Chapter 3

Independent random variables

As we have seen in the proof of Höffding's inequality, a standard fact that is used to generate tail estimates for a sum of independent random variables is based on the behaviour of the moment generating function: if $Z_1, ..., Z_N$ are independent random variables, then for any $\lambda > 0$,

$$Pr\left(\sum_{i=1}^{N} Z_i \geq t\right) = Pr\left(\exp\left(\lambda \sum_{i=1}^{N} Z_i\right) \geq \exp(\lambda t)\right)$$

$$\leq \exp(-\lambda t)\mathbb{E}\exp\left(\lambda \sum_{i=1}^{N} Z_i\right) = \exp(-\lambda t)\prod_{i=1}^{N}\mathbb{E}\exp(\lambda Z_i). \tag{3.1}$$

Therefore, by estimating each moment generating function $\mathbb{E}\exp(\lambda Z_i)$ and optimizing the choice of λ one may derive nontrivial tail bounds.

In what follows we will show how certain assumptions on the random variable Z can be used to bound its moment generating function $\mathbb{E}\exp(\lambda Z)$.

3.1 The sum of independent ψ_2 random variables

Let $Z_1, ..., Z_N$ be independent, centred random variables. If $Z_i \in L_{\psi_2}$ and $a = (a_1, ..., a_N) \in \mathbb{R}^N$, what is the behaviour of the tail of the random variable $\sum_{i=1}^{N} a_i Z_i$? Note that this random variable is just the linear functional defined by the random vector $\mathcal{Z} = (Z_1, ..., Z_N)$, acting on $a \in \mathbb{R}^N$.

We will answer this question using two different arguments. The first is based on estimates on the moment generating function of Z_i and the other will be based on a comparison argument.

Lemma 3.1.1 *There is an absolute constant c for which the following holds. If $Z \in L_{\psi_2}$ is a centred random variable then for any $\lambda > 0$,*

$$\mathbb{E}\exp(\lambda Z) \leq \exp(c\lambda^2 \|Z\|_{\psi_2}^2).$$

Before proving the lemma, us introduce an idea that will appear frequently in what follows. It allows one to replace a mean-zero random variable Z with its symmetric counterpart, εZ, where ε is a symmetric, $\{-1,1\}$-valued random variable that is independent of Z.

Lemma 3.1.2 *Let ϕ be a convex function and let Z be a mean-zero random variable. Then*

$$\mathbb{E}\phi(Z) \leq \mathbb{E}\phi(2\varepsilon Z).$$

Proof. Let Z_1, Z_2 be independent copies of Z. By Jensen's inequality,

$$\mathbb{E}_{Z_1}\phi(Z_1) = \mathbb{E}_{Z_1}\phi(Z_1 - \mathbb{E}_{Z_2}Z_2) \leq \mathbb{E}_{Z_1}\mathbb{E}_{Z_2}\phi(Z_1 - Z_2) = (*). \tag{3.2}$$

The random variable $Z_1 - Z_2$ is symmetric, and in particular has the same distribution as $-1\cdot(Z_1 - Z_2)$. Therefore, $\mathbb{E}_{Z_1}\mathbb{E}_{Z_2}\phi(Z_1 - Z_2) = \mathbb{E}_{Z_1}\mathbb{E}_{Z_2}\phi(\varepsilon(Z_1 - Z_2))$ for every realization of the symmetric random variable ε that is independent of Z_1 and Z_2. Taking the expectation with respect to ε and using the convexity of ϕ, Fubini's Theorem, and that Z_1 and Z_2 have the same distribution as Z,

$$\begin{aligned}(*) =& \mathbb{E}\phi(\varepsilon(Z_1 - Z_2)) = \mathbb{E}\phi\left(\frac{1}{2}\cdot 2\varepsilon Z_1 + \frac{1}{2}\cdot(-2\varepsilon Z_1)\right) \\ \leq& \mathbb{E}_\varepsilon \mathbb{E}_{Z_1}\mathbb{E}_{Z_2}\cdot\frac{1}{2}\left(\phi(2\varepsilon Z_1) + \phi(2\varepsilon Z_2)\right) = \mathbb{E}\phi(2\varepsilon Z). \end{aligned} \tag{3.3}$$

∎

In particular, if Z is a mean-zero random variable then $\mathbb{E}\exp(\lambda Z) \leq \mathbb{E}\exp(2\lambda\varepsilon Z)$, and so from here on we may assume, if needed, that Z is a symmetric random variable.

