Printed: 2 March 2024 at 14:33

Chapter 7

Subgaussian distributions

SECTION 7.1 *defines the subgaussian family of distributions, providing two useful examples.*
SECTION 7.2 *shows that the subgaussian property is equivalent to several other distributional properties.*
SECTION 7.3 *discusses the hoeffding inequality for sums of independent bounded random variables.*
SECTION 7.4 *shows how the hoeffding inequality can be extended to sums of martingale differences.*
SECTION 7.5 *presents an application of the martingale version of the hoeffding inequality, although the resulting bound falls far short of what is possible using other methods.*

7.1 Definition of subgaussian

Section 3.3 noted that every random variable X for which $\mathbb{P}e^{\lambda X} \le e^{\sigma^2\lambda^2/2}$, with $\sigma > 0$ and all $\lambda \in \mathbb{R}$, shares with the $N(0,\sigma^2)$ a two-sided exponential tail bound for $r \ge 0$,

<1> $$\begin{aligned} \mathbb{P}\{X \ge r\} &\le \inf_{\lambda>0} \mathbb{P}e^{\lambda(X-r)} = \exp(-r^2/2\sigma^2) \\ \mathbb{P}\{X \le -r\} &\le \inf_{\lambda\le 0} \mathbb{P}e^{\lambda(X+r)} = \exp(-r^2/2\sigma^2) \end{aligned}$$

As shown by Problems [1] and [2], the constant σ^2 might need to be strictly greater than $\operatorname{var}(X)$. To avoid any hint of the convention that σ^2 denotes a variance, I feel it is safer to replace σ by another symbol.

<2> **Definition.** *Say that a random variable X has a* ***subgaussian distribution*** *with scale factor $0 < \tau < \infty$, denoted by $X \in \text{SUBG}(\tau^2)$, if $\mathbb{P}e^{\lambda X} \le \exp(\tau^2\lambda^2/2)$ for all real λ. Write $\tau(X)$ for the smallest τ for which such a bound holds.*

Remark. Notice that $\text{SUBG}(\tau^2)$ denotes not a single distribution but rather a whole family of distributions, hence the $\in$ instead of $\sim$. This means that it makes no sense to speak of a maximum likelihood estimator (mle) for the parameter of $\text{SUBG}(\tau^2)$ based on independent

observations $X_1, \ldots, X_n$, although it does make sense to speak of the behavior of the mle, $\widehat{\tau}^2 = n^{-1}\sum_i (X_i - \overline{X})^2$, derived for the $N(0, \tau^2)$ model when the distributional assumption is relaxed to subgaussianity. In the same way, one could examine the behavior of the mle under other departures from normality, as the Robustniks like to do.

Some authors call the $\text{SUBG}(\tau^2)$ the centered subgaussian, allowing an extra parameter μ for which $\mathbb{P}e^{\lambda X} \le \exp(\lambda\mu + \lambda^2\tau^2/2)$ for all real λ when $X \in \text{SUBG}(\mu, \tau^2)$. In this case μ must equal $\mathbb{P}X$.

If X has a $\text{SUBG}(\beta^2)$ disribution, inequality <1> gives a pair of tail bounds,

<3> $$\mathbb{P}\{\pm X \ge r\} \le \exp\left(-\frac{r^2}{2\beta^2}\right) \qquad \text{for all } r \ge 0,$$

In fact (Theorem <6>) existence of such a tail bound for some $\beta > 0$ is equivalent to subgaussianity.

<4> **Example.** Suppose $Y = (Y_1, \ldots, Y_n)$ has a multivariate normal distribution. Define $M := \max_{i\le n} Y_i$ and $\sigma^2 := \max_{i\le n} \text{var}(Y_i)$. From Section 6.1 we have □ That is, $M - \mathbb{P}M \in \text{SUBG}(\sigma^2)$, 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 13). The next example captures the main idea.

<5> **Example.** Suppose $S = \sum_{j\le n} c_j \mathfrak{s}_j$, where the c_j's are constants and the $\mathfrak{s}_j$'s are independent random variables with $\mathbb{P}\{\mathfrak{s}_j = +1\} = 1/2 = \mathbb{P}\{\mathfrak{s}_j = -1\}$ (so-called ***rademacher*** variables). Via the fact that $k!2^k \le (2k)!$ for each positive integer k we have

$$\mathbb{P}e^{\lambda c_j \mathfrak{s}_j} = \tfrac{1}{2}\left(e^{\lambda c_j} + e^{-\lambda c_j}\right) = \sum\nolimits_{k=0}^{\infty} (\lambda c_j)^{2k}/(2k)! \le \exp(\lambda^2 c_j^2/2)$$

so that $\mathbb{P}e^{\lambda S} = \prod_{j\le n} \mathbb{P}e^{\lambda c_j \mathfrak{s}_j} \le e^{\lambda^2\tau^2/2}$ with $\tau^2 := \sum_{j\le n} c_j^2$. That is, the sum S has a $\text{SUBG}(\tau^2)$ distribution, so that

$$\mathbb{P}\{\pm S \ge r\} \le \exp\left(-\tfrac{1}{2}r^2/\sum\nolimits_i c_i^2\right) \qquad \text{for } r \ge 0.$$

The $\pm$ is just my lazy way of indicating that the inequality holds for both S □ and $-S$, that is, it holds for both upper and lower tails.

More generally, if S is a sum of independent subgaussians $X_1, \ldots, X_n$ then the equality $\mathbb{P}e^{\lambda S} = \prod_j \mathbb{P}e^{\lambda X_j}$ shows that $X \in \text{SUBG}(\tau^2)$ for $\tau^2 = \sum_j \tau^2(X_j)$. For sums of independent variables we need consider only the subgaussianity of each summand.

7.2 Characterizations of subgaussianity

The calculations at the end of Section 7.1 show the advantage of defining subgaussianity using moment generating functions. For other purposes equivalent definitions, as identified by the following Theorem, are sometimes more convenient.

