HANDOUT #2
MASSKONZENTRATION UND GROSSE ABWEICHUNGEN

FOLKMAR BORNEMANN

1. Motivation

Es seien $X_1, \dots, X_n$ iid reelle Zufallsvariablen mit Mittelwert $\mathbb{E}X_1 = 0$ und Varianz $\mathrm{var}(X_1) = \sigma^2$. Nach dem zentralen Grenzwertsatz gilt für $n \to \infty$

$$\mathbb{P}(X_1 + \cdots + X_n \geqslant t\sigma\sqrt{n}) \to \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \leqslant \frac{1}{2} e^{-t^2/2} \qquad (t \geqslant 0). \tag{1}$$

Abweichungen der Summe $X_1 + \cdots + X_n$ vom Mittelwert 0, welche im Verhältnis zur *typischen* Abweichung $O(\sigma\sqrt{n})$ (d.h. für $t \gg 1$) *groß* sind, sind also äußerst unwahrscheinlich (zumindest asymptotisch für $n \to \infty$). Allgemeiner betrachtet man Wahrscheinlichkeiten großer Abweichungen der Form

$$\mathbb{P}(F(X_1, \dots, X_n) - \bar{F} \geqslant t\sigma),$$

wobei $\bar{F}$ ein Mittel des (nichtlinearen) Funktionals $F(X)$ bezeichnet und σ die typische Abweichung von diesem Wert. In der *Theorie großer Abweichungen* studiert man optimale obere Abschätzungen und präzise Asymptotiken für einen geeigneten zweiskaligen Limes $n \to \infty$ und $t \to \infty$. Solch präzise Ergebnisse hängen aber von den Details der Verteilung von X und von n ab; sie sind oft nur für sehr einfache Funktionale erhältlich. Alternativ studiert man für größere Klassen nichtlinearer Funktionale nicht-optimale obere Abschätzungen, die unabhängig von solchen Details sind. Wenn diese Abschätzungen subexponentielle Verteilungsenden (also extrem straffe Wahrscheinlichkeitsverteilungen) liefern, spricht man vom Phänomen der „Maßkonzentration“.

2. Das Prinzip grosser Abweichungen

Das typische Beispiel ist der klassische Satz von Cramér (1938). Diesen und viel mehr findet man in der Monographie von Dembo and Zeitouni (2010). Es seien $X_1, \dots, X_n$ iid reelle Zufallsvariablen mit $\bar{x} = \mathbb{E}X_1$. Wir nehmen an, dass die Kumulanten-erzeugende Funktion (engl.: CGF = cumulant generating function)

$$\Lambda(\theta) = \log \mathbb{E}\, e^{\theta X_1}$$

für alle $\theta \in \mathbb{R}$ einen endlichen Wert liefert ($\Lambda : \mathbb{R} \to \mathbb{R}$ ist dann konvex). Der Satz von Cramér besagt, dass die *Intensität* großer Abweichungen,[1] d.i.

$$\lim_{n\to\infty} n^{-1} \log \mathbb{P}(X_1 + \cdots + X_n \geqslant xn) = -\Lambda^*(x), \tag{2}$$

für $x > \bar{x}$ durch die (negative) Legendre–Fenchel Transformierte von Λ gegeben ist; diese ist allgemein durch

$$\Lambda^*(x) = \sup_{\theta\in\mathbb{R}} (x\theta - \Lambda(\theta))$$

Date: 6. Oktober 2010.

[1] In der Notation von §1 wird also $t = x\sqrt{n} \to \infty$ betrachtet.

definiert; tatsächlich gilt unter den Voraussetzungen an die CGF sogar, dass

$$\Lambda^*(x) = \sup_{\theta>0}(x\theta - \Lambda(\theta)) \qquad (x \geqslant \bar{x}). \tag{3}$$

Da $\log \mathbb{P}(X_1 + \cdots + X_n \geqslant xn)$ in n superadditiv ist, lässt sich der monotone Limes in (2) durch das Supremum ersetzen und wir erhalten die obere Abschätzung

$$\mathbb{P}(X_1 + \cdots + X_n \geqslant xn) \leqslant e^{-n\Lambda^*(x)} \qquad (x > \bar{x}). \tag{4}$$

Insbesondere zeigt sich, dass hier der Exponent $\Lambda^*(x)$ für festes x asymptotisch bestmöglich ist.