An almost identical argument can be used to show the following:

Theorem 3.1.3 *Let $Z_1, ..., Z_N$ be independent, centred random variables and let $\phi : \mathbb{R}^N \to \mathbb{R}$ be a convex function. If $(\varepsilon_i)_{i=1}^N$ are independent, symmetric, $\{-1,1\}$-valued random variables then*

$$\mathbb{E}\phi\left(\sum_{i=1}^N Z_i\right) \leq \mathbb{E}\phi\left(2\sum_{i=1}^N \varepsilon_i Z_i\right).$$

We will later see more sophisticated symmetrization results.

Proof of Lemma 3.1.1. Applying Lemma 3.1.2 to the convex function $t \to \exp(\lambda t)$,

$$\mathbb{E}\exp(\lambda Z) \leq \mathbb{E}\exp(2\lambda\varepsilon Z) = \mathbb{E}\left(1 + 2\lambda\varepsilon Z + \sum_{j\geq 2}(2\lambda)^j \frac{(\varepsilon Z)^j}{j!}\right) = (*).$$

It is straightforward to verify (e.g., by the Monotone Convergence Theorem for the positive and negative parts of εZ) that

$$(*) \leq 1 + \sum_{j\geq 2}(2\lambda)^{2j}\frac{\mathbb{E}Z^{2j}}{(2j)!} \leq 1 + \sum_{j\geq 2}(2\lambda)^{2j}\frac{\|Z\|_{\psi_2}^j(2cj)^j}{j!j^j},$$

where we have used the fact that $\|Z\|_{L_p} \leq c\sqrt{p}\|Z\|_{\psi_2}$ and that $(2j)! \geq j!j^j$. Therefore,

$$\mathbb{E}\exp(\lambda Z) \leq 1 + \sum_{j\geq 2}\frac{1}{j!}\cdot\left(c_1\lambda^2\|Z\|_{\psi_2}^2\right)^j \leq \exp(c_1\lambda^2\|Z\|_{\psi_2}^2).$$

■

Lemma 3.1.1 leads to the desired estimate one $\|\sum_{i=1}^N a_i Z_i\|_{\psi_2}$.

Corollary 3.1.4 *There is an absolute constant c for which the following holds. Let $Z_1, ..., Z_N$ be independent, centred, ψ_2 random variables, and let $a_1, ..., a_N \in \mathbb{R}$. Then*

$$\|\sum_{i=1}^N a_i Z_i\|_{\psi_2} \leq c\left(\sum_{i=1}^N a_i^2\|Z\|_{\psi_2}^2\right)^{1/2}.$$

Proof. Since a norm is a convex function of its argument, it follows from Theorem 3.1.3 that we may assume without loss of generality that the a_i's are nonnegative and that the Z_i's are symmetric. Fix $\lambda > 0$ to be named later and observe that by Lemma 3.1.1,

$$\begin{aligned}\mathbb{E}\exp(\lambda\sum_{i=1}^N a_i Z_i) &= \prod_{i=1}^N \mathbb{E}\exp(\lambda a_i Z_i) \leq \prod_{i=1}^N \exp(c\lambda^2 a_i^2\|Z_i\|_{\psi_2}^2) \\ &= \exp(c\lambda^2\sum_{i=1}^N a_i^2\|Z_i\|_2^2).\end{aligned}$$

