolev-Ungleichung (LSU)

Ziel: Eine Verallgemeinerungen der isoperimetrischen Ungleichung auf BHC $= \{-1, 1\}^n$ (Lemma 4.11), die insbesondere exponentielle Konzentrationsungleichung für Abweichungen vom Erwartungswert für Funktionen auf BHC impliziert.
Dabei erlaubt es das sogenannte Herbst-Argument aus einer logarithmischen Sobolev-Ungleichung (LSU) Konzentrationsungleichungen zu folgern. Diese Ungleichung wird Ira Herbst zugerechnet, obwohl er sie nie publiziert hat.
Unsere Überlegungen im folgenden Abschnitt sind ein Spezialfall einer allgemeinen *Entropie-Methode*, die auf allg. ZVen, nicht nur auf BHC, anwendbar ist.
Insbesondere kann man mit einem Ansatz, wie wir ihn bereits kennengelernt haben, von Bernoulli-ZVen auf Gaußsche ZVen übergehen und für diese auch Konzentrationsungleichungen folgern.
Die Gaußsche-logarithmische-Sobolev-Ungleichung impliziert auch eine Verallgemeinerung des Johnson-Lindenstrauss-Lemmas. Über die Konzentrationsungleichung ist sie auch auf Probleme der Statistik anwendbar, z.B. in der Regressionsanalyse, genauer, in der Analyse des LASSO-Algorithmus und anderer verwander kleinste Quadrate-Schätzer.
Schliesslich können Hyperkontraktive Ungleichungen aus der Harmonischen Analysis ebenfalls mit Hilfe von logarithmischen Sobolev-Ungleichungen bewiesen werden.

5.1. **LSU für symmetrische Bernoulli-Verteilungen.**

Setting: $f \colon \mathrm{BHC} = \{-1, 1\}^n \to \mathbb{R}, X \colon \Omega \to \mathrm{BHC}$ gleichverteilt.

BHC hat Koordinaten $1, \dots, n$,

$$X = (X_1, \dots, X_n)$$

sodass

$$X_i \colon \Omega \to \{-1, 1\} \quad \text{unabhängige Rademacher-ZVen sind.}$$

Wollen nun Beziehung zwischen zwei Funktionalen herstellen:

1. Entropie:

$$\mathrm{Ent}(f) := E\Big(f(X)\ln(f(X))\Big) - E(f(X)) \cdot \ln(E(f(X))).$$

2. Energieform:

$$\mathcal{E}(f) := \frac{1}{2} E\left(\sum_{i=1}^{n} \Big(f(X) - f(\tilde{X}^{(i)})\Big)^2\right),$$

wobei $\tilde{X}^{(i)} = (X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)$ mit unabhängiger identische Kopie X_i' von X_i.

Setze $Z = f(X)$, sowie $\mathrm{Ent}(Z) = \mathrm{Ent}(f)$, d.h unterdrücke Unterscheidung von Maß P und Bildmaß P_X. (D.h. $X = Id \colon \mathrm{BHC} \to \mathrm{BHC}$) Die Energieform ist uns (inkognito) schon bei der Efron-Stein-Ungleichung begegnet

$$\mathrm{Var}(f(X)) \leq \mathcal{E}(f).$$

Interpretation: links Integral der quadrierten ZV, rechts Integral des quadrierten Gradienten.