Bemerkung 1. Die Abschätztung (4) lässt sich (ohne die Optimalität des Exponenten) ganz einfach direkt zeigen: Für $\theta > 0$ erhalten wir nämlich mit der Tschebyscheff'schen Ungleichung, und da $X_1, \ldots, X_n$ iid sind,

$$\begin{aligned}\mathbb{P}(X_1 + \cdots + X_n \geqslant xn) &= \mathbb{P}\left(e^{\theta(X_1+\cdots+X_n)} \geqslant e^{nx\theta}\right) \\ &\leqslant e^{-nx\theta}\mathbb{E}\left(e^{\theta(X_1+\cdots+X_n)}\right) = e^{-nx\theta}\left(\mathbb{E}\, e^{\theta X}\right)^n = e^{-n(x\theta-\Lambda(\theta))};\end{aligned}$$

Wegen (3) liefert eine Optimierung über $\theta > 0$ schließlich (4).

Bemerkung 2. Für $\mathbb{E}X_1 = 0$ und $\operatorname{var}(X_1) = \sigma^2$ können wir (4) in der Form

$$\mathbb{P}(X_1 + \cdots + X_n \geqslant t\sigma\sqrt{n}) \leqslant \exp(-n\Lambda^*(t\sigma/\sqrt{n})) \qquad (t > 0)$$

schreiben. Aus $\Lambda(\theta) = \theta^2\sigma^2/2 + o(\theta^2)$ für $\theta \to 0$ und etwas konvexer Analysis folgt $\Lambda^*(x) = x^2/2\sigma^2 + o(x^2)$ für $x \to 0$; also gilt mit genau dem gleichen Exponenten wie in (1), aber insgesamt um einen Faktor 2 weniger scharf:

$$\limsup_{n\to\infty} \mathbb{P}(X_1 + \cdots + X_n \geqslant t\sigma\sqrt{n}) \leqslant e^{-t^2/2} \qquad (t \geqslant 0).$$

Für *subgaußsche* Verteilungen, definiert durch $\Lambda(\theta) \leqslant \theta^2\sigma^2/2$ ($\theta > 0$) oder äquivalent $\Lambda^*(x) \geqslant x^2/2\sigma^2$ ($x > 0$), gilt die Abschätzung nicht nur im Limes:

$$\mathbb{P}(X_1 + \cdots + X_n \geqslant t\sigma\sqrt{n}) \leqslant e^{-t^2/2} \qquad (t \geqslant 0). \tag{5}$$

Beispiele sind die Normalverteilung mit $\Lambda(\theta) = \theta^2\sigma^2/2$ und die Bernoulli-Verteilung auf $\{-1, 1\}$ mit $p = 1/2$, für die $\sigma = 1$ und (nach dem Einschließungssatz für alternierende Reihen)

$$\Lambda(\theta) = \log\cosh(\theta) = \sum_{k=1}^{\infty} \frac{2^{2k-1}(2^{2k}-1)B_{2k}}{k(2k)!}\theta^{2k} \leqslant \frac{\theta^2}{2}.$$

3. Das Pänomen der Masskonzentration

Der Klassiker unter jenen Abschätzungen, die Abweichungen vom Mittelwert beschreiben, ist zweifellos die Ungleichung von Bienaymé–Tschebyscheff (1853/1866): Für eine reelle Zufallsvariable X mit Varianz $\sigma^2 = \operatorname{var}(X)$ gilt

$$\mathbb{P}(|X - \mathbb{E}X| \geqslant t\sigma) \leqslant t^{-2} \qquad (t > 0).$$

In direkter Anwendung ist die Ungleichung für viele Anwendungen jedoch zu schwach; so hat sie erst nach Exponentiation den exponentiellen Abfall in (5) geliefert. Ein solcher Exponentiationstrick führt auch auf folgende klassische Ungleichung:

Satz 3 (Hoeffding 1963). *Es seien $X_1, \dots, X_n$ unabhängige reelle Zufallsvariable, welche fast sicher den Ungleichungen $a_j \leqslant X_j \leqslant b_j$ $(j = 1, \dots, n)$ genügen. Dann gilt für $S_n = X_1 + \cdots + X_n$, dass*

$$\mathbb{P}(S_n - \mathbb{E}S_n \geqslant t\sigma) \leqslant \exp(-2t^2) \qquad (t > 0);$$

dabei ist $\sigma^2 = \sum_{j=1}^n (b_j - a_j)^2$.

