Printed: 27 November 2015

Chapter 2

A few good inequalities

Section 2.1 *introduces exponential tail bounds.*

Section 2.2 *introduces the method for bounding tail probabilities using moment generating functions.*

Section 2.3 *describes analogs of the usual $\mathcal{L}^p$ norms, which have proved useful in empirical process theory.*

Section 2.4 *discusses subgaussianity, a most useful extension of the gaussianity property.*

Section 2.5 *discusses the Bennett inequality for sums of independent random variables that are bounded above by a constant.*

Section 2.6 *shows how the Bennett inequality implies one form of the Bernstein inequality, then discusses extensions to unbounded summands.*

Section 2.7 *describes two ways to capture the idea of tails that decrease like those of the exponential distribution.*

Section 2.8 *illustrates the ideas from the previous Section by deriving a subgaussian/subexponential tail bound for quadratic forms in independent subgaussian random variables.*

2.1 Tail bounds and concentration

Basic::S:intro

Sums of independent random variables are often approximately normal distributed. In particular, their tail probabilities often decrease in ways similar to the tails of a normal distribution. Several classical inequalities make this vague idea more precise. Such inequalities provided the foundation on which empirical process theory was built.

We know a lot about the normal distribution. For example, if W has a $N(\mu, \sigma^2)$ distribution then (Feller, 1968, Section 7.1)

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \le \mathbb{P}\{W \ge \mu + \sigma x\} \le \frac{1}{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \qquad \text{for all } x > 0.$$

Clearly the inequalities are useful only for larger x: as x decreases to zero the lower bound goes to $-\infty$ and the upper bound goes to $+\infty$. For many purposes the $\exp(-x^2/2)$ factor matters the most. Indeed, the simpler tail bound $\mathbb{P}\{W \ge \mu + \sigma x\} \le \exp(-x^2/2)$ for $x \ge 0$ often suffices for asymptotic arguments. Sometimes we need a better inequality showing that the distribution concentrates most of its probability mass in a small region, such as

$$\mathbb{P}\{|W - \mu| \ge \sigma x\} \le 2e^{-x^2/2} \qquad \text{for all } x \ge 0,$$

a ***concentration*** inequality.

Similar-looking tail bounds exist for random variables that are not exactly normally distributed. Particularly useful are the so-called exponential tail bounds, which take the form

\E@ exp.bnd <1>

$$\mathbb{P}\{X \ge x + \mathbb{P}X\} \le e^{-g(x)} \qquad \text{for } x \ge 0,$$

with g a function that typically increases like a power of x. Sometimes g behaves like a quadratic for some range of x and like a linear function further out in the tails; sometimes some logarithmic terms creep in; and sometimes it depends on $\text{var}(X)$ or on some more exotic measure of size, such as an Orlicz norm (see Section 2.3).

For example, suppose $S = X_1 + \dots + X_n$, a sum of independent random variables with $\mathbb{P}X_i = 0$ and each X_i bounded above by some constant b. The ***Bernstein inequality*** (Section 2.6) tells us that

$$\mathbb{P}\{S \ge x\sqrt{V}\} \le \exp\left(\frac{-x^2/2}{1 + \frac{1}{3}bx/\sqrt{V}}\right) \qquad \text{for } x \ge 0,$$

where V is any upper bound for $\text{var}(S)$. If each X_i is also bounded below by $-b$ then a similar tail bound exists for $-S$, which leads to a concentration inequality, an upper bound for $\mathbb{P}\{|S| \ge x\sqrt{V}\}$.

For x much smaller than $\sqrt{V}/b$ the exponent in the Bernstein inequality behaves like $-x^2/2$; for much larger x the exponent behaves like a negative multiple of x. In a sense that is made more precise in Sections 2.6 and 2.7, the inequality gives a subgaussian bound for x not too big and a subexponential bound for very large x.

2.2 Moment generating functions

Basic::S:mgf

The most common way to get a bound like <1> uses the moment generating function, $M(\theta) := \mathbb{P}e^{\theta X}$. From the fact that exp is nonnegative and $\exp(\theta(X-w)) \ge 1$ when $X \ge w$ and $\theta \ge 0$ we have

\E@ mgf.tail <2>

$$\mathbb{P}\{X \ge w\} \le \inf_{\theta \ge 0} \mathbb{P}e^{\theta(X-w)} = \inf_{\theta \ge 0} e^{-\theta w} M(\theta).$$

Basic::normal <3> **Example.** If X has a $N(\mu, \sigma^2)$ distribution then $M(\theta) = \exp(\theta\mu + \sigma^2\theta^2/2)$ is finite for all real θ. For $x \ge 0$ inequality <2> gives

$$\begin{aligned}\mathbb{P}\{X \ge \mu + \sigma x\} &\le \inf_{\theta \ge 0} \exp(-\theta(\mu + \sigma x) + \theta\mu + \theta^2\sigma^2/2) \\ &= \exp(-x^2/2) \qquad \text{for all } x \ge 0,\end{aligned}$$

the minimum being achieved by $\theta = x/\sigma$.

Analogous arguments, with $X - \mu$ replaced by $\mu - X$, give an analogous bound for the lower tail,

$$\mathbb{P}\{X \le \mu - \sigma x\} \le \exp(-x^2/2) \qquad \text{for all } x > 0,$$

leading to the concentration inequality $\mathbb{P}\{|X - \mu| \ge \sigma x\} \le 2e^{-x^2/2}$.

□

> **Remark.** Of course the algebra would have been a tad simpler if I had worked with the standardized variable $(X - \mu)/\sigma$. I did things the messier way in order to make a point about the centering in Section 2.4.

Exactly the same arguments work for any random variable whose moment generating is bounded above by $\exp(\theta\mu + \sigma^2\theta^2/2)$, for all real θ and some finite constants μ and τ. This property essentially defines the ***subgaussian*** distributions, which are discussed in Section 2.4.

The function $L(\theta) := \log M(\theta)$ is called the ***cumulant-generating function***. The upper bound in <2> is often rewritten as

$$e^{-L^*(\theta)} \qquad \text{where } L^*(\theta) := \sup_{\theta \ge 0}(\theta w - L(\theta)).$$

> **Remark.** As Boucheron et al. (2013, Sections 2.2) and others have noted, the function L^* is the Fenchel-Lagrange transform of L, which is also known as the conjugate function (Boyd and Vandenberghe, 2004, Section 3.3).

Clearly $L(0) = 0$. The Hölder inequality shows that L is convex: if θ is a convex combination $\alpha_1\theta_1 + \alpha_2\theta_2$ then

$$\begin{aligned} e^{L(\theta)} &= M(\alpha_1\theta_1 + \alpha_2\theta_2) \\ &= \mathbb{P}\left(e^{\alpha_1\theta_1 X} e^{\alpha_2\theta_2 X}\right) \\ &\le \left(\mathbb{P}e^{\theta_1 X}\right)^{\alpha_1}\left(\mathbb{P}e^{\theta_2 X}\right)^{\alpha_2} = e^{\alpha_1 L(\theta_1) + \alpha_2 L(\theta_2)}. \end{aligned}$$

More precisely (Problem [1]), the set $J = \{\theta : M(\theta) < \infty\}$ is an interval, possibly degenerate or unbounded, containing the origin. For example, for the Cauchy, $M(\theta) = \infty$ for all real $\theta \neq 0$; for the standard exponential distribution (density $e^{-x}\{x > 0\}$ with respect to Lebesgue measure),

$$M(\theta) = \begin{cases} (1-\theta)^{-1} & \text{for } \theta < 1 \\ \infty & \text{otherwise} \end{cases}.$$

Problem [1] also justifies the interchange of differentiation and expectation on the interior of J, which leads to

$$\begin{aligned} L'(\theta) &= \frac{M'(\theta)}{M(\theta)} = \mathbb{P}X\frac{e^{\theta X}}{M(\theta)} \\ L''(\theta) &= \frac{M''(\theta)}{M(\theta)} - \left(\frac{M'(\theta)}{M(\theta)}\right)^2 = \mathbb{P}X^2\frac{e^{\theta X}}{M(\theta)} - \left(\mathbb{P}X\frac{e^{\theta X}}{M(\theta)}\right)^2. \end{aligned}$$

If we define a probability measure $\mathbb{P}_\theta$ by its density, $d\mathbb{P}_\theta/d\mathbb{P} = e^{\theta X}/M(\theta)$, then the derivatives of L can be re-expressed using moments of X under $\mathbb{P}_\theta$:

$$L'(\theta) = \mathbb{P}_\theta X \qquad \text{AND} \qquad L''(\theta) = \mathbb{P}_\theta\left(X - \mathbb{P}_\theta X\right)^2 = \text{var}_\theta(X) \ge 0.$$

In particular, $L'(0) = \mathbb{P}X$, so that Taylor expansion gives

\E@ L.Taylor <4> $$L(\theta) = \theta\mathbb{P}X + \tfrac{1}{2}\theta^2\text{var}_{\theta*}(X) \qquad \text{with } \theta^* \text{ lying between 0 and } \theta.$$

Bounds on the $\mathbb{P}_\theta$ variance of X lead to bounds on L and tail bounds for $X - \mathbb{P}X$. The Hoeffding inequality (Example <14>) provides a good illustration of this line of attack.

2.3 Orlicz norms

Basic::S:orlicz

There is a fruitful way to generalize the the concept of an $\mathcal{L}^p$ norm by means of a ***Young function***, that is, a convex, increasing function $\Psi : \mathbb{R}^+ \to \mathbb{R}^+$ with $\Psi(0) = 0$ and $\Psi(x) \to \infty$ as $x \to \infty$. The concept applies with any measure μ but I need it mostly for probability measures.

Remark. The assumption that $\Psi(x) \to \infty$ as $x \to \infty$ is needed only to rule out the trivial case where $\Psi \equiv 0$. Indeed, if $\Psi(x_0) > 0$ for some x_0 then, for $x = x_0/\alpha$ with $\alpha < 1$,

$$\Psi(x_0) = \Psi(\alpha x + (1-\alpha)0) \le \alpha\Psi(x)$$

That is, $\Psi(x)/x \ge \Psi(x_0)/x_0$ for all $x \ge x_0$. If Ψ is not identically zero then it must increase at least linearly with increasing x.

