1. Subgaussian tails

<1> **Definition.** *Say that a random variable X has a **subgaussian distribution** with scale factor $\sigma < \infty$ if $\mathbb{P}\exp(tX) \le \exp(\sigma^2 t^2/2)$ for all real t.*

For example, if X is distributed $N(0, \sigma^2)$ then it is subgaussian.

<2> **Example.** Suppose X is a bounded random variable with a symmmetric distribution. That is, $|X| \le M$ for some constant M and $-X$ has the same distribution as X. Then

$$\mathbb{P}\exp(tX) = 1 + \sum_{k\in\mathbb{N}} \frac{t^k \mathbb{P}X^k}{k!}$$

By symmetry, $\mathbb{P}X^k = 0$ for each odd k. For even k, bound $\mathbb{P}X^k$ by M^k, leaving

$$\mathbb{P}\exp(tX) = 1 + \sum_{k\in\mathbb{N}} \frac{t^{2k} M^{2k}}{(2k)!} \le \exp(M^2 t^2/2)$$

□ because $(2k)! \ge 2^k k!$ for each k in $\mathbb{N}$.

The argument for bounding the maximum of normal random variables carries over to subgaussians.

<3> **Theorem.** *Suppose $X_1, \dots, X_n$ are subgaussian with scale factors bounded by a constant σ. Then $\mathbb{P}\max_{i\le n} |X_i| \le \frac{3}{2}\sigma\sqrt{1 + \log(2n)}$.*

Proof. For each $t > 0$,

$$\exp(t\mathbb{P}\max_{i\le n} |X_i|) \le \mathbb{P}\max_{i\le n} \exp(t|X_i|) \le \sum_{i\le n} \left(\mathbb{P}e^{tX} + \mathbb{P}e^{-tX}\right) \le 2n\exp(\tfrac{1}{2}\sigma^2 t^2)$$

□ Choose $t = \log(2n)/\sigma$.

In fact, we could improve the inequality to give similar bounds for various $\mathcal{L}^p$ norms of $\max_{i\le n} |X_i|$ by choosing slightly different convex functions instead of $x \mapsto \exp(tx)$. I won't derive these bounds explicitly because there is an even better inequality obtainable from another characterization of subgaussianity.

<4> **Theorem.** *Suppose $\mathbb{P}X = 0$. Then X is subgaussian if and only if there exists a finite constant C for which $\mathbb{P}\exp(X^2/C^2) < \infty$.*

Proof. If $\mathbb{P}\exp(tX) \le \exp(\sigma^2 t^2/2)$ for all real t then

$$\begin{aligned}
\mathbb{P}\exp(X^2/4\sigma^2) - 1 &= \mathbb{P}\int_0^\infty \{X^2/4\sigma^2 \ge t \ge 0\} e^t \, dt \\
&\le \int_0^\infty \mathbb{P}\exp\left(\frac{|X|\sqrt{t}}{\sigma} - t\right) dt \\
&\le \int_0^\infty \mathbb{P}\left(\exp(X\sqrt{t}/\sigma) + \exp(-X\sqrt{t}/\sigma)\right) e^{-t} \, dt \\
&\le \int_0^\infty 2e^{-t/2} \, dt < \infty.
\end{aligned}$$

Conversely, if $\mathbb{P}\exp(X^2/C^2) = D < \infty$ then, from the inequality $ab \le (a^2 + b^2)/2$, we get

$$\mathbb{P}\exp(tX) \le \mathbb{P}\exp\left(\frac{X^2}{C^2} + \frac{C^2 t^2}{4}\right) = D\exp(C^2 t^2/4).$$

This bound is not quite what we need for subgaussianity. If we bound t away from zero we can eliminate the D: if $D \le \exp(MC^2\delta^2)$ for some constant M then

$$\mathbb{P}\exp(tX) \le \exp((M+1)C^2t^2) \qquad \text{for } |t| \ge \delta.$$

If δ is small enough, the Taylor expansion gives, for small enough δ,

$$\begin{aligned}\mathbb{P}\exp(tX) &= 1 + t\mathbb{P}X + \tfrac{1}{2}t^2\mathbb{P}X^2 + o(t^2) \\ &\le \exp\left(\tfrac{1}{2}t^2(1+\mathbb{P}X^2)\right) \qquad \text{when } |t| \le \delta.\end{aligned}$$

□ The subgaussianity bound follows.

