Printed: 11 September 2021

Chapter 7

Between subgaussian and subexponential

Subexp::Subexp

*Section 7.1 *briefly explains the meaning of the Chapter title.*

Section 7.2 *discusses the Bennett inequality for sums of independent random variables that are bounded above by a constant.*

Section 7.3 *shows how the Bennett inequality implies one form of the Bernstein inequality, then discusses extensions to unbounded summands.*

Section 7.4 *illustrates the ideas from the previous Section by deriving a subgaussian/subexponential tail bound for quadratic forms in independent subgaussian random variables.*

Section 7.5 *discusses the Bennett inequality for sums of martingale differences that are bounded above by a constant with high probability.*

*Section 7.6 *presents an extended application of a martingale version of the Bennett inequality to derive a version of the Kim-Vu inequality for polynomials in independent, bounded random variables.*

7.1 Introduction

Subexp::S:intro

This Chapter mostly discusses some tail bounds of the form

$$\mathbb{P}\{X - \mathbb{P}X \geq t\} \leq \exp\left(-B(t)\right) \qquad \text{for } t \geq 0,$$

where B is an increasing function that behaves like $c_1 t^2$ for $t \approx 0$ (subgaussian) and no worse than $c_2 t$ for large t (subexponential), for positive constants c_1 and c_2. Only in Section 7.6, which is devoted to one example of the so-called Kim-Vu inequalities, will $B(t)$ not fit this general description.

I have included this example because it is closely related to a martingale analog (Section 7.5) of the Bennett inequality.

It seems to me that some authors regard such an $e^{-B(t)}$ bound as universally inferior to a nice clean subgaussian tail bound. This certainly is not true, as I pointed out in Section 6.3 for X distributed $\text{Bin}(n,p)$. For that case, the Hoeffding inequality gives $\mathbb{P}\{X \ge np + t\} \le \exp(-2t^2/n)$, which is derived via an upper bound for the MGF of X, is always inferior to the tail bound derived from the actual MGF. For example, if $p = 3/4$, the tail bound from Section 3.7 gives $\mathbb{P}\{X \ge 3n/4 + t\} \le \exp\big(-(8/3)t^2/n\big)$.

In general, the Hoeffding inequality depends on the somewhat crude bound

$$\text{var}(W) \le (b-a)^2/4 \qquad \text{if } a \le W \le b = a + c,$$

which leads to the MGF bound

\E@ mgf.bdd.range

$$\log \mathbb{P}e^{\lambda W} \le 1 + \lambda \mathbb{P}W + \tfrac{1}{2}\lambda^2 c^2. \tag*{<1>}$$

Such an inequality can be too pessimistic if W has only a small probability of taking values near the endpoints of the interval $[a,b]$. It can sometimes be improved by introducing second moments explicitly into the upper bound.

7.2 Bennett's inequality for independent summands

Subexp::S:Bennett

The one-sided tail bound due to Bennett (1962) provides a particularly elegant illustration of the idea that it pays to use information about variances when bounding tail probabilitie using the MGF method.

The following argument simplifies Bennett's approach. The magic behind the improvement comes from the Fenchel-Legendre dual pair of convex functions defined in Section 2.2:

\E@ fbb

$$\begin{aligned} \mathbb{f}(\lambda) &= e^{\lambda} - 1 - \lambda = \tfrac{1}{2}\lambda^2 \mathbb{\Delta}(\lambda) \\ &\text{where } \mathbb{\Delta}(\lambda) := \int_0^1 2(1-s)e^{\lambda s}\,ds \qquad \text{for } \lambda \in \mathbb{R} \end{aligned} \tag*{<2>}$$

and

\E@ hbb

$$\begin{aligned} \mathbb{h}(t) &= (1+t)\log(1+t) - t = \tfrac{1}{2}t^2 \psi_{\text{Benn}}(t) \\ &\text{where } \psi_{\text{Benn}}(t) := \int_0^1 \frac{2(1-s)}{(1+ts)}\,ds \qquad \text{for } t \ge -1. \end{aligned} \tag*{<3>}$$

The function $\mathbb{h}$ could be extended to a convex function on the whole real line by defining $\mathbb{h}(x) = +\infty$ when $t < -1$. The function $\mathbb{\Delta}$ is continuous, convex, and strictly increasing. The function ψ_{Benn} is continuous, convex, and strictly decreasing on $[-1, \infty)$. The conjugacy relationship is given by

\E@ fh.conj <4>

$$\mathbb{f}^*(t) := \sup_{\lambda \in \mathbb{R}^+} (t\lambda - \mathbb{f}(\lambda)) = \mathbb{h}(t) \qquad \text{for all } t \in \mathbb{R}.$$

So much for the magic; now back to the MGF argument. The function $\mathbb{f}$ provides the bound that replaces <1>, via the rearrangement

$$e^{\lambda W} = 1 + \lambda W + \mathbb{f}(\lambda W) \leq 1 + \lambda W + \tfrac{1}{2}(\lambda W)^2 \mathbb{\Delta}(\lambda W).$$

If the random variable has $\mathbb{P}W^2 = v$ and $W \leq b$ for some constant b then, using the fact that $\mathbb{\Delta}$ is an increasing function and $\lambda W \leq \lambda b$ if $\lambda \geq 0$, we have

$$\mathbb{P}e^{\lambda W} \leq 1 + \lambda \mathbb{P}W + \tfrac{1}{2}\lambda^2 v \mathbb{\Delta}(\lambda b) \leq \exp\left(\lambda \mathbb{P}W + \tfrac{1}{2}\lambda^2 v \mathbb{\Delta}(\lambda b)\right) \qquad \text{for } \lambda \geq 0.$$

Divide by $e^{\lambda \mathbb{P}W}$ to deduce that

\E@ mgf.Bennett <5>

$$\mathbb{P}e^{\lambda(W - \mathbb{P}W)} \leq \exp\left(\tfrac{1}{2}\lambda^2 v \mathbb{\Delta}(\lambda b)\right) = \exp\left(v \mathbb{f}(\lambda b)/b^2\right) \qquad \text{for } \lambda \geq 0.$$

Remark. For the next Theorem it is important that summands X_i all share the same upper bound, $X_i \leq b$ for all i. If we were to subtract off the means the $\mathbb{P}X_i^2$'s would be replaced by true variances but the upper bounds might no longer be equal. We would have to replace b by $\max_i(b - \mathbb{P}X_i)$. This trade-off has always intrigued me. Some authors require zero means to get true variances; others just put up with second moments. I was always left wondering if there was some optimal compromise.

Subexp::Bennett.indep <*6*>

Theorem. *(Bennett's inequality for independent summands) Let $X_1, \dots, X_n$ be independent random variables with $X_i \leq b$, for a positive constant b, and $\mathbb{P}X_i^2 = v_i$ for each i. Then*

$$\mathbb{P}\left\{\sum_{i \leq n}(X_i - \mathbb{P}X_i) \geq t\right\} \leq \exp\left(-\frac{t^2}{2\mathcal{V}}\psi_{\text{Benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) \qquad \text{for } t \geq 0.$$

where $\mathcal{V}$ is any constant with $\mathcal{V} \geq V_n := \sum_{i \leq n} v_i$.

Remark. The case where $b = 0$, which gives a subgaussian tail, could be obtained as a limit as $b \downarrow 0$ if $X_i \leq 0$ for all i.

I introduced the $\mathcal{V}$ to emphasize that we do not need to know the exact value of V_n. It also makes the Theorem look more like the result for martingale differences that will be proved in Section 7.5.

Proof. Define $S = \sum_{i\le n}(X_i - \mathbb{P}X_i)$ so that multiple appeals to <5> give

$$\mathbb{P}e^{\lambda S} = \prod\nolimits_{i\le n} \mathbb{P}e^{\lambda(X_i - \mathbb{P}X_i)} \le \exp\left(V_n \mathbb{f}(\lambda b)/b^2\right).$$

Notice that the upper bound is an increasing function of V_n. Thus,

$$\mathbb{P}\{S \ge t\} \le \inf_{\lambda>0} e^{-\lambda t}\exp\left(\mathcal{V}\mathbb{f}(\lambda b)/b^2\right) = \exp\left(-\sup_{\lambda>0}\left(\lambda t - \mathcal{V}\mathbb{f}(\lambda b)/b^2\right)\right).$$

The expression in the exponent is $\mathcal{V}/b^2$ times

$$\sup_{\lambda>0}\left(\lambda b^2 t/\mathcal{V} - \mathbb{f}(\lambda b)\right) = \sup_{s>0}\left(sbt/\mathcal{V} - \mathbb{f}(s)\right) = \mathbb{h}(bt/\mathcal{V}).$$

□ The exponent in the upper bound equals $-(\mathcal{V}/b^2)\frac{1}{2}(bt/\mathcal{V})^2\psi_{\text{Benn}}(bt/\mathcal{V})$.

For a comparison between the Bennett and Hoeffding inequalities consider the case where $S = \sum_{i\le n} X_i$ where the X_i's are independent with $|X_i| \le b$ and $\mathbb{P}X_i = 0$ and $\text{var}(X_i) = \sigma^2$ for each i. Then we have $\text{var}(S) = n\sigma^2 = \mathcal{V}$ and

$$\mathbb{P}\{S \ge t\sqrt{\text{var}(S)}\} \le \begin{cases} \exp\left(-\frac{1}{2}t^2(\sigma/b)^2\right) & \text{using Hoeffding} \\ \exp\left(-\frac{1}{2}t^2\psi_{\text{Benn}}(bt/\sigma\sqrt{n}\right) & \text{using Bennett}\end{cases}.$$

If σ is smaller than b then the subgaussian Hoeffding bound is penalized by a small factor $(\sigma/b)^2$ in the exponent, whereas the exponent for the Bennett tail is close to $-t^2/2$ for moderate t and slowly degrades towards the nearly linear $-\frac{1}{2}t\sqrt{n}\log(2t/\sqrt{n})$ as t gets large. For example, if $\sigma/b = 1/2$ then the Bennett bound is superior when $\psi_{\text{Benn}}(bt/\sigma\sqrt{n}) \ge 1/4$, that is, when

$$2t/\sqrt{n} \le \psi_{\text{Benn}}^{-1}(1/4) \approx 16.13 \qquad \text{or} \qquad t \le 8.06\sqrt{n}.$$

For arguments that do not involve the extreme tails of the S distribution the Bennett bound provides better control.

Subexp::Binomial <7> **Example.** As a check on the sharpness of Theorem <6> consider the case where $X_i = \xi_i - p$, where $\xi_1, \dots, \xi_n$ are independent $\text{B}\textsc{er}(p)$'s and $S = \sum_i X_i$. Note that $\mathbb{P}X_i = 0$ and $\mathbb{P}X_i^2 = pq$ and $X_i \le q = 1-p$. For $t \ge 0$ the Theorem gives

\E@ Bin.Bennett <8>

$$\mathbb{P}\{S \ge t\} \le \exp\left(-\frac{t^2}{2npq}\psi_{\text{Benn}}\left(\frac{t}{np}\right)\right)$$

Using the fact that $\sum_{i\le n}\xi$ has a $\textsc{Bin}(n,p)$ distributions we also have a bound, from Section 3.7,

\E@ Bin.mgf <9>

$$\mathbb{P}\{S\ge t\}\le \exp\left(-\frac{t^2}{2npq}\left[q\psi_{\text{Benn}}\left(\frac{t}{np}\right)+p\psi_{\text{Benn}}\left(-\frac{t}{nq}\right)\right]\right)\qquad\text{for } t\ge 0$$

As <9> comes from a minimization involving the actual MGF it is better than <8>, which starts from an upper bound for that MGF. The superiority can also be seen directly from the fact that a convex combination of two distinct real numbers is greater than the smaller of the two, together with the inequality

$$\psi_{\text{Benn}}(-t/nq)>1>\psi_{\text{Benn}}(t/np)\qquad\text{if } t>0.$$

The Taylor expansion,

$$\psi_{\text{Benn}}(y)=1-y/3+O(y^2)\qquad\text{for } y \text{ near } 0,$$

sheds some light on the differences between <9> and <8>. For values of $t\ge 0$ with $\max(t/np,t/nq)$ not too large,

$$\begin{aligned}\psi_{\text{Benn}}(t/np)&=1-\frac{qt}{3npq}+O\left(\frac{t^2}{n^2}p^{-2}\right)\\ q\psi_{\text{Benn}}\left(\frac{t}{np}\right)+p\psi_{\text{Benn}}\left(-\frac{t}{nq}\right)&=1-\frac{(q-p)t}{3npq}+O\left(\frac{t^2}{n^2}(p^{-2}+q^{-2})\right).\end{aligned}$$

The general Bennett bound does not capture the skewness contribution from $\mathbb{P}S^3=npq(q-p)$. □

7.3 Bernstein's inequality

Subexp::S:Bernstein

In Section 2.2 you saw that $\psi_{\text{Benn}}(t)\ge(1+t/3)^{-1}$ for $t\ge -1$. Thus the Bennett inequality from Theorem <6> implies a weaker result: if $X_1,\dots,X_n$ are independent with $\mathbb{P}X_i^2=v_i$ and $X_i\le b=3B$ and $\mathcal{V}\ge\sum_i v_i$ then

\E@ Bernstein.indep0 <10>

$$\mathbb{P}\{\textstyle\sum_{i\le n}(X_i-\mathbb{P}X_i)\ge t\sqrt{\mathcal{V}}\}\le\exp\left(-\frac{t^2/2}{1+Bt/\sqrt{\mathcal{V}}}\right)\qquad\text{for } t\ge 0,$$

a version of ***Bernstein's inequality***. When $t\ll\sqrt{\mathcal{V}}/B$ the exponent is close to $-t^2/2$. When $Bt/\sqrt{\mathcal{V}}$ is large, the exponent behaves like $-t/(2B)$. That is, the inequality looks like a subgaussian bound for $Bt/\mathcal{V}$ small and like an exponential bound further out in the tail.