At one time Orlicz norms became very popular in the empirical process literature, in part (I suspect) because chaining arguments (see Chapter 4) are cleaner with norms than with tail probabilities.

The most important Young functions for my purposes are the power functions, $\Psi(x) = x^p$ for a fixed $p \ge 1$, and the exponential power functions $\Psi_\alpha(x) := \exp(x^\alpha) - 1$ for $\alpha \ge 1$. The function Ψ_2 is particularly useful (see Section 2.4) for working with distributions whose tails decrease like a Gaussian. It is also possible (Problem [6]) to define a Young function Ψ_α, for $0 < \alpha < 1$, for which

$$\Psi_\alpha(x) \le \exp(x^\alpha) \le K_\alpha + \Psi_\alpha(x) \qquad \text{for all } x \ge 0,$$

where K_α is a positive constant. Some linear interpolation near 0 is needed to overcome the annoyance that $x \mapsto \exp(x^\alpha)$ is concave on the interval $[0, z_\alpha]$, for $z_\alpha := ((1-\alpha)/\alpha)^{1/\alpha}$.

Basic::Orlicz.norm <5> **Definition.** *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and f be an $\mathcal{A}$-measuable real function on $\mathcal{X}$. For each Young function Ψ the corresponding **Orlicz norm** $\|f\|_\Psi$ (or Ψ-norm) is defined by*

$$\|f\|_\Psi = \inf\{c > 0 : \mu\Psi(|f(x)|/c) \le 1\},$$

with $\|f\|_\Psi = \infty$ if the infimum runs over an empty set.

Remark. It is not hard to show that $\|f\|_\Psi < \infty$ if and only if $\mu\Psi(|f|/C_0) < \infty$ for at least one finite constant C_0. Moreover, if $0 < c = \|f\|_\Psi < \infty$ then $\mu\Psi(|f|/c) \le 1$; the infimum is achieved.

When restricted to the set $\mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$ of all measurable real functions on $\mathcal{X}$ with finite Ψ-norm, $\|\cdot\|_\Psi$ is a seminorm. It fails to be a norm only because $\|f\|_\Psi = 0$ if and only if $\mu\{x : f(x) \neq 0\} = 0$. The corresponding set of μ-equivalence classes is a Banach space. It is traditional and benign to call $\|\cdot\|_\Psi$ a norm even though it is only a seminorm on $\mathcal{L}^\Psi(\mathcal{X}, \mathcal{A}, \mu)$.

The assumption $\Psi(0) = 0$ can be discarded if μ is a finite measure. For each finite constant $A > \Psi(0)$, the quantity $\inf\{c > 0 : \mu\Psi(|f|/c) \le A\}$ also defines a seminorm.

Finiteness of the Ψ-norm for a random variable Z immediately implies a tail bound: if $0 < \sigma = \|Z\|_\Psi < \infty$ then

\E@ orlicz.tail <6>
$$\mathbb{P}\{|Z| \ge x\sigma\} \le \max\left(1, \mathbb{P}\frac{\Psi(|Z|/\sigma)}{\Psi(x)}\right) = \max\left(1, \frac{1}{\Psi(x)}\right) \qquad \text{for } x \ge 0.$$

For example, for each $\alpha > 0$ there is a constant C_α for which

$$\mathbb{P}\{|Z| \ge x\,\|Z\|_{\Psi_\alpha}\} \le C_\alpha e^{-x^\alpha} \qquad \text{for } x \ge 0.$$

The constant can be chosen so that $C_\alpha e^{-x^\alpha} \ge 1$ for all x so small that $\Psi_\alpha(x)e^{-x^\alpha}$ is not close to 1.

When seeking to bound an Orlicz norm I often find it easier to first obtain a constant C_0 for which $\mathbb{P}\Psi(|X|/C_0) \le K_0$, where K_0 is a constant strictly larger than 1. As the next Lemma shows, a simple convexity argument turns such an inequality into a bound on $\|X\|_\Psi$.

Basic::Psi.bound <7> **Lemma.** *Suppose Ψ is a Young function. If $\mathbb{P}\Psi(|X|/C_0) \le K_0$ for some finite constants C_0 and $K_0 > 1$ then $\|X\|_\Psi \le C_0 K_0$.*

Proof For each θ in $[0,1]$ convexity of Ψ gives

$$\mathbb{P}\Psi\left(\frac{\theta|X|}{C_0}\right) \le \theta\mathbb{P}\Psi\left(\frac{|X|}{C_0}\right) + (1-\theta)\Psi(0) \le \theta K_0.$$

The choice $\theta = 1/K_0$ makes the last bound equal to 1.

□

Beware. The Lemma does not give a foolproof way of determining reasonable bounds for the Ψ-norm. For example, if X has a $N(0,1)$ distribution then simple integration shows that

$$\mathbb{P}\Psi_2(|X|/c) = g(c) := c/\sqrt{c^2-2} - 1 \qquad \text{for } c > \sqrt{2}.$$

Thus $\|X\|_{\Psi_2} = \sqrt{8/3}$ but $cg(c) \to \infty$ as c decreases to $\sqrt{2}$.

2.4 Subgaussian tails

Basic::S:subgaussian

As noted in Section 2.2, if X is a random variable for which there exist constants μ and σ^2 for which $\mathbb{P}e^{\theta X} \le \exp(\mu\theta + \frac{1}{2}\sigma^2\theta^2)$ for all real θ then

\E@ subg.oneside <8>
$$\mathbb{P}\{X \ge \mu + \sigma x\} \le e^{-x^2/2} \qquad \text{for all } x \ge 0.$$

The approximations for θ near zero,

$$\mathbb{P}e^{\theta X} = 1 + \theta\mathbb{P}X + \tfrac{1}{2}\theta^2\mathbb{P}X^2 + o(\theta^2),$$
$$\exp(\mu\theta + \tfrac{1}{2}\sigma^2\theta^2) = 1 + \mu\theta + \tfrac{1}{2}(\sigma^2 + \mu^2)\theta^2 + o(\theta^2),$$

would force $\mathbb{P}X = \mu$ and $\mathbb{P}X^2 \le \sigma^2 + \mu^2$, that is, $\mathrm{var}(X) \le \sigma^2$. As Problem [11] shows, σ^2 might need to be much larger than the variance. To avoid any hint of the convention that σ^2 denotes a variance, I feel it is safer to replace σ by another symbol.

Basic::subgaussian <9> **Definition.** *Say that a random variable X with $\mathbb{P}X = \mu$ has a* ***subgaussian distribution*** *with scale factor $\tau < \infty$ if $\mathbb{P}e^{\theta X} \le \exp(\mu\theta + \tau^2\theta^2/2)$ for all real θ. Write $\tau(X)$ for the smallest τ for which such a bound holds.*

> **Remark.** Some authors call such an X a τ-subgaussian random variable. Some authors require $\mu = 0$ for subgaussianity; others use the qualifier *centered* to refer to the case where μ is zero.

For a subgaussian X, the two-sided analog of <8> gives

\E@ beta.def <10>
$$\mathbb{P}\{|X - \mu| \ge x\} \le 2\exp\left(-\frac{x^2}{2\beta^2}\right) \qquad \text{for all } x \ge 0,$$

for $0 < \beta \le \tau(X)$. In fact (Theorem <16>) existence of such a tail bound is equivalent to subgaussianity.

Basic::gaussian <11> **Example.** Suppose $Y = (Y_1, \dots, Y_n)$ has a multivariate normal distribution. Define $M := \max_{i \le n} |Y_i|$ and $\sigma^2 := \max_{i \le n} \mathrm{var}(Y_i)$. Chapter 6 explains why

$$\mathbb{P}\{|M - \mathbb{P}M| \ge \sigma x\} \le 2\exp(-x^2/2) \qquad \text{for all } x \ge 0.$$

That is, M is subgaussian—a most surprising and wonderful fact.

□

Many useful results that hold for gaussian variables can be extended to subgaussians. In empirical process theory, conditional subgaussianity plays a key role in symmetrization arguments (see Chapter 8). The first example captures the main idea.

Basic::Rademachers <12> **Example.** Suppose $X = \sum_{i \le n} b_i \mathfrak{s}_i$, where the b_i's are constants and the $\mathfrak{s}_i$'s are independent ***Rademacher*** random variables, that is, $\mathbb{P}\{\mathfrak{s}_i = +1\} = \frac{1}{2} = \mathbb{P}\{\mathfrak{s}_i = -1\}$. Then

$$\mathbb{P}e^{\theta b_i \mathfrak{s}_i} = \tfrac{1}{2}\left(e^{\theta b_i} + e^{-\theta b_i}\right) = \sum_{k=0}^{\infty} \frac{(\theta b_i)^{2k}}{(2k)!} \le \exp(\theta^2 b_i^2/2)$$

so that $\mathbb{P}e^{\theta X} \le \prod_{i\le n} \mathbb{P}e^{\theta b_i \mathfrak{s}_i} \le e^{\theta^2\tau^2/2}$ with $\tau^2 = \sum_{i\le n} b_i^2$. That is, X has a centered subgaussian distribution with $\tau(X)^2 \le \sum_{i\le n} b_i^2$.

□

More generally, if X is a sum of independent subgaussians $X_1,\dots,X_n$ then the equality

$$\mathbb{P}e^{\theta(X-\mathbb{P}X)} = \prod\nolimits_i \mathbb{P}e^{\theta(X_i-\mathbb{P}X_i)}$$

shows that X is subgaussian with $\tau^2(X) \le \sum_i \tau^2(X_i)$. For sums of independent variables we need consider only the subgaussianity of each summand.