Therefore,

$$Pr(\sum_{i=1}^N a_i Z_i > t) \leq \exp(-\lambda t)\mathbb{E}\exp(\lambda \sum_{i=1}^N a_i Z_i) \leq \exp(-\lambda t + c\lambda^2 \sum_{i=1}^N a_i^2 \|Z_i\|_2^2).$$

By setting $\lambda = t/2c\sum_{i=1}^N a_i^2\|Z_i\|_2^2$ it follows that for every $t > 0$,

$$Pr(\sum_{i=1}^N a_i Z_i > t) \leq \exp\left(-\frac{t^2}{2}c\sum_{i=1}^N a_i^2\|Z_i\|_2^2\right),$$

and recalling that the Z_i's are symmetric, it is evident that

$$Pr\left(\left|\sum_{i=1}^N a_i Z_i\right| > t\cdot\left(c_1\sum_{i=1}^N a_i^2\|Z_i\|_{\psi_2}^2\right)^{1/2}\right) \leq 2\exp\left(-t^2\right).$$

The claim follows from the characterization of ψ_2 random variables using their tail decay from Theorem 2.3.5. ■

Remark 3.1.5 *Corollary 3.1.4 leads to a more general version of Khintchine's inequality (Theorem 2.3.9): $\mathcal{Z} = (Z_1, ..., Z_N)$ that has independent, mean-zero, variance 1 coordinates that satisfy $\max\|Z_i\|_{\psi_2} \leq M$ then for $p \geq 2$,*

$$\|\sum_{i=1}^N a_i Z_i\|_{L_2} \leq \|\sum_{i=1}^N a_i Z_i\|_{L_p} \leq cM\sqrt{p}\|\sum_{i=1}^N a_i Z_i\|_{L_2}.$$

Let us present a different proof of Corollary 3.1.4 – a proof that is based on a comparison argument.

Lemma 3.1.6 *Let $X_1, ..., X_N$ be centred, independent random variables and assume that for every $1 \leq i \leq N$ and any $1 \leq p \leq q$, $\|Z_i\|_{L_p} \leq \|X_i\|_{L_p}$. Then for every $a \in \mathbb{R}^N$,*

$$\|\sum_{i=1}^N a_i Z_i\|_{L_q} \leq c_1 L\|\sum_{i=1}^N a_i X_i\|_{L_q}.$$

Proof. Thanks to the natural hierarchy of the L_p norms we may assume that q is an even integer, and by the symmetrization argument of Theorem 3.1.3 we may assume that $Z_1, ..., Z_N$ are symmetric. Therefore,

$$\mathbb{E}(\sum_{i=1}^N a_i Z_i)^q = \mathbb{E}\sum_{\vec{\beta}} c_{\vec{\beta}} \prod_{i=1}^N a_i^{\beta_i} Z_i^{\beta_i} = \sum_{\vec{\beta}}\prod_{i=1}^N a_i^{\beta_i}\mathbb{E}Z_i^{\beta_i},$$

with the sum taken over all choices of $\vec{\beta} = (\beta_1, ..., \beta_N) \in \{0, ..., q\}^N$ that sum to q, and $c_{\vec{\alpha}}$ is the appropriate multinomial coefficient. Since $Z_1, ..., Z_N$ are symmetric, each product does not vanish only when $\beta_1, ..., \beta_N$ are even. Hence,

$$\prod_{i=1}^N a_i^{\beta_i} \mathbb{E} Z_i^{\beta_i} \leq \prod_{i=1}^N a_i^{\beta_i} L^{\beta_i} \mathbb{E} X_i^{\beta_i},$$

and therefore,

$$\sum_{\vec{\beta}} \prod_{i=1}^N a_i^{\beta_i} \mathbb{E} Z_i^{\beta_i} \leq L^q \sum_{\vec{\beta}} \prod_{i=1}^N a_i^{\beta_i} \mathbb{E} X_i^{\beta_i} = L^q \mathbb{E}(\sum_{i=1}^N a_i X_i)^q.$$