Da X gleichverteilt auf BHC gilt

$$\begin{aligned}\mathcal{E}(f) &= \frac{1}{4} E\left(\sum_{i=1}^{n} (f(X) - f(\overline{X}^{(i)}))^2\right) \\ &= \frac{1}{2} E\left(\sum_{i=1}^{n} (f(X) - f(\overline{X}^{(i)}))_+^2\right),\end{aligned}$$

wobei $\overline{X}^{(i)} := (X_1, \dots, X_{i-1}, -X_i, X_{i+1}, \dots, X_n)$ Vektor mit geflippte $i$ter Koordinate ist. Setzen wir $\nabla_i f(x) := \frac{1}{2}(f(x) - f(\overline{x}^{(i)}))$, so kann man

$$\nabla f(x) := (\nabla_1 f(x), \dots, \nabla_n f(x))$$

als diskreten Gradienten interpretieren und E-S-U schreiben als:

$$\mathrm{Var}(f(X)) \leq E(\|\nabla f(X)\|^2) \text{ (eine Art Poincare oder Sobolev-Ungl.)}$$

Satz 5.1. *(LSU für Rademacher-ZVen)*
Sei $f\colon \{-1,1\}^n \to \mathbb{R}$ *beliebig,* $X_1, \dots, X_n \colon \Omega \to \{-1,1\}$ *unabhängige Rademacher-ZVen. Dann gilt:*

$$\operatorname{Ent}(f^2(X)) \leq 2\mathcal{E}(f(X))$$

Der Satz enthält Spezialfälle:

- Iisoperimetrische Ungleichung in Theorem 4.14: Ist $A \subset \mathrm{BHC}$ und $f = \mathbb{1}_A$, so gilt $f = f^2$

$$\begin{aligned} \Rightarrow \quad \operatorname{Ent}(f^2) &= E(\mathbb{1}_A(X)\ln(\mathbb{1}_A(X))) - E(\mathbb{1}_A(X)) \cdot \ln(E(\mathbb{1}_A(X))) \\ &= 0 - P_X(A)\ln(P_X(A)) \end{aligned}$$

  Andererseits gilt für totalen Einfluß/Instabilität von A

$$\begin{aligned} I(A) &= 4E(\|\nabla \mathbb{1}_A(X)\|^2) = 4\mathcal{E}(f) \\ &\geq 2\operatorname{Ent}(f^2) = 2P_X(A)\ln\left(\frac{1}{P_X(A)}\right) \end{aligned}$$

- E-S-U für Rademacher-ZVen (Übung)

Beweis. Wir schreiben $Z = f(X)$ und wenden die Subadditivität der Entropie aus Theorem 4.25 an:

$$\operatorname{Ent}(Z^2) \leq E\left(\sum_{i=1}^n \operatorname{Ent}^{(i)}(Z^2)\right)$$

wobei $\operatorname{Ent}^{(i)}(Z^2) = E^{(i)}(\phi(Z^2)) = E^{(i)}\left(Z^2\ln(Z^2)\right) - E^{(i)}(Z^2)\cdot\ln(E^{(i)}(Z^2))$ wie gehabt bzgl. des univariaten Erwartungswerts $\int \dots dP_{X_i}$ gebildet wird. Es reicht zu zeigen:

$$\forall i \in \{1,\dots,n\} : \operatorname{Ent}^{(i)}(Z^2) \leq \frac{1}{2}\frac{2}{2}E^{(i)}\left((f(X) - f(\overline{X}^{(i)}))^2\right). \tag{5.1}$$

Denn dann folgt:

$$\begin{aligned} \operatorname{Ent}(Z^2) &\leq 2\sum_{i=1}^n E\left(E^{(i)}\left(\left(\frac{f(X) - f(\overline{X}^{(i)})}{2}\right)^2\right)\right) \\ &= 2\sum_{i=1}^n E((\nabla_i f(X))^2) \qquad \text{wg. Turmeigenschaft} \\ &= 2E\left(\sum_{i=1}^n (\nabla_i f(X))^2\right) = 2E\left(\|\nabla f(X)\|^2\right) = 2\mathcal{E}(f). \end{aligned}$$