<6> **Theorem.** *For each random variable X with expected value 0, the following assertions are equivalent.*

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

(ii) $L(X) := \sup_{k\in\mathbb{N}} \|X\|_{2k} / \sqrt{k}$ *is finite*

(iii) $X \in \mathcal{L}^{\Psi_2}$ *for the orlicz function* $\Psi_2(x) = \exp(x^2) - 1$

(iv) there exists a positive constant β *for which*

$$\mathbb{P}\{|X| \ge t\} \le 2\exp\left(-\tfrac{1}{2}t^2/\beta^2\right) \qquad \text{for all } t \ge 0.$$

More precisely,

$$\|X\|_{\Psi_2} / \sqrt{6} \le \beta(X) \le \tau(X) \le \sqrt{8}\, L(X)$$

and $L(X) \le \|X\|_{\Psi_2} \le \sqrt{2e} L(X)$*, where* $\beta(X)$ *denotes the smallest* β *for which (iv) holds.*

Remarks.

(i) Implicit in the sequences of inequalities is the assertion that finiteness of one quantity implies finiteness of another quantity.

(ii) It might be more natural use $\sqrt{2k}$ instead of $\sqrt{k}$ in the definition of $L(X)$. The choice $\sqrt{k}$ ensures that $\|X\|_p \le L(X)\sqrt{p}$ for all real $p \ge 1$: if $k = \lceil p/2 \rceil$ then $\|X\|_p \le \|X\|_{2k}$ and $p \ge k$.

Proof. Simplify notation by writing β for $\beta(X)$ and τ for $\tau(X)$. and L for $L(X)$. Also, assume that X is not a constant, to avoid annoying trivial cases.

For $\beta \le \tau$:
Use inequality <1>.

For $\|X\|_{\Psi_2}$ and β:

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

For L versus $\|X\|_{\Psi_2}$:
This result is mostly an exercise in applying the inequality $k/e \le (k!)^{1/k} \le k$. It does not need $\mathbb{P}X = 0$. First note that $\mathbb{P}\Psi_2(|X|/D) = \sum_{k\in\mathbb{N}} \|X/D\|_{2k}^{2k}/k!$. If $\infty > D > \|X\|_{\Psi_2}$ then the sum is bounded by 1, which implies $\|X\|_{2k} \le D(k!)^{1/2k} \le D\sqrt{k}$ for each k. Conversely,

$$\mathbb{P}\Psi_2(|X|/D) \le \sum\nolimits_{k\in\mathbb{N}} (L\sqrt{k}/D)^{2k}(e/k)^k = 1 \qquad \text{if } D = \sqrt{2e}\,L.$$

For $\tau \le 4L$:
Let X' be a random variable with the same distribution as X but independent of X. By the triangle inequality, $\|X - X'\|_{2k} \le 2\|X\|_{2k} \le 2L\sqrt{k}$ for all $k \in \mathbb{N}$. By independence, $\mathbb{P}_X X' = \mathbb{P}X' = 0$. Via Jensen's inequality for the conditional expectation operator $\mathbb{P}_X$ we have

$$\begin{aligned}
\mathbb{P}e^{\lambda X} &= \mathbb{P}e^{\lambda(X - \mathbb{P}_X X')} \le \mathbb{P}e^{\lambda(X-X')} = 1 + \sum\nolimits_{j\in\mathbb{N}} \lambda^j \mathbb{P}(X - X')^j/j! \\
&= 1 + \sum\nolimits_{k\in\mathbb{N}} \frac{\lambda^{2k}\mathbb{P}(X-X')^{2k}}{(2k)!} \qquad \text{symmetry kills odd moments} \\
&\le 1 + \sum\nolimits_{k\in\mathbb{N}} \frac{(\lambda 2L\sqrt{k})^{2k}}{k^k k!} = 1 + \sum\nolimits_{k\in\mathbb{N}} \frac{(4\lambda^2 L^2)^k}{k!} = e^{4L^2\lambda^2},
\end{aligned}$$

and so on. If you worry about the legitimacy of taking $\mathbb{P}$ inside $\sum_j$, note that

$$\sum\nolimits_{j\in\mathbb{N}} \mathbb{P}|\lambda^j(X-X')^j|/j! \le \sum\nolimits_{j\in\mathbb{N}} |\lambda|^j (2L\sqrt{j}\,)^j/(j/e)^j < \infty$$

□ then appeal to Dominated Convergence.

Remark. As mentioned in Chapter 3, the name 'subgaussian' is often applied in a loose sense to any tail bound that decreases like $\exp(-Ct^2)$. For example, Section 3.7 showed that the upper tail for the $\text{Bin}(n,p)$ is subgaussian in this sense, with even the best constant $C^{-1} = 2np(1-p))$ when $p \ge 1/2$.

I tried without much success to extend Theorem <6> to an analogous result for 'subgaussianity of a single tail'. Part of the difficulty was the possibility that the centering constant need not be equal to the mean. See Problem [1]. Another difficulty was the failure of symmetrization when only one tail was required to decrease in a subgaussian way.

Before giving up completely I did have some partial success with the idea that subgaussian in the upper tail should correspond to subgaussianity of $(X-\nu)^+$ for some constant ν. Unfortunately, the fact that $\mathbb{P}(X-\nu)^+ = 0$ only when $\mathbb{P}\{X > \nu\} = 0$ got in the way of clean Taylor expansions near the origin.

7.3 Hoeffding's inequality for independent summands

In a very famous paper, Hoeffding (1963) proved several exponential tail bounds for sums of independent random variables. He also extended the

results to a number of dependent settings (including martingales—see Section 7.4). The following result is probably the best known from his paper.

<7> **Theorem.** *(Hoeffding, 1963, Theorem 2) Suppose $S_n = \sum_i X_i$, a sum of independent random variables with $a_i \le X_i \le b_i$ for each i, for constants a_i and b_i. Then*

$$S_n - \mathbb{P}S_n \in \text{SUBG}(\tau^2) \qquad \text{for } \tau^2 := \textstyle\sum_i c_i^2, \text{ where } c_i := (b_i - a_i)/2.$$

Consequently,

<8> $$\mathbb{P}\{S_n - \mathbb{P}S_n \ge r\} \le \exp\left(-r^2/2\tau^2\right) \qquad \text{for each } r \ge 0,$$

with an analogous inequality for the lower tail.