Remarks.

(i) I wrote the bounding constant as $3B$ rather than b in order to avoid a lot of powers of 3. It is also very convenient that, for each $\ell \in \mathbb{N}$,

$$\tfrac{1}{2}(\ell+2)! = 3 \times 4 \times \cdots \times (\ell+2) \geq 3^{\ell}.$$

This little trick also puts the moment bound in a form closer to the one introduced by Bernstein.

(ii) The Bernstein inequality, with attention focussed on the subgaussian region, played a key role in the early history of empirical process theory: Dudley (1978) used it to prove his functional central limit theorems.

The boundedness assumption behind <10> can be replaced by a weaker moment assumption, due to Bernstein, without affecting the tail bound.

ubexp::Bernstein.condition <11> **Definition.** *Say that a random variable X satisfies the Bernstein condition if there are positive constants v and B for which $\mathbb{P}|X|^k \leq \frac{1}{2}vB^{k-2}k!$ for $k \geq 2$. Equivalently, $\mathbb{P}|X/B|^k \leq \frac{1}{2}vk!/B^2$ for each $k \geq 2$.*

Remark. If X is a random variable for which $\mathbb{P}X^2 \leq v$ and $|X| \leq b = 3B$ then $\mathbb{P}|X|^k \leq \mathbb{P}X^2(3B)^{k-2} \leq vB^{k-2}3^{k-2} \leq \frac{1}{2}vB^{k-2}k!$ for $k \geq 2$.

The constant B plays the role of a scaling parameter. It is often cleaner to remove the main effect of B by working with the variable $Y = X/B$ and reparametrizing, $\alpha = v/B^2$. For example, for $0 \leq \lambda < 1$,

$$\mathbb{P}\left(e^{\lambda Y} - 1 - \lambda Y\right) = \sum\nolimits_{k\geq 2} \frac{\lambda^k \mathbb{P}Y^k}{k!} \leq \frac{\alpha}{2}\sum\nolimits_{k\geq 2} \frac{\lambda^k k!}{k!} = \frac{\alpha\lambda^2/2}{1-\lambda}.$$

which rearranges to

$$\mathbb{P}e^{\lambda Y} \leq 1 + \lambda\mathbb{P}Y + \frac{\alpha\lambda^2/2}{1-\lambda} \leq \exp\left(\lambda\mathbb{P}Y + \frac{\alpha\lambda^2}{2(1-\lambda)}\right) \qquad \text{for } 0 \leq \lambda < 1$$

and

$$\mathbb{P}e^{\lambda(Y-\mathbb{P}Y)} \leq \frac{\alpha\lambda^2}{2(1-\lambda)} \qquad \text{for } 0 \leq \lambda < 1.$$

If you have read Section 3.6 you will recognize the final inequality as the defining property for membership in SUBGAMMA(α), for which several tail bounds were derived. Tail bounds for X can be recovered by means of the trivial equality $\mathbb{P}\{X \geq t\} = \mathbb{P}\{Y \geq t/B\}$.

If $W_1, \dots, W_n$ are independent random variables with $W_i \in \textsc{subGamma}(\alpha_i)$ for each i then

$$\mathbb{P}e^{\lambda \sum_i W_i} = \prod\nolimits_i \mathbb{P}e^{\lambda W_i} \le \frac{\sum_i \alpha\lambda^2/2}{(1-\lambda)} \qquad \text{for } 0 \le \lambda < 1.$$

That is, $\sum_i W_i \in \textsc{subGamma}(\sum_i \alpha_i)$. Using this fact, together with the tail bounds from Section 3.6, we then get a result that contains Bernstein's version of his inequality.

Subexp::Bernstein.unbdd <12> **Theorem.** *Suppose $X_1, \dots, X_n$ are independent random variables with*

$$\mathbb{P}|X_i|^k \le \tfrac{1}{2} v_i B^{k-2} k! \qquad \text{for } k \ge 2, \text{ for each } i,$$

for some positive constants v_i and B, the same for each i. Then

$$\sum\nolimits_i (X_i - \mathbb{P}X_i)/B \in \textsc{subGamma}(\mathcal{V}/B^2) \qquad \text{where } \mathcal{V} = \sum\nolimits_i v_i$$

and

$$\mathbb{P}\left\{\sum\nolimits_{i\le n}(X_i - \mathbb{P}X_i) \ge t\right\} \le e^{-H_1} \le e^{-H_2} \qquad \text{for } t \ge 0.$$

where

$$H_1 = \frac{t^2}{\mathcal{V} + Bt + \sqrt{\mathcal{V}^2 + 2B\mathcal{V}t}} \ge H_2 = \frac{t^2}{2(\mathcal{V} + Bt)}.$$

Similar bounds also hold for the lower tails, because $\mathbb{P}|X_i|^k = \mathbb{P}|-X_i|^k$.

Remarks.

(i) The H_1 and H_2 refer to functions defined in Section 3.6.

(ii) I learned of the H_1 version from Bennett (1962, page 39). Bernstein only proved the H_2 form of the inequality. See Uspensky (1937, pages 204–206) for an English translation of Berstein's argument. Bennett cited a 1924 paper by Bernstein (which I have not seen) as the primary source. See the Notes to Chapter 3 for more about why the Bernstein result 'seemed to have escaped notice in the English-speaking world', an oversight perpetuated by modern authors who still credit the whole MGF method to Chernoff.

(iii) The derivation of the H_1 version of the inequality by Bennett (1962, page 37) contains the amazing assertion that the minimizing value of a parameter c (corresponding to my λ) is equal to $t/(\sigma F)$, where F is actually a function of c. Nevertheless, if one ignores this curious circularity it does lead to a legitimate choice for c, albeit not the minimizing value.

(iv) The Bernstein condition for individual X_i's is used only to establish the inequality $\sum_{i\le n}\mathbb{P}|X_i|^k \le \frac{1}{2}\mathcal{V}B^{k-2}k!$ for $k\ge 2$. See Problem [6] for an example where the added generality helps.

(v) The $\mathcal{V}$ in the Theorem plays the role of a variance. The H_2 bound from the Theorem can also be written as

$$\mathbb{P}\{\sum\nolimits_i(X_i-\mathbb{P}X_i)\ge t\sqrt{\mathcal{V}}\}\le \exp\left(-\frac{t^2/2}{1+Bt/\sqrt{\mathcal{V}}}\right).$$

For a given ϵ in $(0,1)$, the upper bound is $\le \exp(-(1-\epsilon)t^2/2)$ precisely when $0\le t\le \epsilon\sqrt{\mathcal{V}}/\left((1-\epsilon)B\right)$, which shows the importance of having $\sqrt{\mathcal{V}}$ much larger than B if the Bernstein bound is to provide a usefully large interval of subgaussian tail behavior.

If we are really only interested in a one-sided, upper bound then it is a tad superfluous to impose both upper and lower bounds on X_i via control of $\mathbb{P}|X_i|^k$. Boucheron, Lugosi, and Massart (2013, page 37) derived such a bound from an assumption on the moments of the positive parts $X_i^+ = \max(0,X_i)$. They attributed the main idea to Emmanuel Rio. They replaced the assumptions of Theorem <12> by

(i) the X_i's are independent,

(ii) $\sum_{i\le n}\mathbb{P}X_i^2\le\mathcal{V}$,

(iii) $\sum_{i\le n}\mathbb{P}\left(X_i^+\right)^k\le \mathcal{V}B^{k-2}k!/2$ for $k\ge 3$.

The proof is almost the same as the proof of the Theorem. One starts from the inequality

$$e^x\le 1+x+x^2/2+\sum\nolimits_{k\ge3}(x^+)^k/k! \qquad \text{for } x\in\mathbb{R},$$

which follows from $e^x-1-x=\frac{1}{2}x^2\Delta(x)$ with $\Delta(x)\le\Delta(0)=1$ if $x<0$ and is otherwise an equality.

Define $Y_i=X_i/B$ and $\alpha=\mathcal{V}/B^2$. For $\lambda>0$ we have

$$\mathbb{P}e^{\lambda Y_i}\le 1+\lambda\mathbb{P}Y_i+\lambda^2\mathbb{P}Y_i^2/2+\sum\nolimits_{k\ge3}\mathbb{P}\lambda^k\left(Y_i^+\right)^k/k!,$$

which, for $0\le\lambda<1$, implies

$$\mathbb{P}\exp\left(\lambda\sum\nolimits_i(Y_i-\mathbb{P}Y_i)\right)\le\exp\left(\lambda^2\alpha/2+\sum\nolimits_{k\ge3}\alpha\lambda^k/2\right)\le\exp\left(\frac{\alpha\lambda^2/2}{1-\lambda}\right).$$

That is, $\sum_i(Y_i-\mathbb{P}Y_i)\in \text{SUBGAMMA}(\alpha)$. And so on.

So far in this Section I have mostly been treating the Bernstein inequality as a tail bound for a SUBGAMMA distribution, following the approach taken by Boucheron et al. (2013, Sections 2.4, 2.7). In contrast, Vershynin (2018, Section 2.8) derived Bernstein-style bounds for sums of independent subexponential random variables, that is, random variables with finite $\mathcal{L}^{\Psi_1}$ norms. Earlier, van der Vaart and Wellner (1996, page 103) had noted the connection between the Bernstein condition <11> for a random variable X and the behavior of $\mathbb{P}\mathbb{f}(|X|/rB)$, for $r = 1, 2$. See Problems [2] and [4] for the inequalities relevant to the connections between $\mathcal{L}^{\Psi_1}$ (or the closely related $\mathcal{L}^{\mathbb{f}}$) and various Bernstein inequalities.

It took me some time to realize that the works of the eminent authors cited in the previous paragraph were actually closely related to each other. Contemplation of the following Example helped put me on the right track.

Subexp::double.exp <13> **Example.** Suppose X has a double exponential distribution, which has density $g(x) = \frac{1}{2}e^{-|x|}$ with respect to Lebesgue measure. These facts follow from easy Calculus exercises:

(i) $\mathbb{P}\{|X| \geq t\} = e^{-t}$ for $t \geq 0$;

(ii) $\mathbb{P}e^{\lambda X} = (1-\lambda^2)^{-1}$ for $|\lambda| < 1$;

(iii) $\mathbb{P}X^k = 0$ when k is odd and $\mathbb{P}|X|^k = k!$ for $k \in \mathbb{N}$;

(iv) $\pm X \in$ SUBGAMMA(2) because

$$\begin{aligned}\log M_X(\lambda) &= -\log(1-\lambda^2) = \lambda^2 + \lambda^4 + \dots \\ &\leq \left(\lambda^2 + \lambda^3 + \dots\right) = \frac{2\lambda^2/}{1-\lambda} \qquad \text{for } |\lambda| < 1.\end{aligned}$$

Property (iv) leads to the subGamma tail bound

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp\left(-\frac{t^2}{2(2+t)}\right) \qquad \text{for } t \geq 0,$$

which might seem to contradict property (i). However, a little algebra shows that $t^2/(4+2t) < t$ for all $t > 0$.