Subgaussian random variables can also be characterized by an exponential tail bound. Take $t = x/\sigma^2$ in the inequality

$$\mathbb{P}\{X \ge x\} \le \exp(-tx)\mathbb{P}\exp(tX) \le \exp(-tx + \sigma^2t^2/2)$$

to deduce that

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

Replace X by $-X$, which is also subgaussian, then add, to derive the analogous two-sided bound. Conversely, if $\mathbb{P}\{|X| \ge x\} \le C\exp(-x^2/2\sigma^2)$ then

$$\begin{aligned}\mathbb{P}\exp(X^2/9\sigma^2) - 1 &= \mathbb{P}\int_0^\infty \{X^2 \ge 9\sigma^2 t \ge 0\}e^t\,dt \\ &= \int_0^\infty \mathbb{P}\{|X| \ge 3\sigma^2\sqrt{t}\}e^t\,dt \\ &\le \int_0^\infty C\exp(-9t/2 + t)\,dt < \infty\end{aligned}$$

which, via Theorem <4>, gives subgaussianity.

2. Orlicz norms

The convexity argument used to prove Theorem <3> also works for higher moments.

$$\left(\mathbb{P}\max_{i\le N}|X_i|\right)^p \le \mathbb{P}\max_{i\le N}|X_i|^p \le \sum_{i\le N}\mathbb{P}|X_i|^p \le N\max_{i\le N}\mathbb{P}|X_i|^p.$$

Thus

<5>
$$\mathbb{P}\max_{i\le N}|X_i| \le \left\|\max_{i\le N}|X_i|\right\|_p \le N^{1/p}\max_{i\le N}\|X_i\|_p \qquad \text{for } p \ge 1.$$

More generally, if ψ is a nonnegative, convex, strictly increasing function on $\mathbb{R}^+$, then, for each $\sigma > 0$,

$$\begin{aligned}\psi\left(\mathbb{P}\max_{i\le N}\frac{|X_i|}{\sigma}\right) &\le \mathbb{P}\max_{i\le N}\psi\left(\frac{|X_i|}{\sigma}\right) \\ &\le \sum_{i\le N}\mathbb{P}\psi\left(\frac{|X_i|}{\sigma}\right) \\ &\le N\max_{i\le N}\mathbb{P}\psi\left(\frac{|X_i|}{\sigma}\right).\end{aligned}$$

If σ is such that $\mathbb{P}\psi(|X_i|/\sigma) \le 1$ for each i then we have

$$\mathbb{P}\max_{i\le N}|X_i| \le \sigma\psi^{-1}(N).$$

<6> **Definition.** *An **Orlicz function** is a convex, increasing function ψ on $\mathbb{R}^+$ with $0 \le \psi(0) < 1$. Define the **Orlicz norm** $\|X\|_\psi$ (seminorm actually, unless one identifies random variables that are almost everywhere equal) by*

Most authors actually require $\psi(0) = 0$

$$\|X\|_\psi = \inf\{c > 0 : \mathbb{P}\psi(|X|/c) \le 1\},$$

with the understanding that $\|X\|_\psi = \infty$ if the infimum runs over an empty set.

It is not hard to show (Pollard 2001, Problems 2.22 through 2.24) that $\|X\|_\psi < \infty$ if and only if $\mathbb{P}\psi(|X|/C) < \infty$ for at least one finite constant C. The infimum defining $\|X\|_\psi$ is achieved when the norm is finite.

<7> **Example.** Let $\psi(x) = \exp(x^2) - 1$. Then $\|X\|_\psi < \infty$ if and only if $X - \mathbb{P}X$ is subgaussian. □

Notice that a bound on an Orlicz norm, $\|X\|_\psi \le \sigma$, automatically gives a tail bound,

$$\mathbb{P}\{|X| \ge x\} \le \mathbb{P}\psi(|X|/\sigma)/\psi(x/\sigma) \le 1/\psi(x/\sigma) \qquad \text{for } x \ge 0.$$

For example, if $\psi(x) = \frac{1}{2}\exp(x^2)$ then we get a subgaussian tail bound.

Sometimes it is possible to find δ such that $\mathbb{P}\psi(|X|/\delta) \le K$, for a constant $K > 1$. It then follows from convexity of ψ that

<8> $$\|X\|_\psi \le \delta/\theta \qquad \text{where } \theta = \frac{1 - \psi(0)}{K - \psi(0)},$$

because

$$\mathbb{P}\psi\left(\theta|X|/\delta\right) \le \theta\mathbb{P}\psi\left(|X|/\delta\right) + (1-\theta)\psi(0) \le \theta K + (1-\theta)\psi(0) = 1.$$

<9> **Example.** (Compare with page 96 of van der Vaart & Wellner (1996).) Let ψ be an Orlicz function (such as $\exp(x^2) - 1$, as in Problem [1]) for which there exists a finite constant C_0 such that

$$\psi(\alpha)\psi(\beta) \le \psi(C_0\alpha\beta) \qquad \text{for } \psi(\alpha) \wedge \psi(\beta) \ge 1.$$