Proof. First show that $X_i - \mathbb{P}X_i$ has a $\text{SUBG}(c_i^2)$ distribution, for each i. The simplest approach uses facts about $L_i(\lambda) := \log \mathbb{P}e^{\lambda X_i}$ that were established in Section 2.3:

$$\dot{L}_i(\lambda) = \mathbb{P}_{\lambda,i} X_i \qquad \text{AND} \qquad \ddot{L}_i(\lambda) = \text{var}_\lambda(X_i) := \mathbb{P}_{\lambda,i}\left(X_i - \mathbb{P}_{\lambda,i}X_i\right)^2,$$

where $\mathbb{P}_{\lambda,i}$ is the probability measure defined by the tilted density $e^{\lambda X_i}/\mathbb{P}e^{\lambda X_i}$ with respect to $\mathbb{P}$. As $\mathbb{P}_{\lambda,i}\{a_i \le X_i \le b_i\} = 1$ it follows that

$$\text{var}_\lambda(X_i) \le \mathbb{P}_{\lambda,i}\left(X_i - (a_i + b_i)/2\right)^2 \le c_i^2 \qquad \text{for every } \lambda.$$

In particular, for some λ^* between 0 and λ,

$$L_i(\lambda) = L_i(0) + \lambda \dot{L}_i(0) + \tfrac{1}{2}\lambda^2 \ddot{L}_i(\lambda^*) \le 0 + \lambda \mathbb{P}X_i + \tfrac{1}{2}\lambda^2 c_i^2.$$

It then follows that

$$\mathbb{P}e^{\lambda(S_n - \mathbb{P}S_n)} = \prod\nolimits_{i \le n} \mathbb{P}e^{\lambda(X_i - \mathbb{P}X_i)} \le e^{\lambda^2 \sum_{i \le n} c_i^2/2},$$

□ the deired subgaussian bound.

Remark. In the paper, Hoeffding first used convexity of the exponential function to get

<9> $$e^{\lambda X_i} \le \frac{b - X_i}{b - a} e^{\lambda a} + \frac{X_i - a}{b - a} e^{\lambda b}.$$

Then he took expected values:

$$\begin{aligned} \mathbb{P}e^{\lambda X_i} &\le \theta_i e^{\lambda a_i} + (1 - \theta_i) e^{\lambda b_i} && \text{where } \theta_i := (b_i - \mathbb{P}X_i)/(b_i - a_i) \\ &\le \exp\left(\lambda \mathbb{P}X_i + c_i^2 \lambda^2/2\right) && \text{by brute force calculus.} \end{aligned} \tag{<10>}$$

The inequality for <10> could also be established by noting that $\theta_i e^{\lambda a_i} + (1 - \theta_i) e^{\lambda b_i}$ is the MGF of a random variable W for which $\theta_i = \mathbb{P}\{W = a_i\} = 1 - \mathbb{P}\{W = b_i\}$ and $\mathbb{P}W = \theta_i a_i + (1 - \theta_i) b_i = \mathbb{P}X_i$. The brute force calculation is essentially equivalent to the representation of the second derivate of $\log \mathbb{P}e^{\lambda W}$ as a variance under a tilted distribution.

<11> **Example.** The result from Example <5> is a special case of Theorem <7> with $X_j = c_j \mathfrak{s}_j$, a random variable that takes values in the interval $[-c_j, c_j]$, with $\mathbb{P}X_j = 0$ and $\operatorname{var}(X_j) = c_j^2$. In a strong sense this case is extreme, because the distribution of X_j concentrates at the endpoints of the interval.

In general, one quarter of the squared length of the range can be a vast overestimate of the variance. For example, suppose $S \sim \text{BIN}(n, p)$. Then $\mathbb{P}S = np$ and $S - np = \sum_{j \le n} X_j$ where $X_j = \xi_j - p$ with $\xi_j \sim \text{BER}(p)$. The random variable X_j takes values in the interval from $-p$ to $1-p$, which has length 1. From Theorem <7>, $\mathbb{P}\{S - np \ge r\} \le \exp(-2r^2/n)$ for $r \ge 0$. The bound for the lower tail is the same. Compare with the tail bounds from Section 3.7:

<12> $$\mathbb{P}\{S \ge np + r\} \le \exp\left(-\frac{r^2}{2npq} g_{n,p}(r)\right) \qquad \text{for } 0 \le r \le nq$$

where $q = 1 - p$ and, for $r \ge 0$,

$$\begin{aligned} g_{n,p}(r) &:= q\psi_{\text{benn}}\left(\frac{r}{np}\right) + p\psi_{\text{benn}}\left(\frac{-r}{nq}\right) \\ &\ge \psi_{\text{benn}}\left(q\frac{r}{np} + p\frac{-r}{nq}\right) \qquad \text{by convexity of } \psi_{\text{benn}} \\ &= \psi_{\text{benn}}\left(\frac{(q-p)r}{npq}\right) \gtreqless 1 \text{ if } p \gtreqless 1/2. \end{aligned}$$

Inequality <12>, which is derived from a minimization with the true mgf, cannot be worse than the bound from Theorem <7>, which carried out an analogous minimization that started from an upper bound for the same mgf. For p equal to $1/2$ and r much smaller than n the bound from Theorem <7> is comparable to the bound given by <12>; for $p \ne 1/2$ the $2r^2/n$ is markedly inferior to the $r^2/(2npq)$ (particularly so if $p > 1/2$, which ensures that $g_{n,p}(r) \ge 1$). □

The previous Example highlights a weakness of Theorem <7> when the distribution of a bounded random variable concentrates most of its probability a long way from the endpoints of its support. The squared range times $1/4$ can then be a very poor substitute for a variance term. Chapter 8 will show how better control is possible when the actual variances enter a tail bound.