Um (5.1) zu zeigen, beachten wir, dass bei fixierten $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, $Z = f(X) \in \{a, b\}$ nur zwei Werte annimmt und (5.1) äquivalent ist zu

$$\begin{aligned} &\frac{a^2}{2}\ln(a^2) + \frac{b^2}{2}\ln(b^2) - \frac{a^2+b^2}{2}\ln\left(\frac{a^2+b^2}{2}\right) \\ &\leq \frac{1}{2}\left(\frac{1}{2}(a-b)^2 + \frac{1}{2}(b-a)^2\right) = \frac{1}{2}(a-b)^2 \end{aligned} \tag{5.2}$$

Um letzteres zu zeigen, wenden wir folgende Reduktionen an:

- Während für linke Seite die Vorzeichen von a, b keine Rolle spielen, gilt rechts dagegen $(|a| - |b|)^2 \leq (a-b)^2$ also ist das Vorzeichen relevant.

Also impliziert (5.2) für $a, b \geq 0$ auch allgemeines (5.1).

- a, b gehen in (5.2) symmetrisch ein, also

$$(5.2) \text{ für } a \leq b \quad \Leftrightarrow \quad (5.2) \text{ für } a \geq b$$

Ab jetzt $a \geq b \geq 0$ o.B.d.A. Für festes $b \geq 0$ definiert die Funktion

$$h\colon [b, \infty) \to \mathbb{R}, \quad h(a) = \frac{a^2}{2}\ln(a^2) + \frac{b^2}{2}\ln(b^2) - \frac{a^2+b^2}{2}\ln\left(\frac{a^2+b^2}{2}\right) - \frac{1}{2}(a-b)^2.$$

Um (5.2) nachzuweisen reicht es demnach nachzuweisen, dass $h \leq 0$ überall. Betrachte Ableitungen von h, insbesondere in $a = b$:

$$\begin{aligned} h(b) &= b^2\ln(b^2) - b^2\ln(b^2) - 0 = 0 \\ h'(a) &= a\ln\left(\frac{2a^2}{a^2+b^2}\right) - (a-b) \quad \Rightarrow \quad h'(b) = 0 \end{aligned}$$

$\ln(x) - x \leq 1$ impliziert: $\quad h''(a) = 1 + \ln\left(\frac{2a^2}{a^2+b^2}\right) - \frac{2a^2}{a^2+b^2} \leq 0$

Taylorentwicklung: $h(a) \leq 0$ für alle $a > b$ □

Nun erster Schritt zu einem viel allgemeinerterem Fall: Erlaube asymmetrische Bernoulli-ZVen $X_1, \ldots, X_n \in \{-1, 1\}$ u.i.v. mit $p := P(X_i = 1) \in [0, 1]$. In diesem Fall gilt:

$$\mathcal{E}(f) := \frac{1}{2}E\left(\sum_{i=1}^n (f(X) - f(\tilde{X}^{(i)}))^2\right) = p(1-p)E\left(\sum_{i=1}^n (f(X) - f(\overline{X}^{(i)}))^2\right)$$

Satz 5.2. *Seien $X_1, \ldots, X_n \sim$ Ber(p) u.i.v. und $f\colon \{-1, 1\}^n \to \mathbb{R}$ beliebig. Dann gilt:*

$$\mathrm{Ent}(f^2) \leq c(p)\mathcal{E}(f)$$

mit $c(p) = \frac{1}{1-2p}\ln\left(\frac{1-p}{p}\right)$.

Beweis. Übung. □

Theorem 5.1 folgt als Spezialfall, denn Regel von L'Hospital liefert

$$\lim_{p\to\frac{1}{2}} c(p) = \lim_{p\to\frac{1}{2}} \frac{\ln\frac{1-p}{p}}{1-2p} = \lim_{p\to\frac{1}{2}} \frac{\frac{-1}{(1-p)p}}{-2} = \frac{\frac{1}{1/4}}{2} = 2$$

Die Funktion $c\colon (0,1)\to[2,\infty)$ ist konvex und divergiert an den Rändern.