■

Proof of Corollary 3.1.4 - version 2. Recall that if g is a standard gaussian random variable then $\|g\|_{L_p} \sim \sqrt{p}$. Thus, if $Z_i \in L_{\psi_2}$ then for every $p \geq 1$,

$$\|Z_i\|_{L_p} \leq c\sqrt{p}\|Z_i\|_{\psi_2} \leq c_1 \|Z_i\|_{\psi_2} \|g\|_{L_p}.$$

Let $g_1, ..., g_N$ be independent standard gaussian random variables and set $X_i = c_1 \|Z_i\|_{\psi_2} g_i$. Applying Lemma 3.1.6, it is evident that for every $p \geq 1$ and every $a \in \mathbb{R}^N$,

$$\| \sum_{i=1}^N a_i Z_i \|_{L_p} \leq \| \sum_{i=1}^N a_i X_i \|_{L_p} = c_1 \| \sum_{i=1}^N a_i \|Z_i\|_{\psi_2} g_i \|_{L_p}.$$

Finally, using the rotation invariance of the standard gaussian vector, $\sum_{i=1}^N a_i \|Z_i\|_{\psi_2} g_i$ has the same distribution as $(\sum_{i=1}^N a_i^2 \|Z_i\|_{\psi_2}^2)^{1/2} g$; therefore,

$$\| \sum_{i=1}^N a_i \|Z_i\|_{\psi_2} g_i \|_{L_p} \leq c_2 \sqrt{p} \left(\sum_{i=1}^N a_i^2 \|Z_i\|_{\psi_2}^2 \right)^{1/2}$$

and the claim follows from the moment characterization of the ψ_2 norm. ■

3.2 Bernstein type inequalities

Let us return to another application of (3.1), leading to *Bernstein type inequalities.*

Lemma 3.2.1 *Let Z be a mean-zero random variable, and assume that there are some M and σ for which, for every integer $p \geq 2$,*

$$\mathbb{E}|Z|^p \leq p! \cdot M^{p-2}\sigma^2.$$

Then for every $0 < \lambda \leq 1/2M$,

$$\mathbb{E}\exp(\lambda Z) \leq 1 + 2\lambda^2\sigma^2 \leq \exp(2\lambda x).$$

Before proving the lemma, let us examine the condition on the growth of the moments of Z. The condition fits two standard situations.

- Let Z be a bounded random variable and set $M = \|Z\|_{L_\infty}$ and $\mathbb{E}Z^2 = \sigma^2$. Observe that $\mathbb{E}|Z|^p \leq \|Z\|_{L_\infty}^{p-2}\mathbb{E}Z^2 = M^{p-2}\sigma^2$, which is far better than the required condition (there is no additional factor of $p!$).

- If $Z \in L_{\psi_1}$ then by Theorem 2.3.5, $\|Z\|_{L_p} \leq cp\|Z\|_{\psi_1}$ for a suitable absolute constant c. Recall that $p^p \leq e^p \cdot p!$, and therefore one may select $\sigma = M = c_1\|Z\|_{\psi_1}$ for an absolute constant c_1.

Proof. Using Taylor's expansion, $\exp(x) = \sum_{i=0}^{\infty} x^p/p!$, and by a standard argument, (e.g. the Monotone Convergence Theorem applied to the positive and negative parts of Z),

$$\mathbb{E}\exp(\lambda Z) = \mathbb{E}\sum_{p=0}^{\infty}\frac{\lambda Z^p}{p!} = 1 + \sum_{p=1}^{\infty}\frac{\lambda^p \cdot \mathbb{E}Z^p}{p!} = (*).$$

Recall that $\mathbb{E}Z = 0$ and that $\mathbb{E}|Z|^p \leq p!M^{p-2}\sigma^2$; therefore, since $\lambda M \leq 1/2$,

$$(*) \leq 1 + \frac{\sigma^2}{M^2}\sum_{p=2}^{\infty}(\lambda M)^p \leq 1 + 2\lambda^2\sigma^2.$$