The moral of the story is: a tail bound that is close to subgaussian for some range of t need not be better than an exponential bound $2e^{-t}$. My brain had a hard time unlearning the false fact that subgaussian is better than subexponential. This Example reinforces the message that it is unwise to interpret Theorem <12> as a promise of useful subgaussian properties if $\sqrt{\mathcal{V}}/B$ is not large. Also, the fact that $t^2 \leq t$ for $0 \leq t \leq 1$ should be another reminder that subgaussianity is only interesting for larger values of t. □

7.4 The Hanson-Wright inequality for subgaussians

Subexp::S:HW

This Section is based on an elegant argument by Rudelson and Vershynin (2013) supplemented by Vershynin (2018, Chapter 6). Both these sources discussed the history of the Hanson-Wright inequality

The problem is to prove concentration of a quadratic $X'AX$ around its expected value, where $X = (X_1, \dots, X_n)$ is a vector of independent, subgaussians and A is any $n \times n$ real matrix. To simplify the argument I assume that $\max_i \tau(X_i) \leq 1$ and that the A matrix has zeros down its diagonal, which ensures $\mathbb{P} \sum_{i,j} X_i A[i,j] X_j = 0$. For the part of the argument that would deal with $\sum_i A[i,i](X_i^2 - \mathbb{P}X_i^2)$ if $\operatorname{diag}(A)$ were nonzero see Problem [6], which shows there are universal constants c_1 and c_2 for which

$$\mathbb{P}\{\sum\nolimits_i a_i(X_i^2 - \mathbb{P}X_i^2) \geq t\} \leq \exp\left(\frac{-t^2}{c_1|a|_2^2 + c_2|a|_\infty t}\right) \qquad \text{for } t \geq 0,$$

for each $a \in \mathbb{R}^n$, again under the assumption that $\max_i \tau(X_i) \leq 1$.

Subexp::HW <14> **Theorem.** *Let $X_1, \dots, X_n$ be independent subgaussian random variables with $\tau(X_i) \leq 1$ for each i and let A be an $n \times n$ matrix of real numbers with $A[i,i] = 0$ for each i. Then, for a universal constant C,*

$$\mathbb{P}\{X'AX \geq t\} \leq \exp\left(-C \min\left(t^2/\|A\|_{\mathrm{F}}^2, t/\|A\|_2\right)\right) \qquad \text{for } t \geq 0,$$

□ *where $\|A\|_{\mathrm{F}}^2 := \operatorname{trace}(A'A) = \sum_{i,j} A_{i,j}^2$ and $\|A\|_2 := \sup_{\|u\|_2 \leq 1} \|Au\|_2$.*

Remarks.

(i) 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 $\mathbb{P}\{|X'AX| \geq t\}$. To get the analogous result for subgaussians with different scale factors $\tau(X_i) = \tau_i$, replace A by TAT where $T = \operatorname{diag}(\tau_1, \dots, \tau_n)$. The constants C is not particularly important.

(ii) Both Rudelson and Vershynin (2013) and Vershynin (2018) wrote the upper bound as

$$\exp\left(-C \min\left(t^2/\|A\|_{\mathrm{F}}^2, t/\|A\|_2\right)\right).$$

The asserted inequality could be put in a form closer to the inequalities in Theorem <12> by using the fact that

$$\min\left(t^2/\mathcal{V}, t/B\right) \geq t^2/(\mathcal{V} + tB) \geq \tfrac{1}{2} \min\left(t^2/\mathcal{V}, t/B\right).$$

(iii) The quantity $\|A\|_{\mathrm{F}}$ is called the Frobenius (or Hilbert-Schmidt) norm of the matrix and $\|A\|_2$ is the norm of A as a linear map from $\mathbb{R}^n$ (under its usual Euclidean norm $\|u\| = \sqrt{\sum_{i\le n} u_i^2}$) back into itself. In general, if G is a $d\times m$ matrix of rank $k \le \min(d,m)$ then there exist orthonormal bases $\{v_1,\dots,v_m\}$ for $\mathbb{R}^m$ and $\{u_1,\dots,u_d\}$ for $\mathbb{R}^d$ and singular values $\sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_k > 0$ for which $G = \sum_{j=1}^k \sigma_j u_j v_j'$. The representation implies that $Gv_j = \sigma_j u_j$ for $j \le k$ and $Gv_j = 0$ for $j > k$. It is also easy to deduce that $\|G\|_{\mathrm{F}} = \sqrt{\sum_j \sigma_j^2}$ and $\|G\|_2 = \|Gv_1\|_2 = \sigma_1$. See Trefethen and Bau (1997, Lectures 4, 5) if you are not familiar with the singular value decomposition (SVD).

The following Example provides a preview of the method of proof for Theorem <14>: bound a quadratic function of independent subgaussians by an analogous quadratic function of independent standard normals. The rotational invariance of the standard multivariate normal then leads to a decomposition of the problem into a set of problems for one-dimensional normals.

Subexp::NolanPollard <15> **Example.** Suppose $W := \sum_{i,j\in[\![n]\!]} \mathfrak{s}_i\mathfrak{s}_j A[i,j]$, where A is an $n\times n$ symmetric matrix with vector of eigenvalues $\zeta = (\zeta_1,\dots,\zeta_n)$ and $\mathfrak{s}_1,\dots,\mathfrak{s}_n$ are independent Rademacher random variables ($\mathbb{P}\{\mathfrak{s}_i = \pm 1\} = 1/2$). The diagonal contributes only the constant $\sum_i A[i,i]$ to W, so we lose no generality in assuming $\mathrm{diag}(A) = 0$.

Here is a method, which I learned (in 1984, I think) from Gilles Pisier, for bounding $\mathbb{P}\{W \ge t\}$ when $t \ge 0$. Define $\mathfrak{s}_i = \mathrm{sgn}(g_i)$, where $g = (g_1,\dots,g_n) \sim N(0, I_n)$. Write c for $(2/\pi)^{1/2}$, the expected value of $|g_i|$. Then, by invoking Jensen's inequality conditionally on the $\mathfrak{s}_i$'s and using the fact that $g_i = \mathfrak{s}_i|g_i|$, we get

$$\begin{aligned}\mathbb{P}e^{\lambda W} &= \mathbb{P}\exp\left(\lambda \sum\nolimits_{i,j\in[\![n]\!]} \mathfrak{s}_i\mathfrak{s}_j \mathbb{P}|g_i|\,\mathbb{P}|g_j| A[i,j]/c^2\right)\\ &\le \mathbb{P}\exp\left(\lambda\sum\nolimits_{i,j} \mathfrak{s}_i\mathfrak{s}_j |g_i|\,|g_j| A[i,j](\pi/2)\right)\\ &= \mathbb{P}\exp\left(\tfrac{1}{2}\lambda\pi \sum\nolimits_{i,j} g_i g_j A[i,j]\right).\end{aligned}$$

By symmetry of A there exists an orthogonal matrix L for which $A =$

$L'\text{diag}(\zeta)L$. Most conveniently we also have

$$\begin{aligned} \textstyle\sum_j \zeta_j &= \text{trace}(\text{diag}(\zeta)) = \text{trace}(A) = 0, \\ |\zeta|_2^2 = \textstyle\sum_j \zeta_j^2 &= \text{trace}(\text{diag}(\zeta)^2) = \text{trace}(A^2) = \|A\|_{\text{F}}^2, \\ |\zeta|_\infty = \max_j |\zeta_j| &= \|\text{diag}(\zeta)\|_2 = \|A\|_2 . \end{aligned}$$

Thus $g'Ag = h'\text{diag}(\zeta)h = \sum_j \zeta_j h_j^2 = \sum_j \zeta_j(h_j^2 - 1)$, where $h = Lg \sim N(0, I_n)$ and

$$\mathbb{P}e^{\lambda W} \le \mathbb{P}\exp(\lambda Y) \qquad \text{where } Y = \tfrac{1}{2}\pi \textstyle\sum_j \zeta_j h_j^2.$$

It follows that

$$\mathbb{P}\{W \ge t\} \le \inf_{\lambda \ge 0} e^{-\lambda t}\mathbb{P}e^{\lambda W} \le \inf_{\lambda \ge 0} e^{-\lambda t}\mathbb{P}e^{\lambda Y}.$$

If you read Section 3.6 you already know the value for the second infimum

START HERE

□

The main trick in the proof of the Theorem is a decoupling argument, which Rudelson and Vershynin (2013) attributed to Bourgain (1999), that lets us replace $X'AX$ by a simpler quadratic form. The idea is a more refined and general version of the idea behind Pisier's argument. The key matrix facts are contained in the next Lemma, which use the properties of singular value decompositions given in Remark (iii) following the statement of the Theorem

Subexp::submatrix *<16>* **Lemma.** *Suppose A can be written in block form as*

$$A = \begin{bmatrix} B_1 & G \\ B_2 & B_3 \end{bmatrix}$$

where G is a $d \times m$ matrix with singular value decomposition $G = \sum_{i \le k} \theta_i u_i v_i'$ with $\theta_1 \ge \cdots \ge \theta_k > 0$. Then

(i) $\|A\|_{\text{F}} \ge \|G\|_{\text{F}} = \sqrt{\sum_{i \le k} \theta_i^2}$,

□ *(ii)* $\|A\|_2 \ge \|G\|_2 = \|Gv_1\|_2 = \theta_1$.

Proof. For (i) note that

$$\|A\|_{\mathrm{F}}^2 = \sum\nolimits_{i,j\in[\![n]\!]} A[i,j]^2 \ge \sum\nolimits_{i\in[\![d]\!],j\in[\![m]\!]} G[i,j]^2 = \|G\|_{\mathrm{F}}^2 = \sum\nolimits_{i\le k} \theta_i^2.$$

For (ii), remember the SVD fact that $\|G\|_2 = \|Gv_1\|_2 = \theta_1$. Define $\widetilde{v}_1 = [0, v_1]$, a unit column vector in $\mathbb{R}^n$. Then

$$\begin{aligned}\|A\|_2 \ge \|A\widetilde{v}_1\|_2 &= \left\| \begin{bmatrix} Gv_1 \\ B_3v_1 \end{bmatrix} \right\|_2 \\ &= \sqrt{\|Gv_1\|_2^2 + \|B_3v_1\|_2^2} \ge \|Gv_1\|_2 = \|G\|_2 \,.\end{aligned}$$

□

Finally we come to the proof of the main result.

Proof (of Theorem <14>). 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 γ for the $N(0, I_n) = N(0,1)^{\otimes n}$ distribution on $\mathbb{R}^n$, and $\mathbb{Q}$ for the uniform distribution on $\{0,1\}^n$. Under $\mathbb{Q}$, the coordinates of the generic point $\delta = (\delta_1, \dots, \delta_n)$ are independent $\textsc{Ber}(1/2)$ random variables. If $I \subset [\![n]\!]$ define $\mathbb{P}_I = \otimes_{i\in I} P_i$ and $\gamma_I = \otimes_{i\in I} N(0,1)$. Notice that the distribution of X changes when $\mathbb{P}$ is replaced by γ as the probability measure on $\mathbb{R}^n$.

For each $\delta \in \{0,1\}^n$ define a new $n \times n$ matrix A_δ by

$$A_\delta[i,j] = \delta_i(1-\delta_j)A[i,j].$$

Note that $\mathbb{Q}\delta_i(1-\delta_j) = 1/4$ if $i \ne j$ but $\delta_i(1-\delta_i) = 0$. The assumption about $\operatorname{diag}(A)$ was needed so that $\mathbb{Q}A_\delta = \frac{1}{4}A$.

By Jensen and Fubini,

\E@ decouple <17>
$$\mathbb{P}e^{\lambda X'AX} = \mathbb{P}e^{4\lambda X'(\mathbb{Q}A_\delta)X} \le \mathbb{Q}\mathbb{P}e^{4\lambda X'A_\delta X}.$$

For most of the remainder of the proof the δ is held fixed. Write D for $\{i \in [n] : \delta_i = 1\}$ and M for $[n]\backslash D$. Then $X'A_\delta X = X_D'GX_M$ where $G := A[D,M]$ and X_D is made up of the coordinates X_i for $i \in D$. If the rows and columns of A were permuted so that $i < i'$ for all $i \in D$ and $i' \in M$ then A_δ and A would look like

$$A_\delta = \begin{array}{c c} & \begin{array}{cc} D & M \end{array} \\ \begin{array}{c} D \\ M \end{array} & \begin{bmatrix} 0 & G \\ 0 & 0 \end{bmatrix} \end{array} \quad \text{AND} \quad A = \begin{array}{c c} & \begin{array}{cc} D & M \end{array} \\ \begin{array}{c} D \\ M \end{array} & \begin{bmatrix} B_1 & G \\ B_2 & B_3 \end{bmatrix} \end{array} \qquad \text{for some } B_i\text{'s.}$$

The quadratic $X'A_\delta X$ equals $X_D'GX_M = \sum_{i\in D,\, j\in M} X_i G[i,j]X_j$. Under both $\mathbb{P} = \mathbb{P}_D \otimes \mathbb{P}_M$ and $\gamma = \gamma_D \otimes \gamma_M$ the coordinate blocks X_D and X_M are independent. For fixed $\xi \in \mathbb{R}^D$ and $\eta \in \mathbb{R}^M$, independence and subgaussianity imply

$$\begin{aligned}\mathbb{P}_D e^{4\lambda X_D'\xi} &\le \exp\left(\tfrac12 (4\lambda)^2 \|\xi\|_2^2\right) = \gamma_D e^{4\lambda X_D'\xi},\\ \mathbb{P}_M e^{4\lambda \eta' X_M} &\le \exp\left(\tfrac12 (4\lambda)^2 \|\eta\|_2^2\right) = \gamma_M e^{4\lambda \eta' X_M}.\end{aligned}$$

Thus, by taking $\eta = G'X_D$ then $\xi = GX_M$ we get

$$\mathbb{P}_D\mathbb{P}_M e^{4\lambda X_D' G X_M} \le \gamma_M \mathbb{P}_D e^{4\lambda X_D' G X_M} \le \gamma_M e^{8\lambda^2 \|GX_M\|_2^2}.$$

If G has SVD $\sum_{j=1}^k \theta_j u_j v_j'$ with $\theta_1 \ge \dots \ge \theta_k > 0$ then

$$GX_M = \sum\nolimits_{j=1}^k \theta_j u_j v_j' X_M = \sum\nolimits_{j=1}^k \theta_j Z_j u_j \qquad \text{where } Z_j = v_j X_M \text{ for } j = 1,\dots,k$$

so that $\|GX_M\|_2^2 = \sum_{j\le k} \theta_j^2 Z_j^2$.