Da sich Satz genauso auf $-X_1, \dots, -X_n$ anwenden lässt, gilt zudem

$$\begin{aligned}\mathbb{P}(|S_n - \mathbb{E}S_n| \geqslant t\sigma) \\ \leqslant \mathbb{P}(S_n - \mathbb{E}S_n \leqslant -t\sigma) + \mathbb{P}(S_n - \mathbb{E}S_n \geqslant t\sigma) \leqslant 2\exp(-2t^2) \qquad (t > 0).\end{aligned}$$

Dieses Argument trifft in gleicher Art auf alle folgenden Sätze zu.[2]

Al Beispiel betrachten wir Bernoulli-verteilte Zufallsvariablen auf $\{-1, 1\}$ mit $p = 1/2$: Hier reproduziert die Hoeffding'sche Ungleichung extakt die aus dem Satz von Cramér gewonnene Abschätzung (5).

In den 1970er Jahren (unter Aufgriff einer frühen Vorarbeit von Lévy über das normierte Lebesgue'sche Maß auf euklidischen Sphären aus dem Jahre 1919) hat man erkannt, dass sich derartige Konzentrationsungleichungen auf nichtlineare Funktionale gewisser Zufallsvariablen $X_1, \dots, X_n$ verallgemeinern lassen; Pionier auf diesem Gebiet war der Funktionalanalytiker Vitali Milman, der solche Ungleichungen auf die Geometrie von Banachräumen anwandte und den Begriff der „Maßkonzentration" prägte. Eine umfassende Studie dieses Phänomens findet sich in den Monographien von Ledoux (2001) und (für diskrete Maße) von Dubhashi and Panconesi (2009). Wir führen hier nur ein paar typische Ungleichungen auf, die wir in der Vorlesung benutzen werden.

Die unmittelbare Verallgemeinerung der Hoeffding'schen Ungleichung auf nichtlineare F lautet wie folgt (der Beweis verwendet bedingte Erwartungswerte, ist ansonsten genauso elementar wie derjenige der Hoeffding'schen Ungleichung):

Satz 4 (McDiarmid 1989). *Es seien $X_1, \dots, X_n$ unabhängige Zufallselemente $X_j \in A_j$ $(j = 1, \dots, n)$. Es sei $F : A_1 \times \cdots \times A_n \to \mathbb{R}$ eine messbare Abbildung, deren Oszillation im j-ten Argument durch $\sigma_j \geqslant 0$ beschränkt ist $(j = 1, \dots, n)$:*

$$|F(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_n) - F(x_1, \dots, x_{j-1}, x'_j, x_{j+1}, \dots, x_n)| \leqslant \sigma_j$$

für $x'_j \in A_j$, $x_k \in A_k$ $(k = 1, \dots, n)$. Dann gilt

$$\mathbb{P}(F(X) - \mathbb{E}F(X) \geqslant t\sigma) \leqslant \exp(-2t^2) \qquad (t > 0);$$

dabei ist $\sigma^2 = \sum_{j=1}^n \sigma_j^2$.

Die mächtigsten Konzentrationsabschätzungen verwenden jedoch Lipschitz-Bedingungen an F; wir geben zwei fundamentale Beispiele. Das erste basiert auf der Theorie *logarithmischer Sobolev-Ungleichungen* (Royer 2007): Ein Maß μ auf $\mathbb{R}^n$ erfüllt eine solche Ungleichung, falls für alle differenzierbaren $f : \mathbb{R}^n \to \mathbb{R}$ mit der Normierung $\int f^2 \, d\mu = 1$ gilt (links steht die negative *Entropie* der W-Dichte f^2)

$$\int_{\mathbb{R}^n} f^2 \log f^2 \, d\mu \leqslant 2\sigma^2 \int_{\mathbb{R}^n} \|\nabla f\|_2^2 \, d\mu. \tag{6}$$

[2] Nur beim Satz von Talagrand gestattet die Konvexitätsvoraussetzung keine derartige Symmetrie.

Logarithmische Sobolev-Ungleichungen sind invariant unter Produktbildung von Maßen, die Konstante ist das *Maximum* derjenigen der einzelnen Faktoren. Dies verhindert für die Verteilung unabhängiger Zufallsvariable $X_1, \dots, X_n$ eine Abhängigkeit von der Dimension n.