Basic::interval.range <13> **Example.** Suppose X is a random variable with $\mathbb{P}X = \mu$ and $a \le X \le b$, for finite constants a and b. By representation <4>, its moment generating function $M(\theta)$ satisfies

$$\log M(\theta) = \theta\mathbb{P}X + \tfrac{1}{2}\theta^2 \text{var}_{\theta*}(X) \qquad \text{with } \theta^* \text{ lying between 0 and } \theta.$$

Write μ^* for $\mathbb{P}_{\theta^*}X$ and m for $(b-a)/2$. Note that $|X-m| \le (b-a)/2$. Thus

$$\text{var}_{\theta*}(X) = \mathbb{P}_{\theta^*}|X-\mu^*|^2 \le \mathbb{P}_{\theta^*}|X-m|^2 \le (b-a)^2/4.$$

and $\log \mathbb{P}e^{\theta(X-\mu)} \le (b-a)^2\theta^2/8$. The random variable X is subgaussian with $\tau(X) \le (b-a)/2$.

□

Remark. Hoeffding (1963, page 22) derived the same bound on the moment generating function by a direct appeal to convexity of the exponential function:

$$\mathbb{P}e^{\theta X} \le qe^{\theta a} + pe^{\theta b} \qquad \text{where } 1-q = p = \frac{\mathbb{P}X - a}{b-a}$$

In effect, the bound represents the extremal case where X concentrates on the the two-point set $\{a,b\}$. He then calculated a second derivative for L, without interpreting the results as a variance but using an inequality that can be given a variance interpretation.

Basic::Hoeffding.ineq <14> **Example.** (Hoeffding's inequality for independent summands) Suppose random variables $X_1,\dots,X_n$ are independent with $\mathbb{P}X_i = \mu_i$ and $a_i \le X_i \le b_i$ for each i, for constants a_i and b_i.

By Example <13>, each X_i is subgaussian with $\tau(X_i) \le (b_i - a_i)/2$. The sum $S = \sum_{i\le n} X_i$ is also subgaussian, with expected value $\mu = \sum_{i\le n} \mu_i$

and $\tau^2(S) = \sum_{i\le n}\tau^2(X_i) = \sum_{i\le n}(b_i-a_i)^2/4$. The one-sided version of inequality <10> gives

\E@ Hoeff.indep <15>

$$\mathbb{P}\left\{\sum\nolimits_{i\le n}(X_i-\mu_i)\ge x\right\} \le \exp\left(-2x^2/\sum\nolimits_{i\le n}(b_i-a_i)^2\right) \qquad \text{for each } x\ge 0.$$

See Chapter 3 for an extension of Hoeffding's inequality <15> to sums of bounded martingale differences.

□

The next Theorem provides other characterizations of subgaussianity, using the Orlicz norm for the Young function $\Psi_2(x) = \exp(x^2)-1$ and moment inequalities. The proof of the Theorem provides a preview of a "symmetrization" argument that plays a major role in Chapter 8. I prefer to write the argument using product measures rather than conditional expectations. Here is the idea. Suppose X_1 is a random variable with $\mathbb{P}X_1 = \mu$, defined on a probability space $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$. Make an "independent copy" $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ of the probability space then carry X_1 over to a new random variable X_2 on Ω_2. Under $\mathbb{P}_1\otimes\mathbb{P}_2$ the random variables X_1 and X_2 are independent and identically distributed.

Remark. You might prefer to start with an X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ then define random variables on $\Omega\times\Omega$ by $X_1(\omega_1,\omega_2) = X(\omega_1)$ and $X_2(\omega_1,\omega_2) = X(\omega_2)$. Under $\mathbb{P}\otimes\mathbb{P}$ the random variables X_1 and X_2 are independent and each has the same distribution as X (under $\mathbb{P}$). I once referred to independent copies during a seminar in a Math department. One member of the audience protested vehemently that a copy of X is the same as X and hence can only be independent of X in trivial cases.

Note that $\mathbb{P}_2X_2 = \mathbb{P}_1X_1 = \mu$. For each convex function f,

$$\begin{aligned}\mathbb{P}_1 f(X_1-\mu) &= \mathbb{P}_1^{\omega_1} f\left[X_1(\omega_1) - \mathbb{P}_2^{\omega_2}X_2(\omega_2)\right]\\ &= \mathbb{P}_1^{\omega_1} f\left[\mathbb{P}_2^{\omega_2}\left(X_1(\omega_1) - X_2(\omega_2)\right)\right]\\ &\le \mathbb{P}_1^{\omega_1}\mathbb{P}_2^{\omega_2} f\left[X_1(\omega_1) - X_2(\omega_2)\right],\end{aligned}$$

the last line coming from Jensen's inequality for $\mathbb{P}_2$.

Basic::subgauss.equiv <16> **Theorem.** *For each random variable X with expected value μ, the following assertions are equivalent.*

(i) X has a subgaussian disribution (with $\tau(X)<\infty$)

Basic::L.finite *(ii)* $L(X) := \sup_{r\ge1}\dfrac{\|X\|_r}{\sqrt{r}} < \infty$

Basic::Psitwo.finite *(iii)* $\|X\|_{\Psi_2} < \infty$

Basic::beta.def2 *(iv) there exists a positive constant* β *for which* $\mathbb{P}\{|X-\mu| \ge x\} \le 2\exp\left(-\dfrac{x^2}{2\beta^2}\right)$ *for all* $x \ge 0$

More precisely,

Basic::Psi2L.equiv *(a)* $c_0 \|X\|_{\Psi_2} \le L(X) \le \|X\|_{\Psi_2}$ *with* $c_0 = 1/\sqrt{4e}$*;*

Basic::centered *(b) if* $\beta(X)$ *denotes the smallest* β *for which inequality (iv) holds,*

$$c_1 \|X-\mu\|_{\Psi_2} \le \beta(X) \le \tau(X) \le 4L(X-\mu) \qquad \text{with } c_1 = 1/\sqrt{6}.$$

> **Remark.** Finiteness of $L(X)$ is also equivalent to finiteness of $L(X-\mu)$ because $|\mu| \le \|X\|_1 \le L(X)$ and $|L(X)-L(X-\mu)| \le |\mu|$. Consequently $L(X-\mu) \le 2L(X)$, but there can be no universal inequality in the other direction: $L(X-\mu)$ is unchanged by addition of a constant to X but $L(X)$ can be made arbitrarily large.
>
> The precise values of the constants c_i are not important. Most likely they could be improved, but I can see no good reason for trying.

PROOF Implicit in the sequences of inequalities is the assertion that finiteness of one quantity implies finiteness of another quantity.

Proof of (a)

Problem [8] shows that $k^k \ge k! \ge (k/e)^k$ for all $k \in \mathbb{N}$.

Suppose $\mathbb{P}\Psi_2(|X|/D) \le 1$. Then $\sum_{k\in\mathbb{N}} \mathbb{P}(X/D)^{2k}/k! \le 1$ so that

$$\|X\|_{2k} \le D(k!)^{1/2k} \le D\sqrt{k} \qquad \text{for each } k \in \mathbb{N}$$

and $L(X) \le \sqrt{2}\sup_{k\in\mathbb{N}} \|X\|_{2k}/\sqrt{2k} \le D$ by Problem [9].

Conversely, if $\infty > L \ge \|X\|_r/\sqrt{r}$ for all $r \ge 1$ then

$$\mathbb{P}\Psi_2\left(\frac{|X|}{D}\right) = \sum_{k\in\mathbb{N}} \frac{\|X/D\|_{2k}^{2k}}{k!} \le \sum_{k\in\mathbb{N}} (L\sqrt{2k}/D)^{2k}(e/k)^k,$$

which equals 1 if $D = \sqrt{4e}L$.

Proof of (b)

None of the quantities appearing in the inequalities depend on μ, so we may as well assume $\mu = 0$.

Inequality <10> gives $\beta(X) \le \tau(X)$.

Simplify notation by writing β for $\beta(X)$ and τ for $\tau(X)$ and L for $L(X)$, and define $\gamma := \|X\|_{\Psi_2}$. Also, assume that X is not a constant, to avoid annoying trivial cases.

To show that $\gamma \le \beta\sqrt{6}$:

$$\begin{aligned}\mathbb{P}\Psi_2\left(\frac{|X-\mu|}{\beta\sqrt{6}}\right) &= \mathbb{P}\int_0^\infty e^x\{|X-\mu|^2 \ge 6\beta^2 x\}\,dx \\ &= \int_0^\infty e^x \mathbb{P}\{|X-\mu| \ge \beta\sqrt{6x}\}\,dx \\ &\le 2\int_0^\infty \exp\left(x-3x\right)\,dx = 1.\end{aligned}$$

To show that $\tau \le 4L$: Symmetrize. Construct a random variable X' with the same distribution as X but independent of X. By the triangle inequality, $\|X-X'\|_r \le 2\|X\|_r \le 2L\sqrt{r}$ for all $r \ge 1$. By Jensen's inequality

$$\begin{aligned}\mathbb{P}e^{\theta X} &\le \mathbb{P}e^{\theta(X-X')} \\ &= 1 + \sum\nolimits_{k\in\mathbb{N}} \frac{\theta^{2k}\mathbb{P}(X-X')^{2k}}{(2k)!} \qquad \text{symmetry kills odd moments} \\ &\le 1 + \sum\nolimits_{k\in\mathbb{N}} \frac{(2L\sqrt{2k})^{2k}}{k^k k!} \\ &= 1 + \sum\nolimits_{k\in\mathbb{N}} \frac{(8\theta^2 L^2)^k}{k!} = e^{8L^2\theta^2},\end{aligned}$$

and so on.

□

2.5 Bennett's inequality

Basic::S:Bennett

The Hoeffding inequality depends on the somewhat crude bound

$$\text{var}(W) \le (b-a)^2/4 \qquad \text{if } a \le W \le b.$$

Such an inequality can be too pessimistic if W has only a small probability of taking values near the endpoints of the interval $[a,b]$. The inequalities can sometimes be improved by introducing second moments explicitly into the upper bound. Bennett's one-sided tail bound provides a particularly elegant illustration of the principle.