Under γ_M, the random variables $Z_1,\dots,Z_k$ are independent $N(0,1)$'s and their squares are independent χ_1^2's, with $\gamma_M \exp(\lambda Z_j^2) = (1-2\lambda)^{-1/2}$ for $2\lambda < 1$ and $\gamma_M \exp(8\lambda^2\theta_j^2 Z_j^2) = (1-16\lambda^2\theta_j^2)^{-1/2}$ for $16\lambda^2\theta_j^2 < 1$. It follows that

\E@ logMGF <18>

$$\log \mathbb{P}e^{4\lambda X' A_\delta X} \le -\tfrac12 \sum\nolimits_{j\le k} \log(1-16\lambda^2\theta_j^2) \qquad \text{if } 16\lambda^2\theta_j^2 < 1 \text{ for each } j.$$

The last inequality looks a little like the bound from Section 3.6 for the logMGF of a GAMMA(α) distributed random variable. Of course α is now replaced by $1/2$ and the $-\lambda - \log(1-\lambda)$ is replaced by $\log(1-16\lambda^2\theta_j^2)$. The good news is that we already have a λ^2 in the bound. The bad news is that the sneaky tricks that led to the SUBGAMMA no longer lead to a manageable upper bound for the MGF. However there is an alternative way to get a good upper bound.

Note that, for each constant c with $0 < c < 1$

$$-\log(1-t) = t + \frac{t^2}{2} + \frac{t^3}{3} + \dots \le \frac{t}{1-t} \le t/(1-c^2) \qquad \text{for } 0 \le t \le c^2.$$

With for $t = 16\lambda^2\theta_j^2 \le 16\lambda^2\theta_1^2$ for $1 \le j \le k$ and an appeal to Lemma <16>, this inequality bounds the right-hand side of <18> by

\E@ MGF.bnd <19>

$$8\lambda^2/(1-c^2)\sum\nolimits_j \theta_j^2 \le \|A\|_{\text{F}}^2 \qquad \text{provided } 4\lambda\,\|A\|_2 \le c.$$

Here I am using the facts that $\theta_1 = \|G\|_2 \le \|A\|_2$ is the largest of the θ_j's.

To summarize:

$$\mathbb{P}e^{4\lambda X' A_\delta X} \le \exp\left(D_1\lambda^2\right) \qquad \text{for } 0 \le \lambda \le D_2$$
$$\text{where } D_1 := 8\,\|A\|_{\text{F}}^2/(1-c^2) \text{ and } D_2 := c/(4\,\|A\|_2).$$

Notice that the upper bound does not depend on δ. The integration with respect to $\mathbb{Q}$ has no effect; the left-hand side can be replaced by $\mathbb{P}e^{\lambda X'AX}$. That leaves us with a simple minimization to bound the tail probabilities: for $t \ge 0$,

$$\log \mathbb{P}\{X'AX \ge t\} \le \min\{D_1\lambda^2 - \lambda t : 0 \le \lambda \le D_2\}.$$

An appeal to Lemma <**Subexp::constrained.min**> completes the proof. (Readers with spare time on their hands might want to optimize over the choice of c.) □

7.5 A Bennett-style inequality for martingales

Subexp::S:BennettMG

Theorem <6> has a very useful generalization where the assumption of independence of $X_1, \dots, X_n$ is replaced by the assumption that they are martingale differences for a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots \subseteq \mathcal{F}_n$.

> **Remark.** Note that Theorem <6> allowed the possibility that $\mathbb{P}X_i$ might not be zero, so that $\mathbb{P}X_i^2$ might be larger that $\text{var}(X_i)$. This time the martingale difference property requires the conditional epectation $\mathbb{P}_{i-1}X_i$ to be zero so that $\mathbb{P}_{i-1}X_i^2$ is the conditional variance.
>
> As before, I will omit the various 'almost sure' qualifications.

Subexp::Bennett.mda <20> **Theorem.** *(Bennett for martingales) Suppose $X_1, \dots,$ X_n are martingale differences for which there exist nonnegative, $\mathcal{F}_{j-1}$-measurable random variables M_j and v_j such that $X_j \le M_j$ and $\mathbb{P}_{j-1}X_j^2 \le v_j$ for each j. Define $S_n = X_1 + \dots + X_n$ and $V_n = v_1 + \dots + v_n$. For positive constants b and $\mathcal{V}$ define $G_{b,\mathcal{V}} := \{\max_{j\le n} M_j \le b, \quad V_n \le \mathcal{V}\}$. Then*

$$\mathbb{P}\left(\{S_n \ge t\} \cap G_{b,\mathcal{V}}\right) \le \exp\left(-\frac{t^2}{2\mathcal{V}}\psi_{\text{Benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) \qquad \text{for each } t \ge 0.$$

> **Remark.** The constants b and $\mathcal{V}$ can be allowed to depend on t,

Proof. Define $S_i = \sum_{j\le i} X_j$ and $V_i = \sum_{j\le i} v_j$ for $1 \le i \le n$. To allow a clean recursive argument also define $S_0 = V_0 = 0$.

As with many martingale proofs, the main idea is to multiply the increments X_j by predictable weights, which sets up an appeal to the analog of inequality <5> for conditional expectations. The weights come from the stopping time $\tau := \inf\{j \ge 1 : M_j > b \text{ or } V_j > \mathcal{V}\}$. As usual, the infimum of an empty set is taken to be $+\infty$; that is, $\{\tau = +\infty\} = G_{b,\mathcal{V}}$. More generally,

\E@ pred.st <21>
$$\{\tau > j\} = \{\max\nolimits_{i\le j} M_i \le b \text{ and } V_j \le \mathcal{V}\} \in \mathcal{F}_{j-1} \qquad \text{for } 1 \le j \le n.$$

> **Remark.** The sequences $\{v_j\}$ and $\{M_j\}$ are said to be ***predictable***, because the values v_j and M_j are both determined by what we learn up to step $j-1$. The event $\{\tau \le i\}$ is not just $\mathcal{F}_i$-measurable, as required by the definition of a stopping time; the random variable τ is a ***predictable stopping time***. Compare with Pollard (1984, pages 172–179). The corresponding concept in continuous time is exceedingly subtle. See Dellacherie and Meyer (1978, § IV.69) for example.

The $\mathcal{F}_{j-1}$-measurability of the event $\{\tau > j\}$ lets us stop the $\{S_i\}$ process before the event $G^c_{b,\mathcal{V}}$ occurs, without destroying the martingale property. Define new increments $Y_j = X_j\{\tau > j\}$ for $1 \le j \le n$. Then, with probability one,

\E@ Y.bnd <22>
$$Y_j \le \{\tau > j\}M_j \le b,$$

\E@ Y.mg <23>
$$\mathbb{P}_{j-1}Y_j = \mathbb{P}_{j-1}\{\tau > j\}X_j = \{\tau > j\}\mathbb{P}_{j-1}X_j = 0,$$

\E@ Y.condit.var <24>
$$w_j := \mathbb{P}_{j-1}Y_j^2 = \mathbb{P}_{j-1}\{\tau > j\}X_j^2 = \{\tau > j\}\mathbb{P}_{j-1}X_j^2 \le \{\tau > j\}v_j.$$

Inequalities <22> and <23> lead to the conditional analog of inequality <5>,

\E@ Y.mgf.Dele <25>
$$\mathbb{P}_{j-1}e^{\lambda Y_j} \le 1 + 0 + \tfrac{1}{2}\lambda^2 w_j \Delta(\lambda b) \le \exp\left(\tfrac{1}{2}\lambda^2 w_j \Delta(\lambda b)\right) \qquad \text{for } \lambda \ge 0,$$

and inequality <24> ensures that

\E@ Wn.bnd <26>
$$W_i := \sum\nolimits_{j\le i} w_j \le \sum\nolimits_{j\le n} \{\tau > j\}v_j \le \mathcal{V} \qquad \text{for all } i.$$

The partial sums $T_j = Y_1 + \dots + Y_j$, with $T_0 = 0$, also define a martingale, for which $T_n = S_n$ on the set $\{\tau = +\infty\} = G_{b,\mathcal{V}}$. Thus

$$\begin{aligned}\mathbb{P}\left(\{S_n \ge t\} \cap G_{b,\mathcal{V}}\right) &= \mathbb{P}\left(\{T_n \ge t\} \cap G_{b,\mathcal{V}}\right)\\ &\le \mathbb{P}\{T_n \ge t\} \le \inf\nolimits_{\lambda>0} \mathbb{P}\exp\left(\lambda T_n - \lambda t\right).\end{aligned}$$

We are back to the MGF method for a martingale $\{T_j : 0 \le j \le n\}$ with well behaved increments.

Define $D_j = \exp\left(\lambda T_j - \frac{1}{2}\lambda^2 W_j \Delta(\lambda b)\right)$ with $D_0 = 1$. For $j \ge 1$ we have

$$\begin{aligned}\mathbb{P}_{j-1} D_j &= \mathbb{P}_{j-1} D_{j-1} \exp\left(\lambda Y_j - \tfrac{1}{2}\lambda^2 w_j \Delta(\lambda b)\right)\\ &= D_{j-1} \exp\left(-\tfrac{1}{2}\lambda^2 w_j \Delta(\lambda b)\right) \mathbb{P}_{j-1} \exp\left(\lambda Y_j\right)\\ &\le D_{j-1}\end{aligned}$$

Take expected values then iterate to deduce that

$$\begin{aligned}\mathbb{P}\exp\left(\lambda T_n - \tfrac{1}{2}\lambda^2 \mathcal{V}\Delta(\lambda b)\right) &\le \mathbb{P} D_n \qquad \text{by inequality <26>}\\ &\le \mathbb{P} D_0 = 1.\end{aligned}$$

That is, $\mathbb{P}\exp(\lambda T_n) \le \exp\left(\frac{1}{2}\lambda^2 \mathcal{V}\Delta(\lambda b)\right)$, the same as the bound for $\mathbb{P}\exp(\lambda S)$ in the proof of Theorem <6>. The minimization of $\mathbb{P}\exp\left(\lambda T_n - \lambda t\right)$ follows the argument in that proof. □

The expanded version of inequality asserted by the previous Theorem implies

\E@ mg.upper <27>
$$\mathbb{P}\{S_n \ge t\} \le \exp\left(-\frac{t^2}{2\mathcal{V}}\psi_{\text{Benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) + \mathbb{P}\{\max\nolimits_{j\le n} M_j > b \text{ or } V_n > \mathcal{V}\}.$$

If, in addition to the assumptions of the Theorem we actually have two-sided control over the increments, $|X_j| \le M_j$, then we get a two-sided bound covering both tails,

\E@ mg.two-sided <28>
$$\mathbb{P}\{|S_n| \ge t\} \le 2\exp\left(-\frac{t^2}{2\mathcal{V}}\psi_{\text{Benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) + \mathbb{P}\{\max\nolimits_{j\le n} M_j > b \text{ or } V_n > \mathcal{V}\}.$$

See Section 7.6 for a surprising application of this two-sided bound, with the final probability bounded by

$$\sum\nolimits_{j\le n} \mathbb{P}\{M_j > b\} + \mathbb{P}\{\max\nolimits_j M_j \le b \text{ and } V_n > \mathcal{V}\}.$$

*7.6 Concentration of random polynomials

Subexp::S:KimVu

The main ideas for the following material come from the papers by Kim and Vu (2000) and Vu (2002). My Theorem <32> is a minor variation on Theorem 4.1 of Vu (2002). See the Notes for an explanation of the liberties I am taking in referring to the "Kim-Vu method".