The fact that $1 + 2\lambda^2\sigma^2 \leq \exp(2\lambda\sigma$ follows because $1 + x^2/2 \leq \exp(x)$. ∎

Theorem 3.2.2 *Let $Z_1, ..., Z_N$ be mean-zero random variables and assume that there is a constant M for which, for every $1 \leq i \leq N$, $\mathbb{E}|Z_i|^p \leq p!M^{p-2}\sigma_i^2$. If $S^2 = \sum_{i=1}^{N}\sigma_i^2$ then for every $t > 0$,*

$$Pr\left(\sum_{i=1}^{N} Z_i > t\right) \leq \exp\left(-\min\left\{\frac{t^2}{8S^2}, \frac{t}{4M}\right\}\right).$$

Proof. Combining (3.1) and Lemma 3.2.1, it follows that for $0 < \lambda \leq 1/2M$,

$$\prod_{i=1}^{N} \mathbb{E}\exp(\lambda Z_i) \leq \prod_{i=1}^{N} \exp(2\lambda^2\sigma_i^2) = \exp(2\lambda^2 S^2).$$

Therefore,

$$Pr\left(\sum_{i=1}^{N} Z_i > t\right) \leq \exp(-\lambda t + 2\lambda^2 S^2) \leq \exp(-\lambda t/2),$$

provided that $2\lambda^2 S^2 \leq \lambda t/2$, i.e., $\lambda \leq t/4S^2$. Let

$$\lambda = \min\left\{\frac{t}{4S^2}, \frac{1}{2M}\right\}$$

and thus

$$Pr\left(\sum_{i=1}^{N} Z_i > t\right) \leq \exp\left(-\min\left\{\frac{t^2}{8S^2}, \frac{t}{4M}\right\}\right).$$

■

Remark 3.2.3 *The constants appearing in Theorem 3.2.2 are not optimal, though this will be of no importance in what follows. From here on we will replace the constants appearing in Theorem 3.2.2 with an unspecified constant c.*

Corollary 3.2.4 *There exists an absolute constant c for which the following holds. Let $Z_1, ..., Z_N$ be independent, mean-zero random variables.*

- *If $\max_{1\leq i\leq N} \|Z_i\|_{L_\infty} \leq M$ and $\sigma_i^2 = \mathbb{E}Z_i^2$, then for every $t > 0$,*

$$Pr\left(\sum_{i=1}^{N} Z_i > t\right) \leq \exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^{N}\sigma_i^2}, \frac{t}{M}\right\}\right).$$

- *If $\max_{1\leq i\leq M} \|Z_i\|_{\psi_1} \leq M$, then*

$$Pr\left(\sum_{i=1}^{N} Z_i > t\right) \leq \exp\left(-c\min\left\{\frac{t^2}{NM^2}, \frac{t}{M}\right\}\right).$$

In particular, if $Z_1, ..., Z_N$ are also identically distributed as a random variable Z then

$$Pr\left(\frac{1}{N}\sum_{i=1}^{N} Z_i > t\right) \leq \exp\left(-cN\min\left\{\frac{t^2}{\|Z\|_{\psi_1}^2}, \frac{t}{\|Z\|_{\psi_1}}\right\}\right).$$

3.3 Sum of squares of ψ_2 random variables

For reasons that have been outlines in the Introduction, the behaviour of a sum of squares of independent random variables

$$\frac{1}{N}\sum_{i=1}^{N} Z_i^2.$$

is of particular interest.

Later, we will give a rather accurate description of this average, highlighting the difference between the upper estimate and the lower one. However, for the time being, let us take a modest step: a two-sided estimate of the from

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} Z_i^2 - \mathbb{E}Z^2\right| > t\right)$$

when the Z_i's are independent copies of a random variable Z that is L-subgaussian; that is, it satisfies $\|Z\|_{\psi_2} \le L\|Z\|_{L_2}$.