7.4 From independence to martingales

Theorem <7> has a martingale analog, which has proved particularly useful for the development of concentration inequalities. The result is often attributed to Azuma (1967), even though Hoeffding (1963, page 18) had already noted that, if $S_m = X_1 + \cdots + X_m$ for $1 \le m \le n$, then

> "..., the inequalities of Theorems 1 and 2 remain true if the assumption that $X_1, X_2, \ldots, X_n$ are independent is replaced by

> the weaker assumption that the sequence $S'_m = S_m - ES_m$, $m = 1, 2, \ldots, n$, is a martingale, that is,
>
> $$E(S'_m \mid S'_1, \ldots, S'_j) = S'_j, \qquad 1 \le j \le m \le n, \quad \text{(H2.18)}$$
>
> with probability one. ... [(H2.18)] implies that the conditional mean of X_m for S'_{m-1} fixed is equal to its unconditional mean. A slight modification of the proofs of Theorems 1 and 2 yields the stated result."

He also noted that martingale methods provide stronger maximal inequalities. (I added the 'H' to indicate that (2.18) was Hoeffding's equation number.)

I am not completely sure about what Hoeffding meant regarding the martingale analog of his Theorem 2. I suspect it involved writing $S_n - \mathbb{P}S_n$ as a sum of random variables $X_1 + \cdots + X_n$ with zero conditional means and $a_i \le X_i \le b_i$, applying inequality <9> to each X_i then taking conditional expectations. In spite of this evidence, the martingale analog of Hoeffding's Theorem 2 is often attributed to Azuma in the literature; 'Hoeffding's inequality' usually refers to the result for independent summands described in Section 7.3.

> **Remark.** Of course, it would be silly to assume $0 \le X_i \le 1$ with a zero conditional expectation for X_i. Comments following the statement of Hoeffding's Theorem 1 suggest he was thinking about the case $a \le X_i \le b$ when he made the comment about the martingale modification.

The extension to martingales might appear rather easy. For a martingale $\{(S_i, \mathcal{F}_i) : 0 \le i \le n\}$ we would have $S_0 = \mathbb{P}(S_n \mid \mathcal{F}_0) = \mathbb{P}S_n$ if $\mathcal{F}_0 = \{\emptyset, \Omega\}$, the trivial sigma-field. We could even replace S_i by $S_i - \mathbb{P}S_n$ and then forget about S_0 altogether, working directly with the martingale differences $X_1, \ldots, X_n$ for which $\mathbb{P}(X_i \mid \mathcal{F}_{i-1}) = 0$ almost surely and $S_i = X_1 + \cdots + X_i$ for $1 \le i \le n$. We would also have $\mathcal{F}_i = \sigma\{X_1, \ldots, X_i\}$ with $\mathcal{F}_0$ again the trivial sigma-field. We would need an assumption like

<13> $$\mathbb{P}\{a_i \le X_i \le b_i \mid \mathcal{F}_{i-1}\} =_{a.s} 1,$$

for possibly random a_i and b_i that depend only on $\{X_j : j \le i-1\}$. One could then hope that the methods from the Proof of Theorem <7> carry over to show, for $\lambda \ge 0$, that

<14> $$\mathbb{P}\left(e^{\lambda X_i} \mid \mathcal{F}_{i-1}\right) \le_{a.s.} \exp\left(\lambda^2 c_i(\omega)^2/2\right) \qquad \text{if } c_i \ge (b_i - a_i)/2.$$

Is the extension really that straightforward?

A message from the measure theory police:
You can ignore this message, at the slight risk of violating the usual laws of measure theoretic probability.

With the Kolmogorov-style interpretations of conditional expectations, assertions <13> and <14> mean

$$\mathbb{P}g(X_1,\ldots,X_{i-1})\{a_i \le X_i \le b_i\} = \mathbb{P}g(X_1,\ldots,x_{i-1}), \tag{15}$$

$$\mathbb{P}g(X_1,\ldots,X_{i-1})e^{\lambda X_i} \le \mathbb{P}g(X_1,\ldots,x_{i-1})e^{\lambda^2 c_i^2/2} \qquad \text{for } g \ge 0, \tag{16}$$

at least for bounded, measurable functions g. The a.s. subscripts in <13> and <14> reflect the fact that <15> and <16> determine the random variables $\mathbb{P}\{a_i \le X_i \le b_i \mid \mathcal{F}_{i-1}\}$ and $\mathbb{P}\left(e^{\lambda X_i} \mid \mathcal{F}_{i-1}\right)$ only up to almost sure equivalences. In particular, the second of these conditional expectations could be changed without warning on a negligible set that depends on λ, a possibility that bodes ill for any attempt to differentiate with respect to λ.

Equalities <15> and <16> also seem to require a_i, b_i and c_i to be measurable functions of their arguments, $X_1,\ldots,X_{i-1}$. Unfortunately, such an assumption would create measurability complications for the setting that will be described in Example <19>.

When delicate measurability issues arise, it is often useful to follow the example of careful authors—such as two of my mathematical heroes, Doob and Dudley—and work with an explicit representation for the underlying probability space $(\Omega,\mathcal{F},\mathbb{P})$. In this case, we could choose $\Omega = \mathbb{R}^{[\![n]\!]}$ with generic member $\omega = x = (x[1],\ldots,x[n])$ and $X_i(x) = x[i]$, the coordinate map. The probability measure $\mathbb{P}$ corresponds to the joint distribution of $X_1,\ldots,X_n$ and $\mathcal{F}_i$ is the sigma-field generated by $x[1],\ldots,x[i]$. Also with this notation, the vector $(X_1,\ldots,X_i)$ could be written as $x[1:i]$ and the vector $(X_{i+1},\ldots,X_n)$ as $x[i+1:n]$.