Their method deals with random variables expressible as polynomials $f(\xi)$ in independent random variables $\xi = (\xi_1, \dots, \xi_n)$, each taking values in $[0,1]$. For example, we might have $f(x) = 3x_1^2x_2^5 + 2x_1^3x_2^3x_3x_4 + 7x_1^3x_4^5x_5x_9 + 6x_8^8x_9^9$ for $n = 9$. The degree of f is defined to be the largest of the degrees of its terms, $17 = \max(2+5, 3+3+1+1, 3+5+1+1, 8+9)$ for the example just given. At the cost of an extra factor of 2 we may assume that all the coefficients in the polynomial are nonnegative.

To simplify notation, I will again identify the ξ_i's with the coordinate maps on $[0,1]^n$ equipped with a product measure $\mathbb{P} = \otimes_{i\le n} P_i$.

In general, the polynomials will be written in the form

\E@ f.rep <29>
$$f(x) = \sum\nolimits_{a\in\mathbb{A}} w_a x^a \qquad \text{for } x \in [0,1]^n$$

where $\mathbb{A}$ is a finite subset of $\mathbb{Z}_+^n$, the set of all n-tuples of nonnegative integers, and for $a = (a_1, \dots, a_n)$

$$x^a = \prod\nolimits_{i\le n} x_i^{a_i} \qquad \text{with the convention that } x_i^0 = 1.$$

Note that the degree of f equals $\max_{a\in\mathbb{A}}\{w_a > 0\} \sum_{j\le n} a_j$.

The challenge is to find useful exponential tail bounds for $|f(\xi) - \mathbb{P}f(\xi)|$. The polynomials do satisfy a bounded difference conditions but, as explained in Section 6.5, the squared ranges are too crude as upper bounds for the variances. Instead, a recursive appeal to Theorem <20> will give a better result that involves two scaling quantities, $\mathcal{E}_0(f)$ and $\mathcal{E}_1(f)$. Their definitions, which involve some subtlety, greatly puzzled me when I first looked at the Kim-Vu and Vu papers. Only after I had extracted the main ideas in the proof did I understand the motivation for the following definition.

Subexp::ee.def *<30>* **Definition.** *For each nonempty subset H of the index set $[\![n]\!] := \{1, 2, \dots, n\}$ and $x \in [0,1]^n$ define*

$$D_H x^a = \begin{cases} 0 & \text{if there exists an } i \in H \text{ for which } a_i = 0 \\ x^{a'} & \text{where } a_i' = 0 \text{ for } i \in H \text{ and } a_i' = a_i \text{ otherwise} \end{cases}.$$

If $H = \{j\}$ is a singleton write D_j instead of $D_{\{j\}}$. For a polynomial f as in <29> define $D_H f = \sum_{a\in\mathbb{A}} w_a D_H x^a$ and $\mathcal{E}_1(f) = \max_{H\ne\emptyset} \mathbb{P} D_H f$ and $\mathcal{E}_0(f) = \max\left(\mathbb{P}f(\xi), \mathcal{E}_1(f)\right)$. □

> **Remark.** The effect of D_j is similar to the effect of the partial derivatives $\partial^r/\partial x_j^r$, except for the introduction of some extra constants. If $a_j = 0$ then both D_j and $\partial/\partial x_j$ would kill the whole $x\xi^a$ term. If $a_j = r > 0$ then both D_j and $\partial^r/\partial x_j^r$ would neatly remove the $x_j^{a_j}$ factor, at least from the ath term. I find it much cleaner to work with D_H rather than with partial derivatives, as Vu (2002) did.

Subexp::partial.g.eg <31> **Example.** For $g(x) = 3x_1^2x_2^5 + 2x_1^3x_2^3x_3x_4 + 7x_1^3x_4^5x_5x_9 + 6x_8^8x_9^9$,

$$\begin{aligned} D_Hg(x) &= 3 + 2x_3x_4 + 0 + 0 \qquad \text{if } H = \{1,2\} \\ D_Hg(x) &= 0 + 2x_1^3x_4 + 0 + 0 \qquad \text{if } H = \{2,3\} \end{aligned}$$

whereas $\partial^2\partial^5 g(x)/\partial x_1^2\partial x_2^5 = 720+0+0+0$ and $\partial^2\partial^3 g(x)/\partial x_1^2\partial x_2^3 = 360\xi_2^2 + 72x_1\xi_3x_3$. It is not possible to achieve the effect of $D_{\{1,2\}}$ (or even D_1) with partial derivatives. □

Subexp::KV-thm <32> **Theorem.** *For every kth degree polynomial f (with nonnegative coefficients),*

\E@ KV.ineq <33>
$$\mathbb{P}\{|f(\xi) - \mathbb{P}f(\xi)| \ge C_k t^k \sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)}\,\} \le C_{n,k}e^{-\mathbb{h}(t)} \qquad \text{for all } t > 0,$$

where the constants are defined recursively by $C_{n,k} = 2 + (n+1)C_{n,k-1}$ with $C_{n,1} = 2$ and $C_k = 2k(1 + C_{k-1})$ with $C_1 = 1$. □

> **Remark.** The original Kim-Vu and Vu proofs gave tail bounds involving factors $\exp(-t/4)$ for $t > 0$. For my version of their results the $t/4$ is replaced by
>
> $$\mathbb{h}(t) := (1+t)\log(1+t) - t = \tfrac{1}{2}t^2\psi_{\text{Benn}}(t),$$
>
> which fits better with appeals to Bennett inequalities. I do not regard the extra log factor as a particulary significant improvement.
>
> The recursion for $C_{n,k}$ implies $C_{n,k}(\lambda) \le 4(n+1)^{k-1}$. The appearance of such powers of n in the final bound is an inevitable consequence of the appeal to a union bound for $\max_j M_j$ in <36>.

The proof of the Theorem, which appears at the end of the Section, will involve a recursive argument (in the form of an induction on k), starting from the usual representation via a sum of martingale differences,

$$S_n = f(\xi) - \mathbb{P}f(\xi) = X_1 + \cdots + X_n \qquad \text{where } X_j = \mathbb{P}_j f(\xi) - \mathbb{P}_{j-1}f(\xi).$$

The $\mathbb{P}_j$ conditional expectation is obtaining by integrating over the variables $\xi_{j+1},\dots,\xi_n$. To avoid an excess of subscripted superscripts I will use a notation (borrowed from $\boldsymbol{R}$),

$$x^{a[I]} = \prod\nolimits_{i\in I} x_i^{a_i} \qquad \text{for } I \subset [\![n]\!],$$

with abbreviations $a[<j]$ for the special case where $I = \{i \in [\![n]\!] : i < j\}$, and so on. If $I = \emptyset$ interpret $\xi^{a[I]}$ to be identically equal to 1. Thus

$$\mathbb{P}_j f(\xi) = \sum\nolimits_{a\in\mathbb{A}} w_a \xi^{a[\le j]}\mathbb{P}\xi^{a[>j]}$$

so that

\E@ Xj.rep <34>

$$X_j = \sum\nolimits_{a\in\mathbb{A}} w_a \xi^{a[<j]} \left(\xi_j^{a_j} - \mathbb{P}\xi_j^{a_j}\right) \mathbb{P}\xi^{a[>j]}.$$

If $a_j = 0$ then $\xi_j^{a_j} = 1$ and jth factor becomes zero. If $a_j > 0$ then the jth factor takes values in $[-1, +1]$, which leads to the predictable upper bound

\E@ Mj.def <35>

$$|X_j| \le M_j := \sum\nolimits_{a\in\mathbb{A}} w_a\{a_j > 0\}\xi^{a[<j]}\mathbb{P}\xi^{a[>j]}.$$

Very conveniently, the bound $M_j = M_j(\xi_1, \dots, \xi_{j-1})$ is a polynomial of degree at most $k-1$ because it contains contributions from those $a \in \mathbb{A}$ for which $a_j > 0$ but with the $\xi_j^{a_j}$ factor excluded. Also note that $\mathbb{P}M_j = \mathbb{P}D_j f(\xi) \le \mathcal{E}_1(f)$, the first clue to the reason for the definition of $\mathcal{E}_1$.

The Bennett-like bound from Section 7.5 for positive constants b and $\mathcal{V}$, and $V_n := \sum_{j\le n}\mathbb{P}_{j-1}X_j^2$,

\E@ recursive <36>

$$\begin{aligned}\mathbb{P}\{|S_n| \ge t\} \le 2\exp\left(-\frac{t^2}{2\mathcal{V}}\psi_{\text{Benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) \\ + \sum\nolimits_{j\le n}\mathbb{P}\{M_j > b\} + \mathbb{P}\{\max_j M_j \le b \text{ and } V_n > \mathcal{V}\},\end{aligned}$$

now points the way to a proof by induction on k. Unfortunately, the squaring of the X_j leads to polynomial of degree greater than k for V_n. However, V_n can be bounded by a polynomial of degree at most $k-1$ by using the fact that

$$\mathbb{P}_{j-1}X_j^2 \le b\mathbb{P}_{j-1}|X_j| \qquad \text{on the set } \{M_j \le b\}.$$

Here I was sorely tempted use the upper bound M_j for $|X_j|$. For subtle reasons, such a simplification would create difficulties later in the proof. It is better to use the inequality

$$\begin{aligned}\mathbb{P}_{j-1}|X_j| &\le \sum\nolimits_{a\in\mathbb{A}} w_a\{a_j > 0\}\xi^{a[<j]}\mathbb{P}_{j-1}|\xi_j^{a_j} - \mathbb{P}\xi_j^{a_j}|\mathbb{P}\xi^{a[>j]} \\ &\le \sum\nolimits_{a\in\mathbb{A}} w_a\{a_j > 0\}\xi^{a[<j]}(2\mathbb{P}\xi_j^{a_j})\mathbb{P}\xi^{a[>j]},\end{aligned}$$

which leads to the bound

\E@ Un.def <37>

$$V_n \le 2bU_n \qquad \text{where } U_n := \sum\nolimits_{j\le n}\sum\nolimits_{a\in\mathbb{A}} w_a\{a_j > 0\}\xi^{a[<j]}\mathbb{P}\xi^{a[\ge j]}.$$

The random variable $U_n = U_n(\xi)$ is a polynomial of degree at most $k-1$ in ξ. The sum over $\mathbb{A}$ in the definition of U_n differs only slightly from the sum for M_j but the difference will be important in the proof of Lemma <42>, which will show that $\mathcal{E}_0(U_n) \le k\mathcal{E}_0(f)$.

The inequality <36> is now replaced by

\E@ recursiveU <38>

$$\mathbb{P}\{|S_n| \ge t\} \le 2\exp\left(-\frac{t^2}{2\mathcal{V}}\psi_{\text{Benn}}\left(\frac{bt}{\mathcal{V}}\right)\right) \\ + \sum\nolimits_{j\le n} \mathbb{P}\{M_j > b\} + \mathbb{P}\{2bU_n > \mathcal{V}\},$$

which is nicely poised for an inductive appeal to the Theorem for polynomials of degree at most $k-1$. At least conceptually, those appeals generate remainder terms involving polynomials of degree at most $k-2$. And so on. In essence we have to deal with a sequence of polynomials of decreasing degree obtained by eliminating one term at a time from each product ξ^a. The $\mathcal{E}_0$ and $\mathcal{E}_1$ are defined in way that controls all of these polynomials.

In my opinion, it is better to understand some simple cases before jumping into the proof of Theorem <32>. If you are too eager to bother with examples you could skip straight to Lemma <42>, which will explain the cleverness in Definition <30>, before presenting the (short) proof of the Theorem. However, you will then have to flip back to Example <39> for the $k=1$ starting point for the induction.

Subexp::k1 <39> **Example.** Calculate the $\mathcal{E}_j(f)$ for the simplest case where f has degree 1, that is, $f(\xi) = w_0 + \sum_{i\le n} w_i\xi_i$ with $0 \le w_i$ for all i.

The constant w_0 has no effect on $f(\xi) - \mathbb{P}f(\xi)$, so it might be cleaner to set w_0 equal to 0. However, repeated appeals to inequality <38> might conceivably lead to a polynomial of degree one for which there is a nonzero constant term w_0, so it is better to retain the w_0.