5.2. **Herbst-Argument.**
Ziel: exponentielle Konzentrationsungleichungen
Zunächst: für unabhängige Rademacher-ZVen

$$X = (X_1,\dots,X_n) \in \{-1,1\}^n = \text{BHC gleichverteilt.}$$
$$f\colon \text{BHC}\to\mathbb{R}, Z = f(X)$$

Idee wieder: Verwende Substitution mit exponentieller Funktion.
Um exponentielle Konzentrationsungleichung zu erhalten, betrachte

$$g\colon \text{BHC}\to\mathbb{R}, g(x) = \exp\left(\frac{\lambda}{2}f(x)\right) \qquad \text{für ein zu optimierendes } \lambda\in\mathbb{R}$$

Vorbereitung:

$$\begin{aligned}\operatorname{Ent}(g^2) &= \operatorname{Ent}(\exp(\lambda f(X))) = \operatorname{Ent}(e^{\lambda Z}) = \lambda E(Ze^{\lambda Z}) - E(e^{\lambda Z})\ln(E(e^{\lambda Z}))\\ &= \lambda M'(\lambda) - M(\lambda)\ln(M(\lambda)),\end{aligned}$$

da $M(\lambda) = M_Z(\lambda) = E(e^{\lambda Z})$ und $M'(\lambda) = E(Ze^{\lambda Z})$. Um eine Schranke an die Entropie von g^2 zu erhalten, stelle eine Differentialungleichung für M auf:

$$\begin{aligned}\operatorname{Ent}(g^2) &\overset{Thm.5.1}{\leq} 2\mathcal{E}(g) = \frac{1}{2}\sum_{i=1}^n E\left(\left(\exp\left(\frac{\lambda f(X)}{2}\right) - \exp\left(\frac{\lambda f(\overline{X}^{(i)})}{2}\right)\right)^2\right)\\ &= \sum_{i=1}^n E\left(\left(\exp\left(\frac{\lambda f(X)}{2}\right) - \exp\left(\frac{\lambda f(\overline{X}^{(i)})}{2}\right)\right)_+^2\right)\\ &= \sum_{i=1}^n E\left(\left(\exp\left(\frac{\lambda f(X)}{2}\right) - \exp\left(\frac{\lambda f(\overline{X}^{(i)})}{2}\right)\right)^2 \mathbb{1}_{\{\lambda f(X)\geq\lambda f(\overline{X}^{(i)})\}}\right)\end{aligned}$$

Für alle $z > y\in\mathbb{R}$ gilt:

$$\frac{\exp\left(\frac{z}{2}\right) - \exp\left(\frac{y}{2}\right)}{\frac{z-y}{2}} = \exp'(\xi) \leq \exp'(z/2) = \exp\left(\frac{z}{2}\right)$$

da $t\mapsto\exp(t)$ konvex bzw. Ableitung isoton.

$$\Rightarrow \quad \mathrm{Ent}(g^2) \leq \frac{\lambda^2}{4} \sum_{i=1}^{n} E\left[(f(X) - f(\overline{X}^{(i)}))_+^2 e^{\lambda Z}\right]$$
$$= \frac{\lambda^2}{4} E(e^{\lambda Z} \sum_{i=1}^{n} (f(X) - f(\overline{X}^{(i)}))_+^2) \leq \nu \frac{\lambda^2}{4} E(e^{\lambda Z})$$

falls wir $\nu := \max_{x \in \mathrm{BHC}} \sum_{i=1}^{n} (f(x) - f(\overline{x}^{(i)}))^2$ setzen. Offensichtlich $\nu < \infty$, da BHC endlich. Einsetzten ergibt Differentialungleichung für MEF von Z:

$$\lambda M'(\lambda) - M(\lambda) \ln(M(\lambda)) = \mathrm{Ent}(g^2) \leq \frac{\nu \lambda^2}{4} M(\lambda)$$