Satz 5 (Herbst 1976, Davies/Simon 1984, Bobkov 1995, Ledoux 1996)**.** *Es seien $X_1, \dots, X_n$ Zufallsvariablen, deren gemeinsame Wahrscheinlichkeitsverteilung auf $\mathbb{R}^n$ ein Maß μ besitzt, welches der logarithmischen Sobolev-Ungleichung (6) mit der Konstanten σ genügt. Dann gilt für jede 1-Lipschitz-stetige Funktion $F : \mathbb{R}^n \to \mathbb{R}$,[3] dass*

$$\mathbb{P}(F(X) - \mathbb{E}F(X) \geqslant t\sigma) \leqslant \exp(-t^2/2) \qquad (t > 0). \tag{7}$$

Nach Gross (1975) erfüllt die gemeinsame Wahrscheinlichkeitsverteilung (d.i. das Produktmaß) von iid normalverteilten Zufallsvariablen $X_1, \dots, X_n$ die logarithmische Sobolev-Ungleichung (6) mit $\sigma^2 = \mathrm{var}(X_1)$. Theorem 5 liefert hier also eine Konzentrationsabschätzung, deren Konstanten von n unabhängig sind.[4] Für die 1-Lipschitz-stetige Funktion $F(X_1, \dots, X_n) = (X_1 + \cdots + X_n)/\sqrt{n}$ erhalten wir erneut die aus dem Satz von Cramér gewonnene Abschätzung (5).

Das zweite Beispiel besitzt den großen Vorteil sehr leicht überprüfbarer Voraussetzungen. Alle bekannten Beweise sind tiefliegend; Talagrands ursprünglicher Beweis verwendet Dimensionsinduktion und elementare, aber „mysteriöse“ad-hoc Abschätzungen; ein systematischerer Zugang wurde von Katalin Marton entwickelt und verwendet die Theorie optimalen Transports von Maßen (Villani 2009).

Satz 6 (Talagrand 1995)**.** *Es seien $X_1, \dots, X_n$ unabhängige reelle Zufallsvariable, für welche fast sicher $a \leqslant X_j \leqslant b$ $(j = 1, \dots, n)$ gilt. Es sei $F : [a, b] \times \cdots \times [a, b] \to \mathbb{R}$ eine 1-Lipschitz-stetige, konvexe Abbildung. Dann gilt mit $\sigma = b - a$*

$$\mathbb{P}(|F(X) - \mathbb{M}F(X)| \geqslant t\sigma) \leqslant 4\exp(-t^2/4) \qquad (t > 0);$$

dabei ist $\mathbb{M}F(X)$ ein Median von $F(X)$. Das gleiche Resultat gilt, wenn der Median $\mathbb{M}F(X)$ durch den Erwartungswert $\mathbb{E}F(X)$ ersetzt wird (wobei statt der Konstante 4 möglicherweiser eine andere absolute Konstante $c > 0$ genommen werden muss). Für das obere[5] Verteilungsende gilt konkret (hier reicht die koordinatenweise Konvexität von F)

$$\mathbb{P}(F(X) - \mathbb{E}F(X) \geqslant t\sigma) \leqslant \exp(-t^2/2) \qquad (t > 0).$$

Literatur

Dembo, A. and Zeitouni, O.: 2010, *Large Deviations Techniques and Applications*, Springer-Verlag, Berlin.

Dubhashi, D. P. and Panconesi, A.: 2009, *Concentration of Measure for the Analysis of Randomized Algorithms*, Cambridge University Press.

Ledoux, M.: 2001, *The Concentration of Measure Phenomenon*, Amer. Math. Soc., Providence.

Pisier, G.: 1989, *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press.

Royer, G.: 2007, *An Initiation to Logarithmic Sobolev Inequalities*, Amer. Math. Soc., Providence.

Villani, C.: 2009, *Optimal Transport*, Springer-Verlag, Berlin.

[3] Eine solche Funktion erfüllt $|F(x) - F(x')| \leqslant \|x - x'\|_2$ für alle $x, x' \in \mathbb{R}^n$.

[4] Für iid normalverteilte Zufallsvariable wird die Abschätzung (7) nach Sudakov/Tsirel´son (1974) und Borell (1975) benannt. Ein sehr eleganter direkter Beweis (mit einer etwas schlechteren Konstante im Exponenten) geht auf Maurey und Pisier zurück (siehe Pisier 1989, Thm. 4.7).

[5] Das Ergebnis lässt sich nicht sofort (und mit der gleichen Konstanten) auf das untere Verteilungsende übertragen, da $-F$ nur dann auch kordinatenweise konvex ist, wenn F linear ist. In diesem Fall wären wir aber im Bereich der Hoeffding'schen Ungleichung (mit einem schlechteren Exponenten).