If $H \subset [\![n]\!]$ has size 2 or more then D_H kills all the terms in f. If $H = \{j\}$ then $D_H f = w_j$. It follows that $\mathcal{E}_1(f) = b := \max_{1\le j\le n} w_j$ and $\mathcal{E}_0(f) = \max(b, \mathbb{P}f(\xi))$, which leads to the scale factor

$$\gamma^2 := \mathcal{E}_0(f)\mathcal{E}_1(f) = \max\left(b^2, b\mathbb{P}f(\xi)\right).$$

Theorem <32> asserts

$$\mathbb{P}\{|f(\xi) - \mathbb{P}f(\xi)| \ge t\gamma\} \le 2e^{-\mathbb{h}(t)} \qquad \text{for } t > 0.$$

Compare with the two-sided version of the Bennett inequality for independent summands, $f(\xi) - w_0 = \sum_{j\ge1} w_j\xi_j$. First note that

$$\begin{aligned} V_n &= \sum\nolimits_{j\ge1} \mathbb{P}(w_j\xi_i)^2 \\ &\le \sum\nolimits_{j\ge1} bw_j\mathbb{P}\xi_j \qquad \text{because } w_j^2 \le bw_j \text{ and } \xi_j^2 \le \xi_j \\ &\le b\mathbb{P}f(\xi) \le \gamma^2 \end{aligned}$$

so that Theorem <6> with $\max_j |w_j\xi_j| \le b$ gives

$$\begin{aligned} \mathbb{P}\{|f(\xi) - \mathbb{P}f(\xi)| \ge t\gamma\} &\le 2\exp\left(-\frac{t^2\gamma^2}{2\gamma^2}\psi_{\text{Benn}}(t\gamma b/\gamma^2)\right) \\ &\le 2\exp\left(-\frac{t^2}{2}\psi_{\text{Benn}}(t)\right) = 2\exp(-\mathbb{h}(t)), \end{aligned}$$

the second inequality coming from the fact that $b/\gamma \le 1$ and ψ_{Benn} is monotone decreasing. That is, Theorem <32> for $k = 1$ follows directly from the Bennett inequality for independent summands. □

Subexp::ER.triangles2 <40> **Example.** For your convenience, I repeat the introduction to the Example from Section 6.5, which pointed out the deficiencies of the bounded-differences McDiarmid inequality in establishing a concentration bound for the number of triangles in an Erdős-Rényi random graph.

> Suppose the edges of a graph with vertex set $[\![n]\!] := \{i \in \mathbb{N} : i \le n\}$ whose edges are a subset of the set $\mathcal{E}$ of all $|\mathcal{E}| := \binom{n}{2}$ pairs of distinct vertices. Two edges are said to be adjacent if they share an endpoint. Three distinct edges form a triangle if together they contain only three vertices: that is, the edges are $\{i, j\}$, $\{j, k\}$, and $\{k, i\}$ for distinct vertices i, j, k. Write $\mathcal{T}$ for the set of all triangles, subsets of $\mathcal{E}$ of size 3 that involve exactly 3 vertices. The set $\mathcal{T}$ has size $\binom{n}{3}$.
>
> The Erdös-Rényi random graph $G_n(p)$ chooses its edges by means of a set of independent random variables $\{\xi_e : e \in \mathcal{E}\}$, with each ξ_e distributed $\text{Ber}(p)$. That is, $G_n(p)$ includes e when $\xi_e = 1$, an event with probability p.
>
> The number of triangles in $G_n(p)$ equals $f(\xi)$ where
>
> $$f(y) := \sum\nolimits_{\{e_1,e_2,e_3\}\in\mathcal{T}} y_{e_1}y_{e_2}y_{e_3} \qquad \text{for } y \in \{0,1\}^{\mathcal{E}}.$$
>
> The expected number of triangles is $\theta := \mathbb{P}f(\xi) = \binom{n}{3}p^3$.

Remember that the set $\mathcal{E}$ was enumerated in an arbitrary fashion as a sequence $(e_j : 1 \le j \le |\mathcal{E}|)$.

The Chen-Stein method had suggested that, under the extraneous assumption that np^2 is small, for each $\epsilon > 0$ there should be some $s > 0$ for which

\E@ triangle.conc <41>
$$\mathbb{P}\{|f(\xi) - \theta| > s\sqrt{\theta}\} < \epsilon.$$

Let us see how close we can get to such an inequality using Theorem <32>. Calculate.

(i) If $H = \{e_i\}$ for some edge e_i then D_H kills all triangles except for those that have e_i as one of their edges. As explained back in Section 6.5, there are $n-2$ such triangles, corresponding to the choices of the third vertex for the triangle. Thus $D_H f(x)$ is a sum of the form $\sum x_{e_j} x_{e_k}$ over all $n-2$ pairs of edges for which $\{e_i, e_j, e_k\} \in \mathcal{T}$. It follows that $\mathbb{P} D_H f = (n-2)p^2$, the magic quantity identified in the post mortem for the previous attempt at Erdös-Rényi.

(ii) If $H = \{e_i, e_j\}$ then only triangles containing both e_i and e_j can survive D_H. If these two edges do not share a common vertex there are no such triangles; otherwise, there is a single triangle containing both edges. It follows that $\mathbb{P} D_H f(\xi)$ is either 0 or p.

(iii) If $H = \{e_i, e_j, e_k\}$ then $D_H f$ is 1 if $H \in \mathcal{T}$ and 0 otherwise.

(iv) If $|H| \geq 4$ then $D_H f = 0$.

It follows that

$$\mathcal{E}_1(f) = \max\big((n-2)p^2, p, 1\big) \quad \text{AND} \quad \mathcal{E}_0(f) = \max\left(\theta, \mathcal{E}_1(f)\right).$$

To simplify the discussion, suppose $np = R$ for an R large enough that θ is the largest of the quantities

$$\theta \approx R^3/6, \quad (n-2)p^2 \approx R^2/n, \quad p = R/n, \quad 1.$$

In that case, $\mathcal{E}_0(f) = \theta$ and $\mathcal{E}_1(f) = (n-2)p^2$ and

$$\gamma := \sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)} \approx \text{const } \times \sqrt{\theta} \times \max(1, \sqrt{np^2})$$

and the bound from Theorem <32> simplifies to

$$\mathbb{P}\{|f(\xi) - \theta| > c_1 t^3 \sqrt{\theta} \max(1, \sqrt{np^2})\} \leq c_2 n^2 e^{-\mathbb{h}(t)}$$

for constants c_1 and c_2. This inequality is remarkably close to the desired <41> with $c_1 t^3$ playing the role of s. The extra n^2 on the right-hand side means that we need t large enough that $\mathbb{h}(t) \approx t\log(t)$ kills off a $2\log n$ term. The presence of the $\sqrt{np^2}$ factor suggests that the requirement that np^2 is small, which came from the Chen-Stein argument in Section 6.5, might not be as extraneous as I suspected.

Maybe one of other Theorems in the Vu (2002) paper could get us even closer to the desired <41>. I have not checked. □

Now I am ready to attack the proof of Theorem <32>. The main difficulty comes with the choice of the constants b and $\mathcal{V}$ in the first step, which involves some knowledge of the $\mathcal{E}_0$ and $\mathcal{E}_1$ constants for the M_j's and U_n.

Subexp::ee.transfer <42> **Lemma.** *For M_j as in <35> and U_n as in <37>,*

(i) $\mathcal{E}_0(M_j) \le \mathcal{E}_1(f)$ *for each* j,

(ii) $\mathcal{E}_0(U_n) \le k\mathcal{E}_0(f)$.

Proof. Start with the inequality for the expected value of $M_j(\xi_1, \dots, \xi_{j-1}) = \sum_{a\in\mathbb{A}} w_a\{a_j > 0\}\xi^{a[<j]}\mathbb{P}\xi^{a[>j]}$, which illustrates the general pattern of the arguments. By independence of the ξ_i's,

$$\mathbb{P}M_j = \sum\nolimits_{a\in\mathbb{A}} w_a\{a_j > 0\}\mathbb{P}\xi^{a[<j]}\xi^{a[>j]}.$$

When $a_j > 0$ the product $\xi^{a[<j]}\xi^{a[>j]}$ equals $D_j\xi^a$. When $a_j = 0$ the ath term makes no contribution to M_j. Thus $\mathbb{P}M_j \le \mathbb{P}D_jf(\xi) \le \mathcal{E}_1(f)$. The final inequality comes from the fact that $\{j\}$ is one of H's over which the max for $\mathcal{E}_0(f)$ is taken.

Similarly, for a nonempty subset H of $[[n]] = \{i \in \mathbb{N} : i \le n\}$

$$D_HM_j = \sum\nolimits_{a\in\mathbb{A}} w_a\{a_j > 0\}\left(D_H\xi^{a[<j]}\right)\mathbb{P}\xi^{a[>j]}.$$

If H is not a subset of $[[j-1]] = \{i \in \mathbb{N} : i < j\}$ then $D_H\xi^{a[<j]} = 0$ by definition. If $a_j > 0$ and $H \subset [[j-1]]$ then

$$\mathbb{P}\left(D_H\xi^{a[<j]}\right)\mathbb{P}\xi^{a[>j]} = \mathbb{P}\left(D_H\xi^{a[<j]}\right)\xi^{a[>j]} = \mathbb{P}D_{H\cup\{j\}}\xi^a.$$

Thus

$$\mathbb{P}D_HM_j \le \sum\nolimits_{a\in\mathbb{A}} w_a\{a_j > 0\}\mathbb{P}D_{H\cup\{j\}}\xi^a \le \mathbb{P}D_{H\cup\{j\}}f(\xi) \le \mathcal{E}_1(f).$$

The argument for $U_n := \sum_{j\le n}\sum_{a\in\mathbb{A}} w_a\{a_j > 0\}\xi^{a[<j]}\mathbb{P}\xi^{a[\ge j]}$ is similar. Again start with the expected value. Again by independence,

$$\mathbb{P}U_n = \sum_{j\le n}\sum_{a\in\mathbb{A}} w_a\{a_j > 0\}\mathbb{P}\xi^{a[<j]}\xi^{a[\ge j]} = \sum_{j\le n}\sum_{a\in\mathbb{A}} w_a\{a_j > 0\}\mathbb{P}\xi^a.$$

Notice how this time we have no D_j inside the sum. Interchange the order of summation and then note that

$$\sum\nolimits_{j\le n}\{a_j > 0\} \le \sum\nolimits_{j\le n} a_j \le k$$

to deduce that $\mathbb{P}U_n \le k\mathbb{P}f(\xi)$.

Similarly, for a nonempty subset H of $[\![n]\!]$,

$$\begin{aligned}
\mathbb{P}D_H U_n &= \sum\nolimits_{j\le n}\sum\nolimits_{a\in\mathbb{A}} w_a\{a_j>0\}\mathbb{P}\left(D_H\xi^{a[<j]}\right)\xi^{a[\ge j]}\\
&= \sum\nolimits_{j\le n}\sum\nolimits_{a\in\mathbb{A}} w_a\{a_j>0\}\mathbb{P}\left(D_H\xi^{a[<j]}\right)\xi^{a[\ge j]}\\
&\le \mathbb{P}\sum\nolimits_{j\le n}\sum\nolimits_{a\in\mathbb{A}} w_a\{a_j>0\}D_H\xi^{a}\\
&\le k\mathbb{P}\sum\nolimits_{a\in\mathbb{A}} w_a D_H\xi^{a} = k\mathbb{P}D_H f(\xi) \le \mathcal{E}_1(f).
\end{aligned}$$

For fixed j, the ath summand is again zero if either $a_j = 0$ or H is not a subset of $[\![j-1]\!]$. □

The Lemma sheds some light on the cunning choices made for Definition <30>. If we were to bound the contribution $\mathbb{P}\{M_j > b\}$ to <38> by the analogous inequality for S_n replaced by M_j then we would have to deal with some new M_k' polynomials derived from the martingale decomposition of the function M_j. Using the Lemma, we could bound the $\mathcal{E}_0(M_k')$ terms by $\mathcal{E}_1(M_j)$, which is smaller than $\mathcal{E}_1(f)$. For that idea to work we need $\mathcal{E}_1(f)$ to control polynomials obtained by knocking out two of the coordinates from ξ. That is, we need $\mathcal{E}_1(f)$ to control contributions from $D_H f$ for an H of size 2. And so on. The story is similar for U_n. Definition <30> ensures that we have control over all the polynomials obtained by knocking out more and more coordinates of ξ. When we get down to a single remaining coordinate then Example <39> takes over. That is why I like to think of the whole proof for the Theorem as a recursive argument moving down a tree with nodes represented by subsets of $[\![n]\!]$, with a polynomial attached to each node. The tree is rooted at $[\![n]\!]$, with leaves corresponding to the singleton subsets.

Proof (of Theorem <32>). The result for degree k equal to 1 was covered in Example <39>: for a polynomial g of degree 1,

$$\mathbb{P}\left\{|g(\xi)-\mathbb{P}g(\xi)| \ge t\sqrt{\mathcal{E}_0(g)\mathcal{E}_1(g)}\right\} \le 2e^{-\mathbb{h}(t)}.$$

The assertion of the Theorem is true with $C_1 = 1$ and $C_{n,1} = 2$.

In what follows we may always assume that $t \ge 1$ because the upper bound will always be bigger than the decreasing function $2e^{-\mathbb{h}(t)}$ and $2e^{-\mathbb{h}(1)} \approx 1.4$.