Basic::Bennett.ineq <*17*> **Theorem.** *Suppose $X_1, \dots, X_n$ are independent random variables for which $X_i \le b$ and $\mathbb{P}X_i = \mu_i$ and $\mathbb{P}X_i^2 \le v_i$ for each i, for nonnegative constants b and v_i. Let W equal $\sum_{i\le n} v_i$. Then*

$$\mathbb{P}\left\{\sum\nolimits_{i\le n}(X_i-\mu_i) \ge x\right\} \le \exp\left(-\frac{x^2}{2W}\psi_{\text{Benn}}\left(\frac{bx}{W}\right)\right) \qquad \text{for } x \ge 0$$

where ψ_{Benn} denotes the function defined on $[-1,\infty)$ by

\E@ Benpsi.def <18>

$$\psi_{\text{Benn}}(t) := \frac{(1+t)\log(1+t)-t}{t^2/2} \qquad \text{for } t \ne 0, \text{ and } \psi_{\text{Benn}}(0) = 1.$$

> **Remark.** By Problem [14], ψ_{Benn} is positive, decreasing, and convex, with $\psi_{\text{Benn}}(-1) = 2$ and $\psi_{\text{Benn}}(t)$ decreasing like $(2/t)\log t$ as $t \to \infty$.
>
> When bx/W is close to zero the ψ_{Benn} contribution to the upper bound is close to 1 and the exponent is close to a subgaussian $-x^2/(2W)$. (If $b = 0$ then it is exactly subgaussian.) Further out in the tail, for larger x, the exponent decreases like $-x\log(bx/W)/b$, which is slightly faster than exponential.

Proof Define $\Phi_1(x) = e^x - 1 - x$, the convex function Ψ_1 adjusted to have zero derivative at the origin, as in Problem [7]. From Problem [14], the positive function

\E@ Del.def <19>

$$\Delta(x) = \frac{\Phi_1(x)}{x^2/2} \qquad \text{for } x \ne 0, \text{ and } \Delta(0) = 1,$$

is continuous and increasing on the whole real line. For $\theta \ge 0$,

$$\begin{aligned}
\mathbb{P}e^{\theta X_i} &= 1 + \theta\mathbb{P}X_i + \mathbb{P}\Phi_1(\theta X_i) \\
&= 1 + \theta\mu_i + \mathbb{P}\left(\tfrac{1}{2}(\theta X_i)^2\Delta(\theta X_i)\right) \\
&\le 1 + \theta\mu_i + \tfrac{1}{2}\theta^2\mathbb{P}(X_i^2)\Delta(\theta b) \qquad \text{because } \Delta \text{ is increasing} \\
&\le \exp\Big(\theta\mu_i + \tfrac{1}{2}\theta^2 v_i\Delta(\theta b)\Big).
\end{aligned}$$

\E@ Bennett.mgf <20>

That is,

\E@ bdd1 <21>

$$\mathbb{P}e^{\theta(X_i-\mu_i)} \le \exp\left(\tfrac{1}{2}\theta^2 v_i\Delta(\theta b)\right) = \exp\left(\frac{v_i}{b^2}\Phi_1(\theta b)\right) \qquad \text{for } \theta \ge 0.$$

> **Remark.** In the last expression I am implicitly assuming that b is nonzero. We can recover the bound when $b = 0$ by a passage to the limit as b decreases to zero. Actually, the $b = 0$ case is quite noteworthy because the Δ contribution disappears from the upper bound for $\mathbb{P}\exp(\theta X_i)$ and the ψ_{Benn} disappears from the final inequality.

Abbreviate $\sum_{i\le n}(X_i-\mu_i)$ to T. By independence and <21>,

$$\mathbb{P}e^{\theta T}=\prod\nolimits_{i\le n}\mathbb{P}e^{\theta(X_i-\mu_i)}\le\exp\left(\frac{W}{b^2}\Phi_1(\theta b)\right)$$

and, by <2>,

$$\begin{aligned}\mathbb{P}\{T\ge x\}&\le\inf\nolimits_{\theta\ge0}\exp\left(-x\theta+\frac{W}{b^2}\Phi_1(\theta b)\right)\\&=\exp\left(-\frac{W}{b^2}\sup_{b\theta\ge0}\left(\frac{bx}{W}b\theta-\Phi_1(\theta b)\right)\right)\\&=\exp\left(-\frac{W}{b^2}\Phi_1^*(bx/W)\right),\end{aligned}$$

where Φ_1^* denotes the conjugate of Φ_1,

$$\Phi_1^*(y):=\sup_{t\in\mathbb{R}}\left(ty-\Phi_1(t)\right)=\begin{cases}(1+y)\log(1+y)-y & \text{if } y\ge-1\\ +\infty & \text{otherwise}\end{cases}.$$

When $y\ge-1$ the supremum is achieved at $t=\log(1+y)$. Thus $\frac{1}{2}y^2\psi_{\text{Benn}}(y)=\Phi_1^*(y)$ for $y\ge-1$ and $(W/b^2)\Phi_1^*(bx/W)=x^2/(2W)\psi_{\text{Benn}}(bx/W)$, as asserted.

□

Remark. The inequality still holds if $0>b\ge-x/W$, although the proof changes slightly. What happens for smaller b?

Unlike the case of the two-sided Hoeffding inequality, there is no automatic lower tail analog for the Bennett inequality without an explicit lower bound for the summands. One particularly interesting case occurs when $X_i\ge0$:

$$\mathbb{P}\{\sum\nolimits_{i\le n}(X_i-\mu_i)\le-x\}=\mathbb{P}\{\sum\nolimits_{i\le n}(\mu_i-X_i)\ge x\}\le\exp(-\epsilon^2/(2W))$$

for all $x\ge0$, a clean (one-sided) subgaussian tail bound.

The Bennett inequality also works for the Poisson distribution. If X has a Poisson(λ) distribution then (Problem [15])

$$\mathbb{P}\{X\ge\lambda+x\}\exp\left(-\frac{x^2}{2\lambda}\psi_{\text{Benn}}\left(\frac{x}{\lambda}\right)\right)\qquad\text{for } x\ge0$$

and, for $0\le x\le\lambda$,

$$\mathbb{P}\{X\le\lambda-x\}=\exp\left(-\frac{x^2}{2\lambda}\psi_{\text{Benn}}\left(-\frac{x}{\lambda}\right)\right)\le\exp\left(-\frac{x^2}{2\lambda}\right),$$

a perfect subgaussian lower tail. See Problem [16] if you have any doubts about the boundedness of Poisson random variables.

See Chapter 3 for an extension of the Bennett inequality to sums of martingale differences.

2.6 Bernstein's inequality

Basic::S:Bernstein

As shown by Problem [14], $\psi_{\text{Benn}}(t) \geq (1+t/3)^{-1}$ for $t \geq -1$. Thus the Bennett inequality implies a weaker result: if $X_1, \dots, X_n$ are independent with $\mathbb{P}X_i = \mu_i$ and $\mathbb{P}X_i^2 \leq v_i$ and $X_i \leq b$ then

\E@ Bernstein.indep0 <22>

$$\mathbb{P}\{\textstyle\sum_{i\leq n}(X_i - \mu_i) \geq x\} \leq \exp\left(-\frac{x^2/2}{W + bx/3}\right) \qquad \text{for } x \geq 0,$$

a form of ***Bernstein's inequality***. When bx/W is close to zero the exponent is again close to $x^2/(2W)$; when bx/W is large, the Bernstein inequality loses the extra log term in the exponent. If the summands X_i are bounded in absolute value (not just bound above) we get an analogous two-sided inequality.

The assumptions $\mathbb{P}X_i^2 \leq v_i$ and $X_i \leq b$ imply

$$\mathbb{P}|X_i|^k \leq \mathbb{P}X_i^2 b^{k-2} \leq v_i b^{k-2} \qquad \text{for } k \geq 2.$$

A much weaker condition on the growth of $\mathbb{P}|X_i|^k$ with k, without any assumption of boundedness, leads to a useful upper bound for the moment generating function and a more genral version of the Bernstein ineqi=uality..

Basic::Bernstein.unbdd *<23>*

Theorem. *(Bernstein inequality) Suppose $X_1, \dots, X_n$ are independent random variables with $\mathbb{P}X_i = \mu_i$ and*

\E@ Bernstein.moment <24>

$$\mathbb{P}|X_i|^k \leq \tfrac{1}{2} v_i B^{k-2} k! \qquad \text{for } k \geq 2.$$

Define $W = \sum_{i\leq n} v_i$. Then, for $x \geq 0$,

$$\mathbb{P}\left\{\textstyle\sum_{i\leq n}(X_i - \mu_i) \geq x\right\} \leq e^{-H_{\text{Benn}}(x,B,W)} \leq e^{-H_{\text{Bern}}(x,B,W)}$$

where

$$H_{\text{Bern}}(x,B,W) = \frac{x^2/W}{2(1+xB/W)}$$

$$H_{\text{Benn}}(x,B,W) = \frac{x^2/W}{1 + xB/W + \sqrt{1+2xB/W}}$$

Remark. The B here does not represent an upper bound for X_i. The traditional form of Bernstein's inequality, with H_{Bern}, corresponds to the bound <22> derived from the Bennett inequality under the assumption $X_i \le b = B/3$.