Our starting point is the straightforward observation that $\|Z^2\|_{\psi_1} = \|Z\|_{\psi_2}^2$, which follows from the definition of the ψ_α norms. Also, because $\|\cdot\|_{\psi_\alpha}$ are norms for $1 \le \alpha \le 2$, then

$$\|Z^2 - \mathbb{E}Z^2\|_{\psi_1} \le \|Z^2\|_{\psi_1} + \|\mathbb{E}Z^2\|_{\psi_1} \le 2\|Z^2\|_{\psi_1};$$

indeed, by the definition of the ψ_α norm and Jensen's inequality it is evident that for a random variable Y, $\|\mathbb{E}Y\|_{\psi_\alpha} \le \|Y\|_{\psi_\alpha}$.

Thus, the random variable $Z^2 - \mathbb{E}Z^2$ is centred and satisfies that $\|Z^2 - \mathbb{E}Z^2\|_{\psi_1} \le 2\|Z\|_{\psi_2}^2$. Applying the ψ_1 version of Bernstein's inequality,

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} Z_i^2 - \mathbb{E}Z^2\right| > t\right) \le 2\exp\left(-cN\min\left\{\frac{t^2}{\|Z\|_{\psi_2}^4}, \frac{t}{\|Z\|_{\psi_2}^2}\right\}\right).$$

Setting $t = \varepsilon\mathbb{E}Z^2$, it follows that with probability at least

$$1 - 2\exp\left(-cN\min\left\{\varepsilon^2\left(\frac{\|Z\|_{L_2}}{\|Z\|_{\psi_2}}\right)^4, \varepsilon\left(\frac{\|Z\|_{L_2}}{\|Z\|_{\psi_2}}\right)^2\right\}\right), \tag{3.4}$$

$$(1-\varepsilon)\mathbb{E}Z^2 \le \frac{1}{N}\sum_{i=1}^{N} Z_i^2 \le (1+\varepsilon)\mathbb{E}Z^2. \tag{3.5}$$

Now recall that $\|Z\|_{\psi_2}/\|Z\|_{L_2} \leq L$ for some $L \geq 1$. Hence, if $0 < \varepsilon < 1$ then (3.4) becomes

$$1 - 2\exp\left(-cN\min\left\{\frac{\varepsilon^2}{L^4}, \frac{\varepsilon}{L^2}\right\}\right) = 1 - 2\exp(-c_1(L)\varepsilon^2 N), \tag{3.6}$$

and on a high probability event, $N^{-1}\sum_{i=1}^N Z_i^2$ is almost-isometrically equivalent to $\mathbb{E}Z^2$.

As an example, let X be an isotropic, L-subgaussian random vector in $\mathbb{R}^n$, which means that for any $t \in \mathbb{R}^n$, the random variable $Z = \langle t, X\rangle$ is exactly as we described above: it has mean-zero by the symmetry of X, it satisfies $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$ because X is isotropic and $\|\langle t, X\rangle\|_{\psi_2} \leq L\|\langle X, t\rangle\|_{L_2}$ because X is an L-subgaussian random vector.

Let $T \subset \mathbb{R}^n$ be a finite set, let $X_1, ..., X_N$ be independent copies of X, and consider the random matrix

$$\Gamma = \frac{1}{\sqrt{N}}\sum_{i=1}^N \langle X_i, \cdot\rangle e_i.$$

The following result is the celebrated Johnson-Lindenstrauss embedding lemma:

Lemma 3.3.1 *For $L \geq 1$ there exist constants c_1 and c_2 that depend only on L and for which the following holds. If $T \subset \mathbb{R}^n$ is a finite set, $0 < \varepsilon < 1$ and $N \geq c_1\varepsilon^{-2}\log|T|$, then with probability at least $1 - 2\exp(-c_2\varepsilon^2 N)$, for every $x, y \in T$,*

$$(1-\varepsilon)\|x-y\|_2^2 \leq \|\Gamma x - \Gamma y\|_2^2 \leq (1+\varepsilon)\|x-y\|_2^2.$$

In other words, Lemma 3.3.1 implies that the random operator Γ almost preserves distances in T, as long as there is 'enough randomness', in the sense that there are enough independent copies of the random vector X, serving as the rows of the random matrix Γ.