For the inductive step, assume the result holds for degree up to $k-1$. From <38> with t replaced by $r = C_k t^k\sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)}$ we have

\E@ recursive.r <43>
$$\begin{aligned}&\mathbb{P}\Big\{|f(\xi)-\mathbb{P}f(\xi)| \ge C_k t^k\sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)}\Big\}\\ &\quad\le 2\exp\left(-\frac{r^2}{2\mathcal{V}}\psi_{\text{Benn}}\left(\frac{br}{\mathcal{V}}\right)\right)+\sum\nolimits_{j\le n}\mathbb{P}\{M_j>b\}+\mathbb{P}\{2bU_n>\mathcal{V}\}.\end{aligned}$$

To reap an inductive benefit we need

\E@ b.Mj <44>
$$b \ge \mathbb{P}M_j + C_{k-1}t^{k-1}\sqrt{\mathcal{E}_0(M_j)\mathcal{E}_1(M_j)} \qquad \text{for each } j,$$

\E@ vv/b <45>
$$\mathcal{V}/(2b) \ge \mathbb{P}U_n + C_{k-1}t^{k-1}\sqrt{\mathcal{E}_0(U_n)\mathcal{E}_1(U_n)}.$$

Of course this is the point at which Lemma <42> comes to the rescue. Assuming $t\ge 1$, the right-hand side of <44> is bounded above by

$$\mathcal{E}_1(f) + C_{k-1}t^{k-1}\mathcal{E}_1(f) \le (1+C_{k-1})t^{k-1}\mathcal{E}_1(f)$$

and the right-hand side of <44> is bounded above by

$$k\mathcal{E}_0(f) + C_{k-1}t^{k-1}k\mathcal{E}_0(f) \le k(1+C_{k-1})t^{k-1}\mathcal{E}_0(f).$$

If we choose $C_k = 2k(1+C_{k-1})$ and

$$\begin{aligned}b &= C_k t^{k-1}\mathcal{E}_1(f)\\ \mathcal{V} &= 2bk(1+C_{k-1})t^{k-1}\mathcal{E}_0(f) = C_k^2 t^{2k-2}\mathcal{E}_0(f)\mathcal{E}_1(f)\end{aligned}$$

then

$$\begin{aligned}r^2/\mathcal{V} &= \frac{C_k^2 t^{2k}\mathcal{E}_0(f)\mathcal{E}_1(f)}{C_k^2 t^{2k-2}\mathcal{E}_0(f)\mathcal{E}_1(f)} = t^2,\\ br/\mathcal{V} &= \frac{C_k t^{k-1}\mathcal{E}_1(f)C_k t^k\sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)}}{C_k^2 t^{2k-2}\mathcal{E}_0(f)\mathcal{E}_1(f)} = t\sqrt{\mathcal{E}_1(f)/\mathcal{E}_0(f)} \le t.\end{aligned}$$

By virtue of the monotonicity of ψ_{Benn} and the inductive hypothesis, the right-hand side of <43> is then bounded above by

$$2\exp\left(-\tfrac{1}{2}t^2\psi_{\text{Benn}}(t)\right) + C_{n,k-1}(n+1)e^{-\mathbb{h}(t)} = \left(2+C_{n,k-1}(n+1)\right)e^{-\mathbb{h}(t)}.$$

The choice $C_{n,k} = 2 + C_{n,k-1}(n+1)$ then leaves the bound asserted by the Theorem for the probability on the left-hand side of <43> □

7.7 Problems

Subexp::S:Problems

Remember that $\mathbb{f}(t) := e^t - 1 - t = \Psi_1(t) - t$ *and* $\Psi_\alpha(x) = e^{x^\alpha} - 1$ *for* $\alpha \geq 1$ *are Young functions, with corresponding Orlicz norms* $\|\cdot\|_{\mathbb{f}}$ *and* $\|\cdot\|_{\Psi_\alpha}$. *Also remember Stirling's formula,*

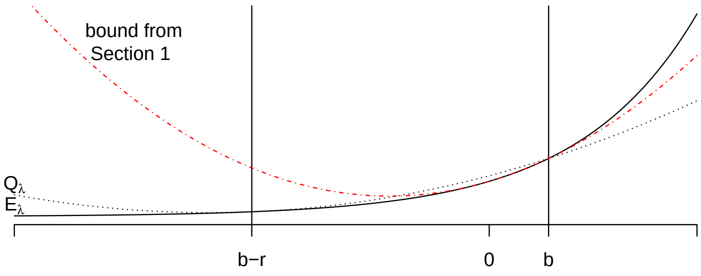
$$k! = \sqrt{2\pi k}\,(k/e)^k e^{r_k} \qquad \text{where } (1+12k)^{-1} < r_k < (12k)^{-1}$$

and its weak analog: $(k/e) \leq k! \leq k^k$.

Subexp::P:Bennett.original [1] Here is an outline of the key step in the original proof for Theorem <6> given by Bennett (1962, page 42). The main task was to bound the MGF for a random variable W with $\mathbb{P}W = 0$, $\mathrm{var}(W) = \sigma^2$, and $W \leq b$ with probability one. Bennett showed that

\E@ max.mgf <46>
$$\mathbb{P}e^{\lambda W} \leq \frac{b^2}{b^2+\sigma^2}e^{-\lambda\sigma^2/b} + \frac{\sigma^2}{b^2+\sigma^2}e^{\lambda b} \qquad \text{for } \lambda > 0.$$

With some brute-force Calculus this inequality can be used to derive an upper bound similar to my inequality <5>. See Hoeffding (1963, page 23) or Pollard (1984, page 194) for condensed accounts of the gruesome details.



Bennett's ingenious idea (for which he acknowledged the help of a referee) was to construct, for a fixed $\lambda > 0$, a quadratic $\mathcal{Q}_\lambda(x) = \alpha x^2 + \beta x + \gamma$ such that

\E@ Qlam <47>
$$\mathcal{Q}_\lambda(x) \geq E_\lambda(x) \qquad \text{for } x \leq b, \text{ with equality only at } x = -\sigma^2/b \text{ and } x = b.$$

Here and subsequently, $E_\lambda(x) = e^{\lambda x}$, considered as a function of x for a fixed λ. The following steps show you how to find the $\mathcal{Q}_\lambda$ that satisfies <47> and then how that inequality implies <46>.

(i) Let P denote the distribution that puts probability θ at $b - r$ and probability $1-\theta$ at b. Show that P has expected value 0 and variance σ^2 if

$$r = (b^2+\sigma^2)/b \quad \text{AND} \quad b - r = -\sigma^2/b \quad \text{AND} \quad \theta = b^2/(b^2+\sigma^2).$$

> **Remark.** Note that the right-hand side of <46> equals $Pe^{\lambda x}$. This P is extremal amongst those distributions on $(-\infty, b]$ with expected value 0 and variance σ^2.

(ii) Write $E_\lambda(x)$ for $e^{\lambda x}$ and x_0 for $b - r$, with r as in the definion of P. Find α, β, γ as functions of λ for which

$$\begin{aligned} E_\lambda(x_0) &= \mathcal{Q}_\lambda(x_0) = \alpha\left(b^2 - 2br + r^2\right) + \beta(b-r) + \gamma, \\ E'_\lambda(x_0) &= \mathcal{Q}'_\lambda(x_0) = 2\alpha(b-r) + \beta, \\ E_\lambda(b) &= \mathcal{Q}_\lambda(b) = \alpha b^2 + \beta b + \gamma. \end{aligned}$$

Hint: Define

$$\begin{aligned} e_2(y) &:= E_\lambda(x_0 + y) - E_\lambda(x_0) - yE'_\lambda(x_0) \\ q_2(y) &:= \mathcal{Q}_\lambda(x_0 + y) - \mathcal{Q}_\lambda(x_0) - y\mathcal{Q}'_\lambda(x_0). \end{aligned}$$

Use the fact that $e_2(r) = q_2(r) = \alpha r^2$, to show that

$$\alpha = e_2(r)/r^2, \qquad \beta = E'_\lambda(x_0) - 2\alpha x_0, \qquad \gamma = e^{\lambda b} - \alpha b^2 - \beta b.$$

(iii) Use the two integral representations (cf. Section 2.2)

$$\begin{aligned} e_2(y) &= y^2 \int_0^1 (1-s)E''_\lambda(x_0 + sy)\, ds, \\ q_2(y) &= y^2 \int_0^1 (1-s)2\alpha\, ds, \end{aligned}$$

to show that

$$\begin{aligned} \mathcal{Q}_\lambda(x_0 + y) - E_\lambda(x_0 + y) &= q_2(y) - e_2(y) \\ &= y^2\mathcal{R}(y) \qquad \text{where } \mathcal{R}(y) = \int_0^1 (1-s)\left(2\alpha - E''_\lambda(x_0 + sy)\right) ds. \end{aligned}$$

Use the fact that E''_λ is strictly increasing to deduce that the function $\mathcal{R}(y)$ is strictly decreasing.

(iv) From the equality $E_\lambda(b) = \mathcal{Q}_\lambda(b)$ show that $0 = r^2\mathcal{R}(r)$. Then deduce that $\mathcal{R}(y) > 0$ for $y < r$, implying $\mathcal{Q}_\lambda(x) \geq E_\lambda(x)$ for $x \leq b$, with equality only at $x = x_0$ and $x = b$.

(v) Finally, argue that

$$\begin{aligned} \mathbb{P}e^{\lambda W} \leq \mathbb{P}\mathcal{Q}_\lambda(W) &= \alpha\mathbb{P}W^2 + \beta\mathbb{P}W + \gamma \\ &= \alpha\sigma^2 + \beta.0 + \gamma \\ &= P\mathcal{Q}_\lambda(x) = Pe^{\lambda x}. \end{aligned}$$

Subexp::P:Psi1

[2] Facts about $\mathcal{L}^{\Psi_1}(\mathbb{P})$:

(i) If $X \in \mathcal{L}^{\Psi_1}(\mathbb{P})$ with $\gamma_1 := \|X\|_{\Psi_1} > 0$, show that

$$\mathbb{P}\{|X| \ge \gamma_1 t\} \le \mathbb{P}\exp\left((|X| - \gamma_1 t)/\gamma_1\right) \le 2\exp(-t) \qquad \text{for } t \ge 0.$$

Conversely, if $\mathbb{P}\{|X| \ge \beta_1 t\} \le 2\exp(-t)$ for all $t \ge 0$ and some $\beta_1 > 0$, show that

$$\mathbb{P}\Psi_1(|X|/3\beta_1) = \mathbb{P}_0^\infty\{|X| \ge 3\beta_1 t\}\, dt \le \int_0^\infty 2e^{-2t} dt = 1.$$

Deduce that $\gamma_1 \le 3\beta_1$.

(ii) Again, if $\infty > \gamma_1 := \|X\|_{\Psi_1} > 0$, deduce from

$$1 \ge \mathbb{P}\Psi_1(|X|/\gamma_1) = \sum\nolimits_{k\in\mathbb{N}} \mathbb{P}|X/\gamma_1|^k/k!$$

that $\|X\|_k \le \gamma_1 (k!)^{1/k} \le \gamma_1 k$ for each $k \in \mathbb{N}$. That is,

$$L_1(X) := \sup\nolimits_{k\in\mathbb{N}} \|X\|_k / k \le \|X\|_{\Psi_1}.$$

Conversely, suppose $L_1 = L_1(X) < \infty$. Show that

$$\mathbb{P}\Phi_1(|X|/c) \le \sum\nolimits_{k\in\mathbb{N}} \mathbb{P}(L_1 k/c)^k/k! \le \sum\nolimits_{k\in\mathbb{N}} (L_1 e/c)^k = 1 \qquad \text{if } c = 2L_1/e.$$

Deduce that $\|X\|_{\Psi_1}/(2e) \le L_1(X) \le \|X\|_{\Psi_1}$.

(iii) Suppose $\|X\|_k \le L_1 k$ for all $k \in \mathbb{N}$, with L_1 a finite constant. For all real λ with $|\lambda| \le 1/(L_1 e)$ show that

$$\mathbb{P}e^{\lambda X} \le 1 + \lambda\mathbb{P}X + \sum\nolimits_{k\ge 2} \frac{(|\lambda| L_1 k)^k}{k!} \le \exp\left(\lambda\mathbb{P}X + \frac{(|\lambda| L_1 e)^2}{1 - |\lambda| L_1 e}\right).$$

Deduce that $\pm(X - \mathbb{P}X)/(L_1 e) \in \textsc{subGamma}(1)$. Conversely, suppose Y is a random variable for which $\pm Y \in \textsc{subGamma}(1)$. Show that

$$1 + \sum_{k\in\mathbb{N}} \lambda^{2k}\mathbb{P}Y^{2k} = \tfrac{1}{2}\left(\mathbb{P}e^{\lambda Y} + \mathbb{P}e^{-\lambda Y}\right) \le \exp\left(\frac{\lambda^2/2}{1-\lambda}\right) \qquad \text{for } 0 \le \lambda < 1.$$

Deduce that $\|Y\|_{2k}^{2k} \le 2^k e^{1/4}(2k)!$ for $k \in \mathbb{N}$, implying $\sup_{p\ge 1} \|Y\|_p / p \le 5.2$.