Proof As before,

$$\bigstar := \mathbb{P}\left\{\sum\nolimits_i X_i \ge x + \sum\nolimits_i \mu_i\right\} \le \inf_{\theta \ge 0} e^{-(x+\sum_i \mu_i)\theta} \prod\nolimits_{i\le n} \mathbb{P}e^{\theta X_i}.$$

Use the moment inequality <24> to bound the moment generating function for X_i.

$$\mathbb{P}e^{\theta X_i} = 1 + \theta\mu_i + \mathbb{P}\Phi_1(\theta X_i) \qquad \text{where } \Phi_1(t) := e^t - 1 - t$$

and, for $\theta \ge 0$,

\E@ Bernstein.mgf

$$\begin{align*}\mathbb{P}\Phi_1(\theta X_i) \le \mathbb{P}\Phi_1(\theta|X_i|) &= \sum\nolimits_{k\ge 2} \frac{\theta^k \mathbb{P}|X_i|^k}{k!} \\ &\le \tfrac{1}{2} v_i \theta^2 \sum\nolimits_{k\ge 2} (\theta B)^{k-2} \\ &= \tfrac{1}{2} \frac{v_i\theta^2}{1 - B\theta} \qquad \text{provided } 0 \le \theta < 1/B. \tag*{<25>}\end{align*}$$

Thus

$$\mathbb{P}e^{\theta X_i} \le 1 + \theta\mu_i + \tfrac{1}{2}\frac{v_i\theta^2}{1 - b\theta} \le \exp\left(\theta\mu_i + \tfrac{1}{2}\frac{v_i\theta^2}{1 - b\theta}\right) \qquad \text{for } 0 \le \theta B < 1$$

and

$$\bigstar \le \exp\left(-\sup\nolimits_{0\le B\theta<1} g(\theta)\right) \qquad \text{where } g(\theta) := x\theta - \tfrac{1}{2}\frac{W\theta^2}{1 - B\theta}.$$

The H_{Bern} bound

As an approximate maximizer for g choose $\theta_0 = x/(W + xB)$ so that

$$\bigstar \le e^{-g(\theta_0)} = \exp\left(-\frac{x^2/2}{W + xB}\right).$$

Bennett (1962, equation (7)) made this strange choice by defining $G = 1 - \theta B$ then minimizing $-x\theta + \frac{1}{2}\theta^2 W/G$ while ignoring the fact that G depends on θ. That gives $\theta = xG/W = (1 - \theta B)x/W$, a linear equation with solution θ_0. Very confusing. As Bennett noted, a boundedness assumption $|X_i| \le b$

actually implies <24> with $B = b/3$. The exponential upper bound then agrees with <22>.

The H_{Benn} bound
Following Bennett (1962, equation (7a)), this time we really maximize $g(\theta)$. The calculation is cleaner if we rewrite g with $u := xB/W$ and $s := 1-\theta B$. With those substitutions,

$$g(\theta) = g_0(s) = \frac{W}{2b^2}\left(2us - (1-s)^2/s\right) = \frac{W}{2B^2}\left(-s^{-1} - (1+2u)s - 2(1+u)\right).$$

The maximum is achieved at $s_0 = (1+2u)^{-1/2}$ giving $\bigstar \le e^{-g_0(s_0)}$ where

\E@ u.bound <26>
$$g_0(s_0) = \frac{W}{B^2}\left(1+u-\sqrt{1+2u}\right) = \frac{W}{B^2}\left(\frac{u^2}{1+u+\sqrt{1+2u}}\right).$$

□

Both inequalities in Theorem <23> have companion lower bounds, obtained by replacing X_i by $-X_i$ in all the calculations. This works because the Bernstein condition <24> also holds with X_i replaced by $-X_i$. The inequalities could all have been written as two-sided bounds, that is, as upper bounds for $\mathbb{P}\{|\sum_i(X_i-\mu_i)| \ge x\}$.

If we are really only interested in a one-sided bound then it seems superfluous to control the size of $|X_i|$. For example, for the upper tail it should be enough to control X_i^+. Boucheron et al. (2013, page 37)=BLM derived such a bound (which they attributed to Emmanuel Rio) by means of the elementary inequality

$$e^x - 1 - x \le \tfrac{1}{2}x^2 + \sum\nolimits_{k\ge 3}(x+)^k/k! \qquad \text{for all real } x.$$

(For $x \ge 0$ both sides are equal. For $x < 0$ it reflects the fact that $g(x) = e^x - (1 + x + \frac{1}{2}x^2)$ has a positive derivative with $g(0) = 0$.) They replaced the assumption <24> by

$$\mathbb{P}X_i^2 = v_i \quad \text{AND} \quad \mathbb{P}|X_i|^k \le \tfrac{1}{2}v_iB^{k-2}k! \qquad \text{for } k \ge 3.$$

For $B > \theta \ge 0$, we then get

$$\mathbb{P}e^{\theta X_i} \le 1 + \theta\mu_i + \tfrac{1}{2}\theta^2 v_i + \tfrac{1}{2}v_i\theta^2\sum_{k\ge 3}(\theta B)^{k-2} \le \exp\left(\theta\mu_i + \tfrac{1}{2}\frac{v_i\theta^2}{1-\theta B}\right),$$

which is exactly the same inequality as in the proof of Theorem <23>. As with the H_{Benn} calculation it follows that

$$\mathbb{P}\left\{\sum\nolimits_i X_i \geq x + \sum\nolimits_i \mu_i\right\} \leq \exp\left(-\frac{W}{B^2}\left(1+u-\sqrt{1+2u}\right)\right)$$

where $W = \sum_i v_i$ and $u = xB/W$. BLM observed that the upper bound can be inverted by solving a quadratic in u,

$$1+2u = \left(1+u-tB^2/W\right)^2,$$

which has a positive solution $u = tB^2/W + \sqrt{2tB^2/W}$, corresponding to $x = Wu/B = tB + \sqrt{2tW}$. The one-sided Bernstein bound then takes an elegant form,

\E@ one.sided.Bernstein <27>

$$\mathbb{P}\left\{\sum\nolimits_i (X_i - \mu_i) \geq tB + \sqrt{2tW}\right\} \leq e^{-t} \qquad \text{for } t \geq 0.$$

By splitting according to whether $tB \geq \sqrt{2tW}$ or not, and by introducing an extraneous factor of 2, we can split the elegant form into two more suggestive inequalities,

$$\begin{aligned} \mathbb{P}\{\sum\nolimits_i (X_i - \mu_i) \geq 2tB\} &\leq e^{-t} \qquad \text{for } t \geq 2W/B^2 \\ \mathbb{P}\{\sum\nolimits_i (X_i - \mu_i) \geq 2t\sqrt{W}\} &\leq e^{-t^2/2} \qquad \text{for } t^2/2 \geq 2W/B^2. \end{aligned}$$

These inequalities make clear that the subexponential part (large t) of the tail is controlled by B and the subgaussian part (moderate t) is controlled by W.

2.7 Subexponential distributions?

Basic::S:subexp

To me, subexponential behavior refers to functions that are bounded above by a constant multiple of $\exp(-c|x|)$, for some positive constant c. For example, the Bernstein inequality shows that, far enough from the mean, the tail probabilities for a large class of sums of independent random variables are subexponential. Closer to the mean the tail are subgaussian, which one might think is better than subexponential. As another example, if a random variable X takes values in $[-1,+1]$ we have $\mathbb{P}\{|X| \geq x\} = 0 \leq \exp(-10^{10}x)$ for all $x \geq 1$, but I would hesitate to call that behavior subexponential.

What then does it mean for a whole distribution to be subexponential? And how is subexponentiality related to the Bernstein moment condition,

\E@ Bernstein.moment0 <28>

$$\mathbb{P}\frac{|X|^k}{B^k k!} \leq \tfrac{1}{2}v/B^2 \qquad \text{for } k \geq 2?$$

As you already know, condition <28> implies

\E@ Bernstein.mgf0 <29>

$$\mathbb{P}e^{\theta X} - 1 - \theta \mathbb{P}X = \mathbb{P}\Phi_1(\theta X) \le \mathbb{P}\Phi_1(\theta |X|) \le \tfrac{1}{2}\frac{v\theta^2}{1-\theta B} \qquad \text{for } 0 \le \theta B < 1,$$

which implies

$$\mathbb{P}e^{\theta X} \le \exp\left(\theta\mu + \tfrac{1}{2}\frac{v\theta^2}{1-\theta B}\right) \qquad \text{for } 0 \le \theta B < 1 \text{ and } \mu = \mathbb{P}X.$$

Van der Vaart and Wellner (1996, page 103) pointed out that condition <28> is implied by

$$\tfrac{1}{2}v/B^2 \ge \mathbb{P}\Phi_1(|X|/B) = \sum_{k\ge 2}\frac{\mathbb{P}|X|^k}{B^k k!} \qquad \text{where } \Phi_1(x) := e^x - 1 - x.$$

And, if <28> holds, then

$$\mathbb{P}\Phi_1(|X|/2B) = \sum_{k\ge 2}\frac{\mathbb{P}|X|^k}{2^k B^k k!} \le \tfrac{1}{2}(2v)/(2B)^2,$$

which is just <29> with $\theta = 1/(2B)$. It would seem that the moment condition is somehow related to $\|X\|_{\Phi_1}$. Indeed, Vershynin (2010, Section 5.2.4) defined a random variable X to be subexponential if $\sup_{k\in\mathbb{N}} \|X\|_k /k < \infty$. By Problems [7] and [10], this condition is equivalent to finiteness of the Orlicz norm $\|X\|_{\Phi_1}$.

Boucheron et al. (2013, Section 2.4) defined a centered random variable X to be *sub-gamma on the right tail with variance factor v and scale factor B* if

\E@ sub.gamma <30>

$$\mathbb{P}e^{\theta X} \le \exp\left(\frac{v\theta^2/2}{1-\theta B}\right) \qquad \text{for } 0 \le \theta B < 1.$$

They pointed out that if X has a gamma(α,β) distribution (with mean $\mu = \alpha\beta$ and variance $v = \alpha\beta^2$) then

$$\begin{aligned}\mathbb{P}e^{\theta(X-\mu)} = e^{-\theta\mu}(1-\theta\beta)^{-\alpha} &= \exp\left(-\theta\alpha\beta - \alpha\log(1-\theta\beta)\right)\\ &\le \exp\left(\frac{\alpha\beta^2\theta^2/2}{1-\theta\beta}\right) \qquad \text{for } 0 \le \theta\beta < 1.\end{aligned}$$

Hence the name. The special case where $\alpha = 1$ corresponds to the exponential.

Let me try to tie these definitions and facts together.