Bis hierhin galten die Aussagen für alle $\lambda \in \mathbb{R}$. Im folgenden interssieren wir uns nur für den Fall $\lambda \neq 0$. Dann gilt $\lambda^2 M(\lambda) \geq 0$ und Division ergibt

$$\Rightarrow \quad \frac{M'(\lambda)}{\lambda M(\lambda)} - \frac{\ln(M(\lambda))}{\lambda^2} \leq \frac{\nu}{4}$$

und setze $\psi(\lambda) := \ln(M(\lambda))$

$$\Rightarrow \quad \psi'(\lambda) = \frac{M'(\lambda)}{M(\lambda)}$$
$$\Rightarrow \quad \frac{d}{d\lambda}\left(\frac{\psi(\lambda)}{\lambda}\right) = \frac{1}{\lambda}\frac{M'(\lambda)}{M(\lambda)} - \frac{1}{\lambda^2}\ln(M(\lambda)) \leq \frac{\nu}{4}.$$

Wir wollen $\frac{\psi(\lambda)}{\lambda}$ durch Integration abschätzen. Wert bei 0: $\lim_{\lambda \to 0} \psi(\lambda) = \lim_{\lambda \to 0} \ln(E(e^{\lambda Z})) = 0$. Nach der Regel von de L'Hospital folgt

$$\lim_{\lambda \to 0} \frac{\psi(\lambda)}{\lambda} = \lim_{\lambda \to 0} \frac{\psi'(\lambda)}{1} = \lim_{\lambda \to 0} \frac{M'(\lambda)}{M(\lambda)} = \frac{M'(0)}{M(0)} = E(Z).$$

Bis hier galten alle Aussagen für $\lambda \neq 0$. Falls $\lambda > 0$, so ergibt Integration von 0 bis λ:

$$\frac{\psi(\lambda)}{\lambda} = E(Z) + \int_0^\lambda \left(\frac{\psi(t)}{t}\right)' dt$$
$$\leq E(Z) + \int_0^\lambda \frac{\nu}{4}\, dt = E(Z) + \frac{\lambda\nu}{4},$$

also $\psi(\lambda) \leq \lambda E(Z) + \frac{\lambda^2\nu}{4} \quad \Rightarrow \quad M(\lambda) \leq e^{\lambda E(Z)+\frac{\lambda^2\nu}{4}}$. Mit der Markov-Ungleichung folgt:

$$\begin{aligned} P(Z \geq E(Z)+t) &\overset{\lambda>0}{=} P(\lambda Z \geq \lambda E(Z) + \lambda t) = P(e^{\lambda Z} \geq e^{\lambda E(Z)+\lambda t}) \\ &\leq \inf_{\lambda>0} \left(M(\lambda) e^{-\lambda E(Z)-t\lambda} \right) \\ &\leq \inf_{\lambda>0} \exp\left(\lambda E(Z) + \frac{\lambda^2\nu}{4} - \lambda E(Z) - t\lambda \right) \\ &\leq \inf_{\lambda>0} \left(e^{\frac{\lambda^2\nu}{4} - \lambda t} \right) \leq e^{-\frac{t^2}{\nu}} \quad \text{falls } \lambda = \frac{2t}{\nu}. \end{aligned}$$

Analoge Rechnung für $\lambda < 0$.

$$\begin{aligned} &\frac{\psi(\lambda)}{\lambda} = E(Z) - \int_\lambda^0 \left(\frac{\psi(t)}{t} \right)' dt \geq E(Z) - \frac{\lambda\nu}{4} \\ \Rightarrow &\psi(\lambda) \leq \lambda E(Z) - \frac{\lambda^2\nu}{4} \\ \Rightarrow &M(\lambda) \leq e^{\lambda E(Z) - \frac{\lambda^2\nu}{4}}. \end{aligned}$$