Subexp::P:Psi1.Psi.alpha

[3] Deducing $\mathcal{L}^{\Psi_\alpha}$ facts from $\mathcal{L}^{\Psi_1}$ facts:

(i) For each $\alpha > 1$ show that $X \in \mathcal{L}^{\Psi_\alpha}(\mathbb{P})$ if and only if $|X|^\alpha \in \mathcal{L}^{\Psi_1}(\mathbb{P})$. Define $\gamma_\alpha := \|X\|_{\Psi_\alpha}$. For such an X show that $\gamma_\alpha^\alpha = \|\,|X|^\alpha\|_{\Psi_1}$.

(ii) Deduce from Problem [2] that

$$\mathbb{P}\{|X| \ge \gamma_\alpha t\} = \mathbb{P}\{|X|^\alpha \ge (\gamma_\alpha t)^\alpha\} \le 2\exp(-t^\alpha) \qquad \text{for } t \ge 0.$$

Conversely, if there is a constant $\beta > 0$ for which $\mathbb{P}\{|X| \ge \beta t\} \le 2\exp(-t^\alpha)$ for all $t \ge 0$ deduce that $\gamma_\alpha \le 3^{1/\alpha}\beta$.

::P:Bernstein.equivalences

[4] (cf. van der Vaart and Wellner, 1996, Section 2.2.2) Suppose X is a random variable for which there exists a positive constant B such that $\mathbb{P}|X|^k \le \frac{1}{2}vB^{k-2}k!$ for $k = 3, 4, \dots$, with equality at $k = 2$. (That is, X satisfies a little more than the Bernstein moment condition, <11> from Section 7.3.) Define $Y = X/B$ and $\alpha = v/B^2$, so that $\mathbb{P}|Y|^k \le \frac{1}{2}\alpha k!$ with equality at $k = 2$.

(i) From the inequality $\alpha = \mathbb{P}Y^2 \le \left(\mathbb{P}|Y|^4\right)^{1/2} \le (12\alpha)^{1/2}$ deduce that $\alpha \le 12$.

(ii) From the inequality

$$\mathbb{P}\mathbb{f}(|Y|/\lambda) = \mathbb{P}\left(e^{\lambda Y} - 1 - \lambda Y\right) \le \frac{\alpha\lambda^2/2}{1-\lambda} \qquad \text{for } 0 \le \lambda < 1$$

deduce that $\|X\|_{\mathbb{f}}/B = \|Y\|_{\mathbb{f}} \le \gamma := \left(1 + \sqrt{1+2\alpha}\right)/2 \le 3$. Hint: γ is the positive root of the equation $c^2 - c = \alpha/2$.

(iii) Suppose W is a random variable for which $\|W/B\|_{\mathbb{f}} \le 3$. Define $v = 9B^2$. Show that $\mathbb{P}|W|^k \le \frac{1}{2}vB^{k-2}k!$ for $k \ge 2$.

Subexp::P:fnorm.Bernstein

[5] Suppose $X_1, \dots, X_n$ are independent random variables in $\mathcal{L}^{\mathbb{f}}$, with $\gamma_i := \|X_i\|_{\mathbb{f}} > 0$. Define $B = \max_i \gamma_i$ and $\mathcal{V} = 2\sum_i \gamma_i^2$ and $W = \sum_i (X_i - \mathbb{P}X_i)$.

(i) From the inequality $\mathbb{P}\mathbb{f}(|X_i|/\gamma_i) \le 1$ deduce that $\mathbb{P}|X|^k \le \gamma_i^k k!$ for $k \ge 2$ so that

$$\mathbb{P}e^{\lambda(X_i - \mathbb{P}X_i)} \le \exp\left(\sum\nolimits_{k\ge2} \lambda^k \gamma_i^k\right) \le \exp\left(\frac{\lambda^2\gamma_i^2}{1 - B\lambda}\right) \qquad \text{for } 0 \le \lambda B < 1.$$

(ii) Deduce that $\pm W/B \in \textsc{subGamma}(\alpha)$ where $\alpha = \mathcal{V}/B^2$.

(iii) Deduce that

$$\mathbb{P}\{|W| \ge t\} \le 2\exp\left(-\frac{t^2}{4\sum_i \gamma_i^2 + 2t\max_i \gamma_i)}\right) \qquad \text{for } t \ge 0.$$

Subexp::P:diag.HW

[6] This problem handles the diagonal term that was omitted from Theorem <14>. Suppose $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ and $X_1, \dots, X_n$ are independent subgaussians

for which $\max \tau(X_i) \le 1$. Define $Y_i = X_i^2 - \mathbb{P}X_i^2$ and $W = \sum_{i\le n} a_i Y_i$. As usual, $|a|_2 := \sqrt{\sum_{i\le n} a_i^2}$ and $|a|_\infty = \max_{i\le n}|a_i|$.

(i) Use the inequality $\sup_{k\in\mathbb{N}} \|X_i\|_{2k}/\sqrt{k} := L(X_i) \le \sqrt{6}\,\tau(X_i)$ from Section 6.2 to show that

$$\mathbb{P}|Y_i|^k = \|Y_i\|_k^k \le \left(2\left\|X_i^2\right\|_k\right)^k = 2^k\mathbb{P}X_i^{2k} \le 2^k\left(\sqrt{6k}\right)^{2k} = (12k)^k.$$

(ii) For $0 \le 12e|a|_\infty\lambda < 1$ show that

$$\begin{aligned}\mathbb{P}e^{\lambda a_i Y_i} &= 1 + \lambda a_i \mathbb{P}Y_i + \sum\nolimits_{k\ge 2}\frac{|a_i|^k\lambda^k(12k)^k}{k!}\\ &\le 1 + \sum\nolimits_{k\ge 2}(12e|a_i|\lambda)^k \le \exp\left(\frac{(12e\lambda)^2 a_i^2}{1-12e|a|_\infty\lambda}\right),\end{aligned}$$

so that $\log \mathbb{P}e^{\lambda W/d} \le \alpha\lambda^2/2(1-\lambda)$ for $0 \le \lambda < 1$, where $d := 12e|a|_\infty$ and $\alpha := 2|a|_2^2/|a|_\infty^2$. That is, $W/d \in \textsc{subGamma}(\alpha)$, in the sense of Section 3.6.

(iii) Deduce that

$$\mathbb{P}\{S \ge t\} \le \exp\left(\frac{-t^2/d^2}{2(\alpha + t/d)}\right) = \exp\left(\frac{-t^2}{576|a|_2^2 + 12e|a|_\infty t}\right) \qquad \text{for } t \ge 0.$$

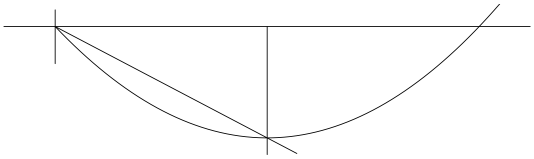
Subexp::P:constrained.min [7] If $\mathcal{Q}(\lambda) = D_1\lambda^2 - \lambda t$ for positive constants D_1, D_2, and t, show that

$$\min\nolimits_{0\le\lambda\le D_2}\mathcal{Q}(\lambda) \le \max\left(-t^2/(4D_1), -tD_2/2\right).$$

> **Remark.** Vershynin (2018) consistently wrote his Berstein-like tail bounds as maxima of two terms, because he reduced the MGF calculations to this constrained minimization problem rather than following the subGamma route.

(i) Show that the global minimum of $\mathcal{Q}(\lambda)$ is $-t^2/(4D_1)$, which is achieved at $\lambda_1 = t/(2D_1)$. If $D_2 \ge \lambda_1$ then the constrained minimum equals the global minimum, otherwise it equals $\mathcal{Q}(D_2)$.

(ii) Show that $\mathcal{Q}(\lambda) \le -t\lambda/2$ for $0 \le \lambda \le \lambda_1$. Deduce that $\mathcal{Q}(D_2) \le -tD_2/2$ if $D_2 < \lambda_1$.



7.8 Notes

Subexp::S:Notes

The Bennett inequality for independent summands comes from Bennett (1962), although with a different proof, which I outlined in Problem [1] because I suspect his ingenious argument might be useful for other constrained optimization problems.

Section 7.3 on the Bernstein inequality is based on the exposition by Uspensky (1937, pages 204–206), who followed the account in Bernstein's 1927 "Theory of Probability" (in Russian), which Elena Khusainova kindly translated for me. As already noted, my treatment was also heavily influenced by Bennett (1962), Boucheron, Lugosi, and Massart (2013, Sections 2.4, 2.7), Vershynin (2018, Section 2.8), and van der Vaart and Wellner (1996, page 103).

A weaker version of Theorem <14> was proved by Hanson and Wright (1971). They assumed symmetry of the matrix A and symmetry around zero for the distributions of the X_i's. They commented: "We would very much like to remove the restriction that the distributions of the X's be symmetric. Unfortunately, our proof depends heavily on this symmetry." Interestingly, that method involved a bounding of moment quantities for the X_i's by analogous moment quantities for the $N(0,1)$, leading to a bound on the MGF for the centered quadratic form by an analogous MGF for a quadratic form in standard normals. From that point on, their proof was similar to the proof of Theorem <14>.

Rudelson and Vershynin (2013) cited Bourgain (1999, page 55) for the clever decoupling idea with the δ_i's. Bourgain jumped straight to the decoupling bound, "we use a standard decoupling trick", without much explanation. Vershynin (2018, Chapter 6) (which I highly recommend) instead cited Bourgain and Tzafriri (1987, page 149), whose proof was prefaced by the comment:

> The proof of Theorem 1.6 requires the use of a variant of the so-called decoupling principle.
>
> This principle can be found in literature, mostly for symmetric matrices. For sake of completeness, we give here a proof of the version needed below.

A slightly cruder form of Example <15> was the key to an analysis (Nolan and Pollard, 1987) of U-processes, stochastic processes whose increments are U-statistics. We used a symmetrization trick (compare with Chapter 15) to control U-statistics by quadratic forms in independent Rademachers. Of course we did not use the subGamma approach. Instead we effectively used a constrained optimization analogous to the one in Problem [7].

The extension of Bennett's inequality to sums of martingale differences with increments that are bounded in absolute value by 1 is largely due to Freedman (1975). The relaxation to predictable upper bounds is implicit in Sections 3.3 and 3.4 of the paper of Vu (2002), although he did not express it via stopping times. He did not appeal to Bennett's inequality, but instead derived the necessary exponential bounds by direct arguments. Something similar to inequality <33> appeared in the Kim and Vu (2000) paper, but only for X_i's taking values in $\{0, 1\}$. Vu (2002) developed more elaborate versions of the inequality. Most of my discussion in Section 7.6 is a translation of Vu's paper into the framework of Bennett's inequality.

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.

Bourgain1998MSRI Bourgain, J. (1999). Random points in isotropic convex sets. *MSRI Publications: Convex geometric analysis 34*, 53–58. Available at https://www.cambridge.org/us/catalogue/catalogue.asp?isbn=9780521155649.

BourgainTzafriri1987 Bourgain, J. and L. Tzafriri (1987). Invertibility of "large" submatrices with applications to the geometry of banach spaces and harmonic analysis. *Israel journal of mathematics 57*(2), 137–224.

DellacherieMeyer78bookA Dellacherie, C. and P. A. Meyer (1978). *Probabilities and Potential*. Amsterdam: North-Holland. (First of three volumes).

Dudley78clt Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability 6*, 899–929.

Freedman1975AnnProb Freedman, D. A. (1975). On tail probabilities for martingales. *Annals of Probability 3*(1), 100–118.

HansonWright1971AMS Hanson, D. L. and F. T. Wright (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics 42*(3), pp. 1079–1083.

Hoeffding1963JASA Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

KimVu2000Combinatorica Kim, J. H. and V. H. Vu (2000). Concentration of multivariate polynomials and its applications. *Combinatorica 20*(3), 417–434.

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.

NolanPollard87Uproc1 Nolan, D. and D. Pollard (1987). U-processes: rates of convergence. *Annals of Statistics 15*, 780–799.

Pollard84book Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

RudelsonVershynin2013ECP Rudelson, M. and R. Vershynin (2013). Hanson-Wright inequality and subgaussian concentration. *Electron. Commun. Probab 18*(82), 1–9.

Trefethen1997numerical Trefethen, L. N. and D. Bau (1997). *Numerical linear algebra*, Volume 50. Siam.

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.

Vershynin2020HDP Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Vu2002RSA Vu, V. H. (2002). Concentration of non-Lipschitz functions and applications. *Random Structures and Algorithms 20*, 262–316.