First note that the upper bound in <29> equals 1 if $\theta = 2/\left(B + \sqrt{B^2 + 2v}\right)$. Thus the Bernstein assumption <28> implies

$$\gamma := \|X\|_{\Phi_1} \le \left(B + \sqrt{B^2 + 2v}\right)/2.$$

If, as is sometimes assumed, $v = \mathbb{P}X^2$ then <28> also implies

$$v^{3/2} \le \mathbb{P}|X|^3 \le \tfrac{1}{2}v3!B^2.$$

That is, $v \le (3B)^2$ and $B < \gamma \le 2.7B$. It appears that the B, or $\|X\|_{\Phi_1}$, controls the subexponential part of the tail probabilities, at least as far as the Bernstein inequality is concerned. Inequality <27> conveyed the same message. Compare with the subexponential bound

$$\mathbb{P}\{|X| \ge x\} \le \min\left(1, 1/\Phi_1(x/\gamma)\right) \qquad \text{for } \gamma = \|X\|_{\Phi_1}.$$

The inequality $\mathbb{P}\Phi_1(|X|/\gamma) \le 1$ also implies

$$\mathbb{P}\frac{|X|^k}{\gamma^k k!} \le 1 \le \tfrac{1}{2}(2\gamma^2)/\gamma^2 \qquad \text{for } k \ge 2.$$

That is, <28> holds with $v = 2\gamma^2$ and $B = \gamma$. Theorem <23> then gives

$$\begin{aligned}\mathbb{P}\{X - \mu \ge x\} &\le \exp\left(-\frac{x^2}{4\gamma^2 + 2x\gamma}\right)\\ &\le \exp(-x^2/(8\gamma^2)) \qquad \text{for } 0 \le x \le 2\gamma.\end{aligned}$$

Is that not a subgaussian bound? Unfortunately, the bound—subgaussian or not—has little value for such a small range near zero. The upper bound always exceeds $1/\sqrt{e}$.

The useful subgaussian part of the bound is contributed by the v in <28>. The effect is clearer for a sum of independent random variables $X_1, \dots, X_n$, each distributed like X:

$$\mathbb{P}\exp(\theta \textstyle\sum_{i\le n} X_i) \le \exp\left(\dfrac{nv\theta^2/2}{1-\theta B}\right) \qquad \text{for } 0 \le \theta B < 1,$$

which leads to tail bounds like

$$\mathbb{P}\{\textstyle\sum_i X_i \ge x\} \le \exp\left(-\dfrac{x^2/2}{nv + xB}\right),$$

useful subgaussian behavior when $0 \le x \le nv/B$.

In my opinion, the sub-gamma property <30> is better thought of as a restricted subgaussian assumption that automatically implies subexponential behavior far out in the tails. For each α in $(0,1)$ it implies

$$\mathbb{P}e^{\theta X} \le \exp\left(\frac{v\theta^2/2}{1-\alpha}\right) \qquad \text{for } 0 \le \theta B \le \alpha < 1.$$

> **Remark.** Uspensky (1937, page 204) used this approach for the Bernstein inequality, before making the choice $\alpha = xB/(W + xB)$ to recover the traditional H_{Bern} form of the upper bound.

More generally, if we have a bound

\E@ restricted.subg <31>

$$\mathbb{P}e^{\theta(X-\mu)} \le e^{v\theta^2/2} \qquad \text{for } 0 \le \theta \le \rho,$$

with v and ρ positive constants and $\mu = \mathbb{P}X$, then for $x \ge 0$,

\E@ gauss.exp <32>

$$\begin{aligned} \mathbb{P}\{X \ge \mu + x\} &\le \min_{0\le\theta\le\rho} \exp\left(-x\theta + \tfrac{1}{2}v\theta^2\right) \\ &= \begin{cases} \exp\left(-\tfrac{1}{2}x^2/v\right) & \text{for } 0 \le x \le \rho v \\ \exp\left(-\rho x + \tfrac{1}{2}v\rho^2\right) < e^{-x\rho/2} & \text{for } x > \rho v \end{cases} . \end{aligned}$$

The minimum is achieved by $\theta = \min(\rho, x/v)$.

The proof in the next Section of a generalized version of the Hanson-Wright inequality—showing that a quadratic $\sum_{i,j} X_i A_{i,j} X_j$ in independent subgaussians has a subgaussian/subexponential tail— illustrates this approach.

2.8 The Hanson-Wright inequality for subgaussians

Basic::S:HW

The following is based on an elegant argument by Rudelson and Vershynin (2013) = R&V. To shorten the proof I assume that the A matrix has zeros down its diagonal, so that $\mathbb{P}\sum_{i,j} X_i A_{i,j} X_j = 0$ and I can skip the part of the argument (see Problem [17]) that deals with $\sum_i A_{i,i}(X_i^2 - \mathbb{P}X_i^2)$.

Basic::HW <33> **Theorem.** *Let $X_1, \dots, X_n$ be independent, centered subgaussian random variables with $\tau \ge \max_i \tau(X_i)$. Let A be an $n \times n$ matrix of real numbers with $A_{ii} = 0$ for each i. Then*

$$\mathbb{P}\{X'AX \ge t\} \le \exp\left(-\min\left(\frac{t^2}{64\tau^4 \|A\|_{\text{HS}}}, \frac{t}{8\sqrt{2}\tau^2 \|A\|_2}\right)\right) \qquad \text{for } t \ge 0,$$

where $\|A\|_{\text{HS}} := \sqrt{\text{trace}(A'A)} = \sqrt{\sum_{i,j} A_{i,j}^2}$ *and* $\|A\|_2 := \sup_{\|u\|_2 \le 1} \|Au\|_2$.

Remark. There are no assumptions of symmetry or positive definiteness on A. The proof also works with A replaced by $-A$, leading to a similar upper bound for $\mathbb{P}\{|X'AX| \ge t\}$. The Hilbert-Schmidt norm $\|A\|_{\text{HS}}$ is also called the Frobenius norm. The constants 64 and $8\sqrt{2}$ are not important.

Proof Without loss of generality, suppose $\tau = 1$, so that $\mathbb{P}e^{\theta X_i} \le e^{\theta^2/2}$ for all i and all real θ and $\mathbb{P}\exp(\alpha' X) \le \exp(\|\alpha\|_2^2/2)$ for each $\alpha \in \mathbb{R}^n$.

As explained at the end of Section 2.7, it suffices to show

$$\mathbb{P}e^{\theta X'AX} \le \exp(\tfrac{1}{2}v\theta^2) \qquad \text{for } 0 \le \theta \le \rho$$

where $v = 16\|A\|_{\text{HS}}^2$ and $\rho = \left(\sqrt{32}\,\|A\|_2\right)^{-1}$.

To avoid conditioning arguments, assume $\mathbb{P} = \otimes_{i\in[n]}P_i$ on $\mathcal{B}(\mathbb{R}^{[n]})$, where $[n] := \{1, 2, \dots, n\}$, and each X_i is a coordinate map, that is, $X_i(x) = x_i$ and $X(x) = x$. Write $\mathbb{G}$ for the $N(0, I_n)$ distribution, another probability measure on $\mathcal{B}(\mathbb{R}^{[n]})$. Of course $\mathbb{G} = \otimes_{i\in[n]}G_i$ with each G_i equal to $N(0,1)$. For each subset J of $[n]$ write $\mathbb{P}_J$ for $\otimes_{i\in J}P_i$ and $\mathbb{G}_J$ for $\otimes_{i\in J}G_i$ and, for generic vectors $w = (w_i : i \in [n])$ in $\mathbb{R}^{[n]}$, write w_J for $(w_i : i \in J) \in \mathbb{R}^J$.

The proof works by bounding $\mathbb{P}e^{\theta X'AX}$ by the $\mathbb{G}$ expected value of an analogous quadratic, using a decoupling argument that R&V attributed to Bourgain (1998).

For each $\delta = (\delta_i : i \in I) \in \{0,1\}^{[n]}$ define A_δ to be the $n \times n$ matrix with (i,j)th element $\delta_i(1-\delta_j)A_{i,j}$. Write D_δ for $\{i \in [n] : \delta_i = 1\}$ and N_δ for $[n]\backslash D_\delta$. Then $X'A_\delta X = X'_{D_\delta}EX_{N_\delta}$ where $E_\delta := A[D_\delta, N_\delta]$. If the rows and columns of A were permuted so that $i < i'$ for all $i \in D_\delta$ and $i' \in N_\delta$ then A_δ would look like

$$A_\delta = \begin{pmatrix} 0 & E_\delta \\ 0 & 0 \end{pmatrix} \qquad \text{where } E_\delta := A[D_\delta, N_\delta]\ .$$

Now let $\mathbb{Q}$ denote the uniform distribution on $\{0,1\}^{[n]}$. Under $\mathbb{Q}$ the δ_i's are independent Ber$(1/2)$ random variables, so that $\mathbb{Q}A_\delta = \frac{1}{4}A$. For each θ, Jensen's inequality and Fubini give

$$\mathbb{P}\exp\left(\theta X'AX\right) = \mathbb{P}^x\exp(4\theta\mathbb{Q}^\delta X'A_\delta X) \le \mathbb{Q}^\delta\mathbb{P}^x\exp(4\theta X'A_\delta X).$$

Write $\mathcal{Q}_\delta$ for $X'A_\delta X$. It therefore suffices to show that

$$\mathbb{P}\exp(4\theta\mathcal{Q}_\delta) \le \exp(\tfrac{1}{2}v\theta^2) \qquad \text{for each } \delta \text{ and all } 0 \le \theta \le \rho.$$

From this point onwards the δ is held fixed. To simplify notation I drop most of the subscript δ's, writing X_D instead of X_{D_δ}, and so on. Of course A_δ cannot drop its subscript.