Die Markov-Ungleichung liefert:

$$\begin{aligned} P(Z < E(Z)-t) &\overset{\lambda<0}{=} P(\lambda Z > \lambda E(Z) - \lambda t) = P(e^{\lambda Z} > e^{\lambda E(Z)-\lambda t}) \\ &\leq \inf_{\lambda>0} \frac{M(\lambda)}{e^{\lambda E(Z)-\lambda t}} \\ &\leq \inf_{\lambda>0} \exp\left(\lambda E(Z) + \frac{\lambda^2\nu}{4} - \lambda E(Z) + t\lambda \right) \\ &\leq e^{-\frac{t^2}{\nu}} \quad \text{falls } \lambda = -\frac{2t}{\nu}. \end{aligned}$$

Damit ist bewiesen:

Satz 5.3. *Sei $f\colon BHC \to \mathbb{R}$ und X gleichverteilt auf BHC. Sei $\nu > 0$ so dass*

$$\forall x \in BHC\colon \quad \sum_{j=1}^n \left(f(x) - f(\bar{x}^{(j)}) \right)_+^2 \leq \nu$$

Dann gilt für $Z := f(X)$ und jedes $t > 0$:

$$\max\left\{ P(Z > EZ + t), P(Z < EZ - t) \right\} \leq \exp(-t^2/\nu)$$

Vergleiche mit E-S-U:

$$\mathrm{Var}(Z) \leq \nu_{ES} := E\left[\sum_{j=1}^n \left((Z - Z_j')_+ \right)^2 \right] \leq \frac{\nu}{2}$$

da:

- $E\{\ldots\} \leq \sup\{\ldots\}$
  (dafür stellen wir hier stärkere, weil pkt-weise Annahme)
- geflippte ZV haben doppelten Effekt im Vergleich zu resampelten

Hier erhalten wir eine Schranke für W'keit seltener Ereignisse mit Abfallrate wie bei normalverteilten ZV. Dagegen konnte man aus E-S-U nur die Rate

$$P(Z > EZ + t) \leq \exp\left(-t/\sqrt{\nu}\right)$$

herleiten.
Wir merken uns die Schritte der Beweisstrategie, weil sie auch in anderen Fällen (modifiziert) implemetiert werden kann:

- logarithmische Sobolev-Ungleichung
- anwenden auf ZV $\exp(\lambda Z)$
- Differentialungleichung herleiten
- aufintegrieren
- Cramer-Chernov-Schranke

5.3. **Gaußsche LSU.** Nachdem wir den Fall BHC bzw. Vektoren von Bernoulli-ZV studiert haben können wir einen Übergang durchführen

$$\text{Summen von Bernoulli ZV} \longrightarrow \text{ Gauß ZV}$$

Satz 5.4. *(Gaußsche log-Sobolev-Ungleichung)*
Sei $X = (X_1, \ldots, X_d) \sim \mathcal{N}(0, Id_d)$, $f\colon \mathbb{R}^d \to \mathbb{R}$, $f \in C^1(\mathbb{R}^d)$ *und* $Z := f(X)$.
Dann gilt

$$\mathrm{Ent}(Z^2) = \mathrm{Ent}(f^2) \leq E\left[\|\nabla f(X)\|^2\right]$$

Aussage gilt mit Dichtheitsargument auch für Funktionen $f \in H^1 = W^{1,2}$ bzgl. des natürlichen Maßes, d.h. $\mu = \mathcal{N}(0, Id_d)$.
Theorem 5.4 verstärkt und impliziert die Gaußsche Poincare Ungleichung Theorem 3.39.
Die Beweisideen ähneln sich.

5.4. **TSI-Konzentrationsungleichungen für Gauß-Zufallsvariablen.**

5.5. **Konzentrationsungleichung für Suprema von Gauß-Prozessen.**

5.6. **Zufällige Gaußsche Projektionen.**
In Kapitel 2.9 wird ein Johnson-Lindenstrauß-Lemma in einem einfacheren Rahmen dargestellt.

5.7. **Hyperkontraktivität.**

Literatur