> **Remark.** *Here I am writing $x[i]$ instead of x_i because my old eyes have trouble with multiple levels of subscripts and superscripts. The shorthand $x[I] := (x[i] : i \in I)$ for $I \subset [\![n]\!]$ also simplifies the notation.*

The great advantage of such a representation is that $\mathbb{P}$ can be built up from an initial distibution P_1 for X_1 and a set of conditional distributions $P_{i|x[1:i-1]}$ for $\mathbb{P}\{X_i \in \cdot \mid X_1,\ldots,X_{i-1}\}$. That is, at least for bounded, measurable functions $g : \mathbb{R}^{[\![n]\!]} \to \mathbb{R}$, we have

$$\mathbb{P}g(X_1,\ldots,X_n) = P_1^{x[1]} P_{2|x[1]}^{x[2]} \ldots P_{n|x[1:n-1]}^{x[n]} g(x),$$

$$\mathbb{P}\left(g(X_1,\ldots,X_n) \mid \mathcal{F}_i\right) =_{a.s.} g_i(x[1:i]) = \mathbb{Q}_{i+1|x[1:i]}^{x[i+1:n]} g(x),$$

*where, symbolically, $\mathbb{Q}_{i+1} := P_{i+1|x[1:i]}^{x[i+1]} \ldots P_{n|x[1:n-1]}^{x[n]}$. The $P_{i|x[1:i-1]}$'s are often called **regular conditional probability distributions** (Breiman, 1968, Section 4.3). Read Pollard (2001, Chap. 5) if this notation worries you.*

Equalities <15> and <16> can now be rewritten as:

$P_{i|x[1:i-1]}$ concentrates on the interval $[a_i, b_i]$,
possibly with a_i and b_i both functions of $x[1:i-1]$;

<17> $P_{i|x[1:i-1]} e^{\lambda x[i]} \le \exp\left(\lambda^2 (b_i - a_i)^2/8\right) \le \exp\left(\lambda^2 c_i^2/2\right)$,
where c_i can depend on $x[1:i-1]$ and $c_i \ge (b_i - a_i)/2$,

without almost sure caveats. The inequality <17> could be proved by either of the methods described in Section 7.3. Neither equation relies on an assumption that a_i, b_i, c_i depend on $x[1:i-1]$ in a measurable way, because the averaging is carried out separately for each fixed $x[1:i-1]$, a small but significant improvement over the almost sure versions.

□ *End of message.*

If a_i and b_i are actually constants then the method used to prove Theorem <7> carries over with minimal changes. For martingale differences $\{X_i\}$ with truly random c_i's a new complication arises. We now need c_i to depend on $X_1, \ldots, X_{i-1}$ in a measurable way. That is, we need c_i to be $\mathcal{F}_{i-1}$-measurable, so that factors can be pulled out one at a time by successive conditioning arguments. Start with the X_n contribution:

$$\mathbb{P}\left(e^{\lambda S_n} \mid \mathcal{F}_{n-1}\right) = e^{\lambda S_{n-1}} \mathbb{P}\left(e^{\lambda X_n} \mid \mathcal{F}_{n-1}\right) \le e^{\lambda S_{n-1}} e^{\lambda^2 c_n^2/2} \qquad \text{a.s.} \quad .$$

The factor $\exp(\lambda^2 c_n^2/2)$ would present an obstacle if we now tried to condition on $\mathcal{F}_{n-2}$. Such an obstacle can be removed by absorbing the $\mathcal{F}_{n-1}$-measurabile factor involving c_n into the left-hand side:

$$\mathbb{P}\left(e^{\lambda S_n - \lambda^2 c_n^2/2} \mid \mathcal{F}_{n-1}\right) \le e^{\lambda S_{n-1}} \qquad \text{a.s.} \quad .$$

A further conditioning on $\mathcal{F}_{n-2}$ would then throw out an $\exp\left(\lambda^2 c_{n-1}^2/2\right)$ factor. Of course this new obstacle could also be removed by also absorbing the c_{n-1} contribution into the left-hand side. And so on. (This is an old trick much used in the literature on martingale central limit theorems.) You should now understand how the proof of the next Theorem works.

<18> **Theorem.** *Let $((X_i, \mathcal{F}_i) : i \in [\![n]\!])$ be a martingale difference sequence, with $\mathbb{P}(X_i \mid \mathcal{F}_{i-1}) =_{a.s.} 0$, with $\mathcal{F}_0$ the trivial sigma-field $\{\emptyset, \Omega\}$. Define $S_i := \sum_{j=1}^{i} X_j$ for $i \in [\![n]\!]$.*

(i) Suppose the conditional distribution of X_i given $\mathcal{F}_{i-1}$ concentrates on an interval $[a_i, b_i]$, with both a_i and b_i being functions of $X_1, \ldots, X_{i-1}$.

(ii) Let c_i be any $\mathcal{F}_{i-1}$-measurable upper bound for $(b_i - a_i)/2$. Define $V_0 = 0$ and $V_i := \sum_{j=1}^{i} c_j^2$ with for $i \in [\![n]\!]$.

Then for each positive constant σ^2,

$$\mathbb{P}\{S_n \ge r, V_n \le \sigma^2\} \le e^{-r^2/(2\sigma^2)} \qquad \text{for } r \ge 0.$$

□ *Consequently, $\mathbb{P}\{S_n \ge r\} \le e^{-\lambda^2 (r/\sigma)^2/2} + \mathbb{P}\{V_n > \sigma^2\}$ for $r \ge 0$.*

Proof. The second inequality comes from adding $\mathbb{P}\{S_n \geq r, V_n > \sigma^2\}$ to both sides of the first inequality.

For the main argument define $W_0 := e^{\lambda^2\sigma^2/2}$ and, for $j \in [\![n]\!]$,

$$W_j := \exp\left(\lambda S_j + \lambda^2(\sigma^2 - V_j)/2\right) = W_{j-1}\exp\left(\lambda X_j - \lambda^2 c_j^2/2\right).$$

Take conditional expectations then invoke inequality <14> to deduce

$$\mathbb{P}\left(W_j \mid \mathcal{F}_{j-1}\right) = W_{j-1}e^{-\lambda^2 c_j^2/2}\mathbb{P}\left(e^{\lambda X_j} \mid \mathcal{F}_{j-1}\right) \leq W_{j-1},$$

which implies $\mathbb{P}W_j \leq \mathbb{P}W_{j-1}$. From the pointwise inequality for $\lambda > 0$,

$$\{S_n \geq r, V_n \leq \sigma^2\} \leq \exp\left(\lambda(S_n - r) + \lambda^2(\sigma^2 - V_n)/2\right) = e^{-\lambda r}W_n,$$

we get

$$\mathbb{P}\{S_n \geq r, V_n \leq \sigma^2\} \leq e^{-\lambda r}\mathbb{P}W_n \leq e^{-\lambda r}\mathbb{P}W_0 = e^{-\lambda r + \lambda^2\sigma^2/2} \qquad \text{for } \lambda \geq 0.$$