Under $\mathbb{P}_D$, with X_N held fixed, the quadratic $\mathcal{Q}_\delta := X'A_\delta X = X_D'EX_N$ is a linear combination of the independent subgaussians X_D. Thus

$$\mathbb{P}_D e^{4\theta\mathcal{Q}_\delta} \le \exp\left(\tfrac{1}{2}(4\theta)^2 \left\|EX_N\right\|_2^2\right) = \mathbb{G}_D \exp\left(4\theta\mathcal{Q}_\delta\right).$$

Similarly, $\mathbb{P}_N \exp\left(4\theta\mathcal{Q}_\delta\right) \le \mathbb{G}_N \exp\left(\theta\mathcal{Q}_\delta\right)$. Thus

$$\begin{aligned}\mathbb{P}\exp(4\theta\mathcal{Q}_\delta) &= \mathbb{P}_D\mathbb{P}_N \exp(4\theta\mathcal{Q}_\delta)\\ &\le \mathbb{P}_D\mathbb{G}_N \exp(4\theta\mathcal{Q}_\delta) = \mathbb{G}_N\mathbb{P}_D \exp(4\theta\mathcal{Q}_\delta)\\ &\le \mathbb{G}_N\mathbb{G}_D \exp(4\theta\mathcal{Q}_\delta) = \mathbb{G}_N e^{8\theta^2 X_N'E'EX_N}.\end{aligned}$$

The problem is now reduced to its gaussian analog, for which rotational symmetry of $\mathbb{G}_N$ allows further simplification by means of a diagonalization of a positive definite matrix, $E'E = L'\Lambda L$, where L is orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, with $\lambda_1 \ge \lambda_2 \ge \dots \lambda_n \ge 0$ are the eigenvalues of $E'E$.

\E@ G.bnd <34>

$$\begin{aligned}\mathbb{G}_N e^{8\theta^2 X_N'E'EX_N} &= \mathbb{G}_N \exp\left(8\theta^2 X_N'\Lambda X_N\right) = \prod\nolimits_{i\in N} G_i \exp(8\theta^2\lambda_i x_i^2)\\ &\le \exp\left(\sum\nolimits_{i\in N} 16\theta^2\lambda_i\right) \qquad \text{provided } 8\theta^2 \max_{i\in N}\lambda_i \le 1/4.\end{aligned}$$

The last inequality uses the fact that, if $Z \sim N(0,1)$,

$$\mathbb{P}e^{sZ^2} = (1-2s)^{-1/2} = \exp\left(-\tfrac{1}{2}\log(1-2s)\right) \le \exp(2s) \qquad \text{if } 0 \le s \le \tfrac{1}{4}.$$

Now we need some matrix facts to bring everything back to an expression involving A. For notational simplicity again assume that the rows D all precede the rows N, so that

$$A = \begin{pmatrix} B_1 & E \\ B_2 & B_3 \end{pmatrix} \qquad \text{AND} \qquad A'A = \begin{pmatrix} B_1'B_1 + B_2'B_2 & B_1'E + B_2'B_3 \\ E'B_1 + B_3'B_2 & E'E + B_3'B_3 \end{pmatrix}.$$

It follows that

$$\|A\|_{\text{HS}}^2 = \operatorname{trace}(A'A) \ge \operatorname{trace}(E'E) = \sum\nolimits_{i\in N} \lambda_i$$

and

$$\|A\|_2 = \sup_{\|v\|_2\le 1} \|Av\|_2 \ge \sup_{\|u\|_2\le 1} \|Eu\|_2 = \lambda_1.$$

Combine the various inequalities to conclude that

$$\mathbb{P}\exp\left(\theta X'AX\right) \le \exp\left(16\theta^2 \|A\|_{\text{HS}}^2\right) \qquad \text{provide } 0 \le \theta \le \left(\sqrt{32}\,\|A\|_2\right)^{-1}.$$

An appeal to inequality <32> completes the proof.

□

2.9 Problems

Basic::S:Problems

Basic::P:convex.interval

[1] Define $a = \sup\{\theta < 0 : M(\theta) = +\infty\}$ and $b = \inf\{\theta > 0 : M(\theta) = +\infty\}$ where $M(\theta) = \mathbb{P}e^{\theta X}$. For simplicity, suppose both a and b are finite and $a < b$.

(i) Show that the restriction of M to the interval $[a, b]$ is continuous. (For example, if $M(b) = +\infty$ you should show that $M(\theta) \to +\infty$ as θ increases to b.)

(ii) Suppose $\theta \in (a, b)$. Choose $\delta > 0$ so that $[\theta - 2\delta, \theta + 2\delta_0] \subset (a, b)$. For $|h| < \delta$, establish the domination bound

$$\begin{aligned} |e^{(\theta+h)X} - e^{\theta X}|/h &\le \left| \int_0^1 X e^{(\theta+sh)X} ds \right| \\ &\le \max\left(e^{(\theta+2\delta)X}, e^{\theta X}, e^{(\theta-2s\delta)X} \right)/\delta. \end{aligned}$$

Deduce via a Dominated Convergence argument that $M'(\theta) = \mathbb{P}(Xe^{\theta X})$.

(iii) Argue similarly to justify another differentiation inside the expectation, leading to the expression for $M''(\theta)$.

Basic::P:e.to.g

[2] Suppose $h(x) = e^{g(x)}$, where g is a twice differentiable real-valued function on $\mathbb{R}^+$.

(i) Show that h is convex on any interval J for which $(g'(x))^2 + g''(x) \ge 0$ for $x \in \text{int}(J)$.

Basic::P:Psia.ge1

[3] For a fixed $\alpha \ge 1$ define $\Psi_\alpha(x) = e^{x^\alpha} - 1$ for $x \ge 0$.

(i) Show that $\Psi_\alpha(x)$ is a Young function..

(ii) Show that

$$\begin{aligned} \Psi_\alpha(x)\Psi_\alpha(y) \le \exp(x^\alpha + y^\alpha) - 1 &\le \Psi_\alpha(x+y) \qquad \text{for all } x, y \ge 0 \\ &\le \Psi_\alpha(2^{1/a}xy) \qquad \text{for } x \wedge y \ge 1. \end{aligned}$$

(iii) Deduce from the inequality $\Psi_\alpha(x)\Psi_\alpha(y) \le \Psi_\alpha(x+y)$ that

$$\Psi_\alpha^{-1}(uv) \le \Psi_\alpha^{-1}(u) + \Psi_\alpha^{-1}(v) \qquad \text{for all } u, v \ge 0.$$

Basic::P:extend.to.Young

[4] Suppose f is a convex, nonnegative, increasing function on an interval $[a, \infty)$, with $a > 0$. Suppose also that there exists an $x_0 \in [a, \infty)$ for which

$f(x_0)/x_0 = \tau := \inf_{x \geq a} f(x)/x$. Show that $h(x) := \tau x\{0 \leq x < x_0\} + f(x)\{x \geq x_0\}$ is a Young function.

Basic::P:extend.range

[5] Suppose $f : \mathbb{R}^+ \to \mathbb{R}^+$ is strictly increasing with $f(0) = 0$ and

$$f(u_1 u_2) \leq c_0 \left(f(u_1) + f(u_2) \right) \qquad \text{for } \min(u_1, u_2) \geq v_0,$$

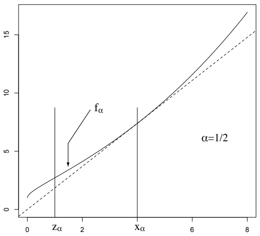
for some constants c_0 and $v_0 > 0$. For each v_1 in $(0, v_0)$ prove existence of a larger constant c_1 for which

$$f(u_1 u_2) \leq c_1 \left(f(u_1) + f(u_2) \right) \qquad \text{for } \min(u_1, u_2) \geq v_1.$$

Hint: For $u_i^* = \max(u_i, v_0)$ show that $f(u_i^*) \leq f(u_i) f(v_0)/f(v_1)$.

[6] For a fixed $\alpha \in (0,1)$ define $f_\alpha(x) = e^{x^\alpha}$ for $x \geq 0$.

Basic::P:Psia.lt1



(i) Show that $x \mapsto f_\alpha(x)$ is convex for $x \geq z_\alpha := (\alpha^{-1} - 1)^{1/\alpha}$ and concave for $0 \leq x \leq z_\alpha$.

(ii) Define $x_\alpha = (1/\alpha)^{1/\alpha}$ and $\tau_\alpha = f_\alpha(x_\alpha)/x_\alpha = (\alpha e)^{1/\alpha}$. Show that

$$\Psi_\alpha(x) := \tau x\{x < x_\alpha\} + f_\alpha(x)\{x \geq x_\alpha\}$$

is a Young function.

(iii) Show that there exists a constant K_α for which

$$\Psi_\alpha(x) \leq \exp(x^\alpha) \leq K_\alpha + \Psi_\alpha(x) \qquad \text{for all } x \geq 0.$$

(iv) Show that $\Psi_\alpha(x)\Psi_\alpha(y) \leq \Psi_\alpha(2^{1/\alpha} xy)$ for $x \wedge y \geq x_\alpha$.

(v) Deduce that there exists a constant C_α for which

$$C_\alpha \left(\Psi_\alpha^{-1}(u) + \Psi_\alpha^{-1}(v) \right) \geq \Psi_\alpha^{-1}(uv) \qquad \text{when } u \wedge v \geq 1.$$

Basic::P:slope.zero

[7] Suppose Ψ is a Young function with right-hand derivative $\tau > 0$ at 0. Define $\Phi(x) = \Psi(x) - \tau x$ for $x \geq 0$. If Φ is not identically zero then it is a Young function. For that case show that

$$\|f\|_\Psi \geq \|f\|_\Phi \geq \|f\|_\Psi / K \qquad \text{where } K = 1 + \tau \Phi^{-1}(1).$$

for each measurable real function f.