Remark 3.3.2 *As we will explain later, $\log|T|$ is a rather crude measure of complexity for the set T, and one may obtain better estimates on the number of rows required (i.e., the number of sample points/linear measurements needed), as well as a more accurate description on the way an operator like Γ acts on a set.*

Proof. Observe that for every $t_1, t_2 \in T$,

$$\|\Gamma(t_1 - t_2)\|_2^2 = \frac{1}{N}\sum_{i=1}^{N}\left\langle X_i, t_1 - t_2\right\rangle^2$$

and obviously, $\mathbb{E}\left\langle X_i, t_1 - t_2\right\rangle^2 = \|t_1 - t_2\|_2^2$. Hence, for $0 < \varepsilon < 1$, with probability at least $1 - 2\exp(-c(L)\varepsilon^2 N)$,

$$(1+\varepsilon)\|t_1 - t_2\|_2^2 \leq \|\Gamma(t_1 - t_2)\|_2^2 \leq (1+\varepsilon)\|t_1 - t_2\|_2^2. \tag{3.7}$$

The number of different pairs of distinct point in T is at most $|T|^2$, and by the union bound, with probability at least

$$1 - 2|T|^2 \exp(-c(L)\varepsilon^2 N),$$

for every $t_i \neq t_j$, $t_i, t_j \in T$, one has

$$(1+\varepsilon)\|t_i - t_j\|_2^2 \leq \|\Gamma(t_i - t_j)\|_2^2 \leq (1+\varepsilon)\|t_i - t_j\|_2^2. \tag{3.8}$$

If $N \geq c_1(L)\varepsilon^{-2}\log|T|$ then (3.8) holds with probability at least $1-2\exp(-c_2(L)\varepsilon^2 N)$. ■

3.4 Bennett's Inequality

Let us return to the first part of Corollary 3.2.4, which deals with bounded random variables. As can be seen from the estimates on the growth of moments in this case, i.e., that $\mathbb{E}|Z|^p \leq M^{p-2}\sigma^2$ – without the additional factor of $p!$, there should be some room for maneuvering. This observation is at the heart of the following improvement to Bernstein's inequality in the bounded case, known as *Bennett's inequality.*

Before we formulate and prove that inequality, let us improve the estimate on the moment generating function of a bounded random variable.

Lemma 3.4.1 *Let Z be a centred random variable. If $\|Z\|_{L_\infty} \leq M$ then for every $\lambda > 0$,*

$$\mathbb{E}\exp(\lambda Z) \leq 1 + \frac{\mathbb{E}Z^2}{M^2}\left(\exp(\lambda M) - 1 - \lambda M\right).$$

Proof. Using the same argument as in Lemma 3.2.1 and recalling that $\mathbb{E}Z = 0$ and that $\mathbb{E}|Z|^p \leq M^{p-2}\mathbb{E}Z^2$,

$$\mathbb{E}\exp(\lambda Z) \leq 1 + \frac{\mathbb{E}Z^2}{M^2}\sum_{p=2}^{\infty}\frac{\lambda^p M^p}{p!} = 1 + \frac{\mathbb{E}Z^2}{M^2}\left(\exp(\lambda M) - 1 - \lambda M\right),$$

where the final step follows from Taylor's expansion of $\exp(x)$ and the choice $x = \lambda M$. ■