The usual subgaussian minimization argument then delivers the asserted tail bound. □

<19> **Example.** Suppose $\xi = (\xi_1, \ldots, \xi_n)$ is a vector of independent random quantities, with each ξ_i taking values in some measurable space $(\mathcal{X}_i, \mathcal{A}_i)$, with ξ_i having distribution P_i. (You will lose no great generality if you assume $\mathcal{X}_i = \mathbb{R}$ equipped with its borel sigma-field.) Suppose also that f is an $\otimes_i\mathcal{A}_i$-measurable real-valued function on the product space $\prod_i \mathcal{X}_i$ with the ***bounded difference*** property: for constants c_i,

$$|f(y) - f(w)| \leq c_i \qquad \text{if } y \text{ and } w \text{ differ only in the } i\text{th coordinate.}$$

More succinctly,

<20> $$|f(y_1, \ldots, y_n) - f(w_1, \ldots, w_n)| \leq \sum\nolimits_{i \leq n} c_i\{y_i \neq w_i\} \qquad \text{for all } y, w \in \mathcal{X}.$$

Let $\mathcal{F}_i$ denote the sigma-field generated by $\xi_1, \ldots, \xi_i$, with $\mathcal{F}_0$ the trivial sigma-field. Define a martingale $\{(S_i, \mathcal{F}_i) : 0 \leq i \leq n\}$ by $S_i := \mathbb{P}\left(f(\xi) \mid \mathcal{F}_i\right) - \mathbb{P}f(\xi)$. Note that $S_0 = 0$ because $\mathcal{F}_0$ is trivial.

An appeal to Theorem <18> with $\sigma^2 = \sum_{i=1}^n c_i^2$ will give the subgaussian bound

<21> $$\mathbb{P}\{f(\xi) \geq r + \mathbb{P}f(\xi)\} \leq \exp\left(-r^2/2\sigma^2\right) \qquad \text{for } r \geq 0.$$

Remark. Inequality <20> is sometimes referred to as a lipschitz condition (for a weighted hamming distance). Inequality <21> is often referred to as ***McDiarmid's inequality***.

To calculate $\mathbb{P}\left(f(\xi) \mid \mathcal{F}_i\right)$ we have only to average out over the independent coordinates $\xi_{i+1}, \ldots, \xi_n$, whose joint distribution is $\mathbb{Q}_{i+1} = \otimes_{i<j\leq n}P_j$:

$$\mathbb{P}\left(f(\xi) \mid \mathcal{F}_i\right) =_{a.s.} f_i(\xi_1, \ldots, \xi_i) := \mathbb{Q}_{i+1}f(\xi_1, \ldots, \xi_i, z_{i+1}, \ldots, z_n),$$

where $\mathbb{Q}_{i+1}$ integrates over the z's.

By construction, $f(\xi) - \mathbb{P}f(\xi) = S_n = \sum_{j=1}^n X_j$, a sum of martingale differences $X_i := S_i - S_{i-1}$ for $i \in [\![n]\!]$. The increment can also be written as

$$\begin{aligned} X_i &= f_i(\xi_1, \ldots, \xi_i) - f_{i-1}(\xi_1, \ldots, \xi_{i-1}) \\ &= \mathbb{Q}_i f(\xi_1, \ldots, \xi_i, z_{i+1}, \ldots, z_n) - \mathbb{Q}_{i-1} f(\xi_1, \ldots, \xi_{i-1}, z_i, z_{i+1}, \ldots, z_n) \\ &= \mathbb{Q}_i P_i \Big[f(\xi_1, \ldots, \xi_i, z_{i+1}, \ldots, z_n) - f(\xi_1, \ldots, \xi_{i-1}, z_i, z_{j+1}, \ldots, z_n) \Big]. \end{aligned}$$

The extra P_i has no effect on the ξ_i.

Let me first show you a *suboptimal* way to proceed. We could bound X_i by arguing that:

$$\begin{aligned} &|f_j(y_1, \ldots, y_j) - f_{j-1}(y_1, \ldots, y_{j-1})| \\ &\quad \le \mathbb{Q}_i P_i |f(y_1, \ldots, y_i, z_{i+1}, \ldots, z_n) - f(y_1, \ldots, y_{i-1}, z_i, z_{i+1}, \ldots, z_n)| \\ &\quad \le \mathbb{Q}_i P_i c_i = c_i \qquad \text{via } <20>. \end{aligned}$$

The thing to notice is that P_i integrates out the z_i but ignores the y_i. This inequality suggests choosing $a_i = -c_i$ and $b_i = c_i$, which would contribute an extra, unwanted factor of $1/4$ in the exponent of $<21>$, as pointed out by McDiarmid (1989, page 159, after Lemma 4.1).

McDiarmid took a better approach, based on a very clever insight. First note that f_i inherits a bounded difference property from f:

$$\begin{aligned} &|f_i(y_1, \ldots, y_i) - f_i(w_1, \ldots, w_i)| \\ &\quad \le \mathbb{Q}_i |f(y_1, \ldots, y_i, z_{i+1}, \ldots, z_n) - f(w_1, \ldots, w_i, z_{i+1}, \ldots, z_n)| \\ &\quad \le \sum\nolimits_{j \le i} c_j \{y_j \ne w_j\} \qquad \text{via } <20>. \end{aligned}$$

Then note that $a_i \le X_i \le b_i$ where

$$\begin{aligned} a_i &= \inf_{y \in \mathcal{X}_i} f_i(\xi_1, \ldots, \xi_{j-1}, y) - f_{i-1}(\xi_1, \ldots, \xi_{i-1}), \\ b_i &= \sup_{w \in \mathcal{X}_i} f_i(\xi_1, \ldots, \xi_{i-1}, w) - f_{i-1}(\xi_1, \ldots, \xi_{i-1}). \end{aligned}$$

Clearly the f_{j-1} term has no effect on the length of the interval $[a_i, b_i]$:

$$\begin{aligned} 0 \le b_i - a_i &= \sup_{w \in \mathcal{X}_i} f_j(\xi_1, \ldots, \xi_{j-1}, w) - \inf_{y \in \mathcal{X}_i} f_j(\xi_1, \ldots, \xi_{i-1}, y) \\ &= \sup_{w, y \in \mathcal{X}_i} |f_j(\xi_1, \ldots, \xi_{i-1}, w) - f_j(\xi_1, \ldots, \xi_{i-1}, y)|, \end{aligned}$$

which is $\le c_i$ by the bounded difference inequality for f_i. And so on.