Basic::P:Stirling [8] For each k in $\mathbb{N}$ prove that $e^k \ge k^k/k! \ge 1$. Compare with the sharper estimate via Stirling's formula (Feller, 1968, Section 2.9),

$$\frac{k^k e^{-k}}{k!} = e^{-r_k}/\sqrt{2\pi k} \qquad \text{where } \frac{1}{12k} > r_k > \frac{1}{12k+1}.$$

Basic::P:interp [9] Suppose $n_0 = 1 \le n_1 < n_2 < \dots$ is an increasing sequence of positive integers with $\sup_{k\in\mathbb{N}} n_k/n_{k-1} = B < \infty$ and $f : [1,\infty) \to \mathbb{R}^+$ is an increasing function with $\sup_{t\ge1} f(Bt)/f(t) = A < \infty$. For each random variable X show that $\sup_{r\ge1} \|X\|_r/f(r) \le A \sup_{k\in\mathbb{N}} \|X\|_{n_k}/f(n_k)$.

Basic::P:moment.Psia [10] With Ψ_α as in Problem [2], prove the existence of positive constants c_α an C_α for which

$$c_\alpha \|X\|_{\Psi_\alpha} \le \sup_{r\ge1} \frac{\|X\|_r}{r^{1/\alpha}} \le C_\alpha \|X\|_{\Psi_\alpha}$$

for every random variable X.

Basic::P:subg.var [11] Theorem <16> suggests that $\|X\|_2$ might be comparable to $\beta(X)$ if X has a subgaussian distribution. It is easy to deduce from inequality <10> that $\|X\|_2 \le 2\beta(X)$. Show that there is no companion inequality in the other direction by considering the bounded, symmetric random variable X for which $\mathbb{P}\{X = \pm M\} = \delta$ and $\mathbb{P}\{X = 0\} = 1 - 2\delta$. If $2\delta = M$ show that $\mathbb{P}X^2 = 1$ but

$$\log\left(1 + \theta^2/2! + (\theta^4 M^2)/4!\right) \le \log \mathbb{P}e^{X\theta} \le \tau^2\theta^2/2 \qquad \text{for all real } \theta$$

would force $\tau^2 \ge 2\log\left(3/2 + M^2/24\right)$.

Basic::P:subg.suff [12] Show that a random variable X that has either of the following two properties is subgaussian.

(i) For some positive constants c_1, c_2, and c_3,

$$\mathbb{P}\{|X| \ge x\} \le c_1 \exp(-c_2 x^2) \qquad \text{for all } |x| \ge c_3.$$

(ii) For some positive constants c_1, c_2, and c_3,

$$\mathbb{P}e^{\theta X} \le c_1 \exp(c_2\theta^2) \qquad \text{for all } |\theta| \ge c_3$$

Basic::P:subg.limit1 [13] Suppose $\{X_n\}$ is a sequence of random variables for which

$$\mathbb{P}\exp(\theta X_n) \le \exp(\tau^2\theta^2/2) \qquad \text{for all } n \text{ and all real } \theta,$$

where τ is a fixed positive constant. If $X_n \to X$ almost surely, show that $\mathbb{P}e^{\theta X} \le \exp(\tau^2\theta^2/2)$ for all real θ.

Basic::P:lagrange [14] Suppose f is a sufficiently smooth real-valued function defined at least in a neighborhood of the origin of the real line. Define

$$G(x) = \frac{f(x) - f(0) - xf'(0)}{x^2/2} \qquad \text{if } x \neq 0, \text{ and } G(0) = f''(0).$$

Prove the following facts.

(i) $f(x) - f(0) - xf'(0) = x^2 \iint f''(xs)\{0 < s < t < 1\}\, ds\, dt$

(ii) If f is convex then G is nonnegative.

(iii) If f'' is an increasing function then so is $G(x)$.

(iv) If f'' is a convex function then so is G. Moreover $G(x) \ge f''(x/3)$. Hint: Jensen and $2\iint s\{0 < s < t < 1\}\, ds\, dt = 1/3$.

(v) For the special case where $f(x) = (1+x)\log(1+x) - x$ show that $G = \psi_{\text{Benn}}$ from <18> and $G(x) \ge (1 + x/3)^{-1}$.

Basic::P:Poisson [15] Recall the notation $F(x) = e^x - 1 - x$ for $x \in \mathbb{R}$ and

$\psi_{\text{Benn}}(y) = F^*(y)/(y^2/2) \qquad$ where $F^*(y) = (1+y)\log(1+y) - y$ for $y \ge -1$.

Suppose X has a Poisson(λ) distribution

(i) Show that $L(\theta) = \log \mathbb{P}^{\theta X} = \lambda\theta + \lambda F(\theta)$.

(ii) For $x \ge 0$, deduce that

$$\mathbb{P}\{X \ge \lambda + x\} \le \exp\left(-\lambda F^*(x/\lambda)\right) = \exp\left(-\frac{x^2}{2\lambda}\psi_{\text{Benn}}(x/\lambda)\right).$$

(iii) For $0 \le x \le \lambda$ show that

$$\begin{aligned}\mathbb{P}\{X \le \lambda - x\} &\le \inf_{\theta \ge 0} \exp\left(\lambda F(-\theta) - \lambda x\right) \\ &= \exp\left(-\frac{x^2}{2\lambda}\psi_{\text{Benn}}(-x/\lambda)\right) \le \exp\left(-\frac{x^2}{2\lambda}\right).\end{aligned}$$

Basic::P:BinPoisson [16] To what extent can the bounds from Problem [15] be recovered as limiting cases of the bounds derived via the Bennett inequality applied to the Bin$(n, \lambda/n)$ distribution?

Basic::P:diag.HW [17] Suppose $\infty > \gamma = \|X\|_{\Psi_2}$. Use Theorem <16> to deduce that $\mathbb{P}X^{2k} \le \left(\sqrt{2k}\gamma\right)^{2k} \le (2e\gamma^2)^k k!$. Use Bernstein's inequality from Theorem <23> to deduce a tail bound analogous to Theorem <33> for $\sum_i \alpha_i(X_i^2 - \mathbb{P}X_i^2)$ with independent subgaussians $X_1, \dots, X_n$ and constants α_i.

2.10 Notes

Basic::S:Notes

Anyone who has read Massart's Saint-Flour lectures (Massart, 2003) or Lugosi's lecture notes (Lugosi, 2003) will realize how much I have been influenced by those authors. Many of the ideas in those lectures reappear in the book by Boucheron, Lugosi, and Massart (2013).

Many authors seem to credit Chernoff (1952) with the moment generating trick in <2>, even though the idea is obviously much older: compare with the treatment of the Bernstein inequality by Uspensky (1937, pages 204–206)). Chernoff himself gave credit to Cramér for "extremely more powerful results" obtained under stronger assumptions.

The characterizations for subgaussians in Theorem <16> are essentially due to Kahane (1968, Exercise 6.10).

Hoeffding (1963) established many tail bounds for sums of independent random variables. He also commented that the results extend to martingales.

The Bennett inequality for independent summands was proved by Bennett (1962) by a more complicated method.

Section 2.6 on the Bernstein inequality is based on the exposition by Uspensky (1937, pages 204–206), with refinements from Bennett (1962) and Massart (2003, Section 1.2.3). See also Boucheron et al. (2013, Sections 2.4, 2.8) for a slightly more detailed exposition, including the idea of using the positive part to get a one-sided bound, which they credited to Emmanuel Rio. Van der Vaart and Wellner (1996, page 103) noted the connection between the moment assumption <24> and the behavior of $\mathbb{P}\Phi_1(\theta|X_i|)$.

Dudley (1999, Appendix H) used the name *Young-Orlicz modulus* for what I am calling a Young function. to such a function as an Orlicz modulus.

Every Young function can be written as $\Psi(t) = \int_0^t \psi(x)\,dx$, where ψ is the right-hand derivative of Ψ. The function ψ is increasing and right-continuous. For their N-functions, Krasnosel'skiĭ and Rutickiĭ (1961, Section 1.3) added the requirements that $\psi(0) = 0 < \psi(x)$ for $x > 0$ and $\psi(x) \to \infty$ as $x \to \infty$, which ensures that Ψ is strictly increasing with $\Psi(t)/t \to \infty$ as $t \to \infty$. Dudley called such a function an Orlicz modulus.

When $\psi(0) = 0$ there is a measure μ on $\mathcal{B}(\mathbb{R}^+)$ for which $\psi(b) = \mu(0, b]$ for all intervals $(0, b]$. In that case $\mathbb{P}\Psi(X) = \mu^s\mathbb{P}(X - s)^+$, a representation closely related to Lemma 1 of Panchenko (2003).

References

Bennett62jasa

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association 57*, 33–45.

BLM2013Concentration

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

bourgain1998random

Bourgain, J. (1998). Random points in isotropic convex sets. *MSRI Publications: Convex geometric analysis 34*, 53–58.

BoydVandenberghe2004

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Chernoff52AMS

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Annals of Mathematical Statistics 23*(4), 493–507.

Dudley99book

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

Feller1

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (third ed.), Volume 1. New York: Wiley.

Hoeffding:63

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Kahane68RSF

Kahane, J.-P. (1968). *Some Random Series of Functions*. Heath. Second edition: Cambridge University Press 1985.

KrasnoselskiiRutickii1961

Krasnosel'skiĭ, M. A. and Y. B. Rutickiĭ (1961). *Convex Functions and Orlicz Spaces*. Noordhoff. Translated from the first Russian edition by Leo F. Boron.

Lugosi2003ANU

Lugosi, G. (2003). Concentration-of-measure inequalities. Notes from the Summer School on Machine Learning, Australian National University. Available at http://www.econ.upf.es/˜lugosi/.

Massart03Flour

Massart, P. (2003). *Concentration Inequalities and Model Selection*, Volume 1896 of *Lecture Notes in Mathematics*. Springer Verlag. Lectures given at the 33rd Probability Summer School in Saint-Flour.

Panchenko2003AP

Panchenko, D. (2003). Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability 31*, 2068–2081.

RudelsonVershynin2013ECP

Rudelson, M. and R. Vershynin (2013). Hanson-Wright inequality and subgaussian concentration. *Electron. Commun. Probab 18*(82), 1–9.

Uspensky1937book

Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill.

vaartwellner96book

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

vershynin2010compressed

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. arXiv:1011.3027.