Corollary 3.4.2 *Let $Z_1, ..., Z_N$ be independent mean-zero random variables that satisfy $\max_{1 \le i \le M} \|Z_i\|_{L_\infty} \le M$. Set $S^2 = \sum_{i=1}^N \mathbb{E} Z_i^2$ and put*

$$\Phi(x) = (1+x)\log(1+x) - x.$$

Then, for every $t > 0$,

$$Pr\left(\sum_{i=1}^N Z_i \ge t\right) \le \exp\left(-\frac{S^2}{M^2}\Phi\left(\frac{tM}{S^2}\right)\right).$$

Proof. Combining (3.1) and Lemma 3.4.1, it follows that

$$\exp(-t\lambda)\prod_{i=1}^N \mathbb{E}\exp(\lambda Z_i) \le \exp\left(-\lambda t + \frac{S^2}{M^2}\left(\exp(\lambda M) - 1 - \lambda M\right)\right) = (*).$$

One may optimize the choice of λ and verify that the minimum is attained for

$$\lambda = \frac{1}{M}\log\left(\frac{tM}{S^2} + 1\right),$$

which yields the desired bound. ■

To understand the meaning of Corollary 3.4.2, one should study the behaviour of the function $\Phi(x)$. It is straightforward to verify that for $x \ge 1$, $\Phi(x) \ge (1/2)x\log(1+x)$, and that for $0 < x < 1$, $\Phi(x) \ge x^2/4$. Thus, the tail behaviour changes according to the value of t: when $0 < t \le S^2/M$, the sum exhibits a better-than-gaussian tail decay, and when $t > S^2/M$ the tail behaviour is better than $\sim \exp(-t)$: thanks to the extra logarithmic term, the tail is actually similar to that of a Poisson random variable.

As an example, let $Z_1, ..., Z_N$ be identically distributed according to a bounded, centred random variable Z. Then $S^2 = N\mathbb{E}Z^2 = N\sigma^2$ and for a suitable absolute constant c,

$$Pr\left(\frac{1}{N}\sum_{i=1}^N Z_i > t\right) \le \exp\left(-\frac{Nt}{M} - \left(\frac{Nt}{M} + \frac{N\sigma^2}{M^2}\right)\log\left(1 + \frac{Mt}{\sigma^2}\right)\right).$$

If $0 < t \le N\sigma^2/M$ then the tail is smaller than

$$\exp(-t^2/2N\sigma^2), \tag{3.9}$$

while for every $t > 0$, it is smaller than

$$\exp\left(-\frac{t}{M}\left(\log\left(1+\frac{Mt}{N\sigma^2}\right)\right)\right) \tag{3.10}$$

One very useful application of Bennett's inequality is for the sum of independent selectors – which are simply $\{0,1\}$-valued random variables. Indeed, let $(\delta_i)_{i=1}^N$ be independent, taking values in $\{0,1\}$ and assume that for every $1 \le i \le N$, $\mathbb{E}\delta_i = \delta$. Set $Z_i = \delta_i - \delta$ and note that $\mathbb{E}Z_i = 0$, $\|Z_i\|_\infty = 1$ and $\mathbb{E}Z_i^2 = \delta - \delta^2 \ge \delta/4$ provided that $\delta \le 3/4$. In that case, applying (3.9) and (3.10) to the Z_i's and $-Z_i$'s, the following is evident: for $t \le N\delta/4$,

$$Pr\left(|\sum_{i=1}^N(\delta_i-\delta)| > t\right) \le 2\exp\left(-c\frac{t^2}{N\delta}\right) \tag{3.11}$$

and for $t \ge N\delta/4$,

$$Pr\left(|\sum_{i=1}^N(\delta_i-\delta)| > t\right) \le 2\exp\left(-ct\left(\log\left(1+\frac{t}{N\delta}\right)-1\right)\right). \tag{3.12}$$

Thus, with probability at least $1-2\exp(-c_1\delta N)$,

$$\frac{1}{2}\delta N \le |\{i : \delta_i = 1\}| \le \frac{3}{2}\delta N$$

and for $u \ge 2$, with probability at least $1-2\exp(-c_1\delta N u\log u)$,

$$|\{i : \delta_i = 1\}| \le u\delta N.$$