If you worry about the effect of a taking a supremum over a possibly uncountably infinite set $\mathcal{X}_i$ you should note the lack of any measurabilty assumptions about a_i and b_i in Theorem $<18>$; it is only c_i that needs to be $\mathcal{F}_{i-1}$-measurable. You might also want to read the police message. □

Remark. McDiarmid (1989, page 168, proof of (6.10)) was obviously aware of the measurability issues. I am not quite sure exactly what he meant by $\operatorname{ess\,inf}[X_k \mid \mathcal{F}_{k-1}]$, but it is clearly intended to avoid a non-measurable supremum. McDiarmid is a careful author.

At this point you are probably expecting to see a triumphal procession of applications of Theorems <7> and <18> to problems in combinatorics and graph theory. I will give only two examples. The first (bin packing) is quite straightforward. I include it here because it was one of the first examples that convinced me concentration is a general topic worth careful study. The second example, which will appear in in Section 7.5, will emphasize the limitations of these methods. For more uplifting applications I would advise you to look at the beautiful books by Steele (1997) and Alon and Spencer (2000), or to read McDiarmid (1989, 1998), or even just google "Hoeffding", "Azuma", "bounded difference", or "McDiarmid inequality". Be warned that your googling might also dredge up results that are discussed in Chapter 8. See Boucheron, Lugosi, and Massart (2013, Chap 6), especially the Notes to the chapter, for some of the generalizations of the method of bounded differences.

<22> **Example.** Suppose objects of independent random sizes $\xi_1, \ldots, \xi_n$, each with distribution P concentrated on $[0,1]$, are packed into bins of capacity 1. No objects may be split and the contents of no bin is allowed to exceed capacity. Define $f(\xi_1, \ldots, \xi_n)$ to be the smallest number of bins needed to contain all the objects.

The function f has the bounded difference property on $[0,1]^n$, with $c_i = 1$ for each i. To see why, suppose vectors y and z differ only in the ith coordinate. Every packing for y can then be converted to a packing for z with one more bin by discarding y_i and putting z_i into a new bin. It follows that $f(z) \le 1 + f(y)$. Reverse the roles of y and z to get the companion inequality $f(y) \le 1 + f(z)$.

Notice that the resulting subgaussian tail bound makes no use of special properties of P such as its variance. For a more refined analysis see Section 17.2, which will use a brilliant idea of Talagrand to bring a variance term into the tail bound. □

7.5 A cautionary example

The following example was mentioned by Kim and Vu (2000) and Vu (2002) as motivation for developing a more general inequality (see Section 8.6) that is analogous to a bennett inequality for martingales.

<23> **Example.** (random graphs) Consider a graph with vertex set $[[n]]$ 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 random graph $G_n(p)$, which oftens carries the names of Erdös and Rényi, chooses its edges by means of a set of independent random vari-

ables $\{\xi_e : e \in \mathcal{E}\}$, with each ξ_e distributed $\text{BER}(p)$. That is, $G_n(p)$ includes e when $\xi_e = 1$, an event with probability p.

Remark. Actually Erdös and Rényi (1960) took the number of edges to be a nonrandom number N. They considered the graph $\Gamma_{n,N}$ whose edge set is a sample of size N selected without replacement from $\mathcal{E}$. As explained by Janson, Łuczak, and Ruciński (2000, Section 1.4), the random graph $G_n(p)$ behaves asymptotically very like $\Gamma_{n,N}$ with $N = \binom{n}{2}p$.

The number of triangles in $G_n(p)$ equals

$$f(\xi) := \sum\nolimits_{\{e_1,e_2,e_3\}\in\mathcal{T}} \xi_{e_1}\xi_{e_2}\xi_{e_3}.$$

The expected number of triangles is $\theta := \mathbb{P}f(\xi) = \binom{n}{3}p^3$.

To fit this example into the bounded difference setting of Example <19> we need to enumerate $\mathcal{E}$ as a sequence $(e_j : 1 \le j \le |\mathcal{E}|)$. I don't know of any particularly clever method for enumerating that can improve the concentration bound for $f(\xi)$.

A change in a single y_e can have a big effect on $f(y)$. For example, if $y_e = 1$ for all $e \in \mathcal{E}$ and just one y_e is changed to zero then $n-2$ triangles disappear. The constants corresponding to the c_i's in inequality <20> must all be equal to $n-2$. The two-sided analog of inequality <21> gives

$$\mathbb{P}\{|f(\xi) - \mathbb{P}f(\xi)| \ge r\} \le 2\exp\left(-\frac{r^2}{2|\mathcal{E}|(n-2)^2}\right) \qquad \text{for } r \ge 0. \tag{<24>}$$

We would need an $r = r_n$ with $r_n/n^2 \to \infty$ to send the tail bound to zero as $n \to \infty$.

Better bounds are available. For example, arguments based on the Chen-Stein method show that $f(\xi)$ is approximately $P_\theta = \text{POISSON}(\theta)$ distributed. More precisely, as shown by Barbour, Holst, and Janson (1992, Section 5.1),

$$\sup_A |\mathbb{P}\{f(\xi) \in A\} - P_\theta A| = O(np^2) \qquad \text{if } np \to \infty.$$

In particular, the choice $A = \{i \in \mathbb{N}_0 : |j - \theta| > r\}$ and the Poisson tail bound from Section 3.5,

$$P_\theta A \le 2\exp\left(-\frac{r^2}{2\theta}\psi_{\text{benn}}(r/\theta)\right) = 2\exp(-\theta\hbar(r/\theta)),$$

gives

$$\mathbb{P}\{|f(\xi) - \theta| > s_n\theta^{1/2}\} \to 0$$

if, for example, $s_n \to \infty$ and $s_n/\theta^{1/2} \to 0$ and $np^2 \to 0$.

Remark. The appeal to Chen-Stein has the drawback that it imposes the extraneous requirement that $f(\xi)$ have an approximate Poisson distribution. My only point in invoking Chen-Stein is to make sure you realize that <24> is very crude. It is possible to have concentration without an approximate Poisson distribution.

Why has inequality <21> fallen so far short of the Chen-Stein result? The blame lies with the pessimistic choice $c_i = n-2$ forced by the unlikely configuration y with $y_e = 1$ all $e \in \mathcal{E}$. It is much more likely that a possible edge, say $e = \{1,2\}$, is involved in only about np^2 potential triangles: the expected number of vertices $k \geq 3$ for which $\xi_{\{1,k\}} = 1 = \xi_{\{2,k\}}$ is equal to $(n-2)p^2$. A change in ξ_e probably changes $f(\xi)$ by an amount of order np^2. If by some modification of the argument we could reduce the effective c_i to a term of order np^2 then the exponent in the bound <24> would be greatly improved. Roughly speaking, this is the idea behind the Kim-Vu method, which will be discussed in Section 8.6. To be continued. □

7.6 Problems

[1] Suppose X is a random variable whose moment generating function $M_X(\lambda)$ is finite in a neighborhood of the origin. Suppose also that there exist constants ν and $\tau > 0$ for which $M_X(\lambda) \leq \exp(\nu\lambda + \frac{1}{2}\tau^2\lambda^2)$ for all $\lambda \geq 0$, but not necessarily for all real λ.

(i) Show that $\mathbb{P}\{X \geq \nu + \tau t\} \leq e^{-t^2/2}$ for all $t \geq 0$.

(ii) Use the Taylor expansions for λ near zero,

$$\begin{aligned} M_X(\lambda) &= 1 + \lambda \mathbb{P}X + o(\lambda), \\ \exp(\nu\lambda + \tfrac{1}{2}\sigma^2\lambda^2) &= 1 + \lambda\nu + o(\lambda), \\ \mathbb{P}e^{\lambda(X-\nu)} &= 1 + \lambda(\mathbb{P}X - \nu) + \tfrac{1}{2}\lambda^2 \mathbb{P}(X-\nu)^2 + o(\lambda^2), \\ \exp(\tfrac{1}{2}\sigma^2\lambda^2) &= 1 + \tfrac{1}{2}(\tau^2)\lambda^2 + o(\lambda^2), \end{aligned}$$

to deduce that $\mathbb{P}X \leq \nu$ and $\operatorname{var}(X) \leq \mathbb{P}(X-\nu)^2 \leq \tau^2$. As Problem [2] shows, τ^2 might need to be strictly larger than the variance.

(iii) If $M_X(\lambda) \leq \exp(\nu\lambda + \frac{1}{2}\tau^2\lambda^2)$ for all real λ, show that $\mathbb{P}X = \nu$.

[2] Theorem <6> suggests that $\|X\|_2$ might be comparable to $\beta(X)$ if X has a subgaussian distribution. It is easy to deduce from inequality <3> that $\|X\|_2 \leq 2\beta(X)$. 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

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

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

[3] 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,

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

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

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

[4] Suppose $\mathbb{P}\{X = -1/2\} = 2/3$ and $\mathbb{P}\{X = 1\} = 1/3$. Show that $X \in \text{SUBG}(1)$. Also show that $\mathbb{P}X = 0$ and $\mathbb{P}X^2 = 1/2 < 1$. You might also find it informative to compare the higher order moments of X with the correponding moments of the $N(0, 1)$.

[5] Suppose $\{X_n\}$ is a sequence of random variables for which

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

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

7.7 Notes

The essentials of the characterizations for subgaussians in Theorem <6> come from Kahane (1968, Exercise 6.10). He did not cite any source. Kahane (1960, pages 4-5) proved similar results. For that paper, it is not clear to me whether Kahane was merely reminding readers of well known facts or whether he was introducing subgaussianity as a new concept.

Apparently Azuma (1967) was unaware of Hoeffding's paper, although he did cite a 1966 paper of Chow, which in turn cited a 1963 paper of Kahane for the properties of "semi-gaussian" random variables.

The proof of Theorem <18> borrows ideas from Freedman (1975).

References

Alon, N. and J. H. Spencer (2000). *The Probabilistic Method* (second ed.). Wiley.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal 19*(3), 357–367.

Barbour, A. D., L. Holst, and S. Janson (1992). *Poisson Approximation*. Oxford University Press.

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

Breiman, L. (1968). *Probability*. Reading, Massachusets: Addison-Wesley.

Erdös, P. and A. Rényi (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci 5*(1), 17–60.

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

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

Janson, S., T. Łuczak, and A. Ruciński (2000). *Random Graphs*. Wiley.

Kahane, J.-P. (1960). Propriétés locales des fonctions à séries de Fourier aléatoires. *Studia Mathematica 19*(1), 1–25.

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

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

McDiarmid, C. (1989). On the method of bounded differences. In J. Siemons (Ed.), *Surveys in Combinatorics*, Volume 141 of *London Mathematical Society Lecture Notes*, pp. 148–188. Cambridge University Press.

McDiarmid, C. (1997). Centering sequences with bounded differences. *Combinatorics, Probability and Computing 6*(1), 79–86.

McDiarmid, C. (1998). Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsen, and B. Reed (Eds.), *Probabilistic Methods for Algorithmic Discrete Mathematics*, pp. 195–248. Springer-Verlag.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Steele, J. M. (1997). *Probability Theory and Combinatorial Optimization*. SIAM.

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