Then

<10> $$\left\| \max_{i \le N} |X_i| \right\|_\psi \le C\psi^{-1}(N) \max_{i \le N} \|X_i\|_\psi \qquad \text{where } C := \frac{2 - \psi(0)}{1 - \psi(0)} C_0$$

To prove the assertion, define $D = C_0\psi^{-1}(N)$ and $\delta = \max_{i \le N} \|X_i\|_\psi$. Notice that $\psi(D/C_0) = N \ge 1$. When $\psi(\max_i |X_i|/D\delta) \ge 1$,

$$\psi\left(\frac{\max_i |X_i|}{D\delta}\right)\psi\left(\frac{D}{C_0}\right) \le \psi\left(\frac{\max_i |X_i|}{\delta}\right) \le \sum_i \psi\left(\frac{|X_i|}{\delta}\right).$$

That is,

$$\psi\left(\frac{\max_i |X_i|}{D\delta}\right) \le \min\left(1, N^{-1}\sum_i \psi\left(\frac{|X_i|}{\delta}\right)\right)$$

Take expectations.

$$\mathbb{P}\psi\left(\frac{\max_i |X_i|}{D\delta}\right) \le 1 + N^{-1}\sum_i \mathbb{P}\psi\left(\frac{|X_i|}{\delta}\right) \le 2.$$

□ Invoke inequality <8>.

Finally, notice that if $\|X\|_\psi = \sigma$ for $\psi(x) = \exp(x^2) - 1$ then

$$\frac{\mathbb{P}|X|^{2p}}{\sigma^{2p}} \le p!\mathbb{P}\exp(X^2/\sigma^2) \le 2p!.$$

A bound on the Orlicz norm, for this particular ψ, gives a bound on moments of all orders.

<11> **Example.** For each event A with $\mathbb{P}A > 0$, write $\mathbb{P}_A$ for the conditional expectation given A. Suppose $\|X\|_\psi \le \delta < \infty$. From Jensen's inequality and the definition of the Orlicz norm we get

$$\psi(\mathbb{P}_A|X|/\delta) \le \mathbb{P}_A\psi(|X|/\delta) = \frac{\mathbb{P}\psi(|X|/\delta)A}{\mathbb{P}A} \le \frac{1}{\mathbb{P}A},$$

from which it follows that

<12> $$\mathbb{P}_A|X| \le \|X\|_\psi \psi^{-1}(1/\mathbb{P}A).$$

With cunning choices of A, this inequality will deliver a useful maximal inequality for finite collections of random variables, namely,

<13> $$\mathbb{P}_A \max_{i\le N}|X_i| \le \delta\psi^{-1}(N/\mathbb{P}A) \qquad \text{if} \quad \max\nolimits_{i\le N}\|X_i\|_\psi \le \delta.$$

Indeed, if $A_1,\dots,A_N$ denotes a partition of A into subsets, such that $|X_i|$ is the largest of the $|X_j|$ on the set A_i, then

$$\mathbb{P}_A \max_{i\le N}|X_i| = \sum_i \mathbb{P}_A|X_i|A_i = \sum_i \frac{\mathbb{P}A_i}{\mathbb{P}A}\mathbb{P}_{A_i}|X_i|.$$

Inequality <12> and concavity of the function ψ^{-1} bound the last sum by

$$\sum_i \frac{\mathbb{P}A_i}{\mathbb{P}A}\delta\psi^{-1}\left(\frac{1}{\mathbb{P}A_i}\right) \le \delta\psi^{-1}\left(\sum_i \frac{\mathbb{P}A_i}{\mathbb{P}A}\frac{1}{\mathbb{P}A_i}\right) = \delta\psi^{-1}\left(\frac{N}{\mathbb{P}A}\right).$$

□

The bound <13> will turn out to be much more powerful than one might at first glance suspect. If we choose $A = \{\max_{i\le N}|X_i| \ge \epsilon\}$ then we get lower bound for $1/\mathbb{P}A$. The full power of this trick will appear in the Chapter on chaining.

3. Problems

[1] Show that $(\exp(x^2)-1)(\exp(y^2)-1) \le \exp(2x^2y^2)-1$ for $x \wedge y \ge 1$.

[2] Suppose X has a symmetric distribution. Show that it is subgaussian if and only if there exists some constant c for which $\|X\|_k \le c\sqrt{k}$ for each k in $\mathbb{N}$. Hints: Note that $\|X\|_k$ is an increasing function of k. For k even, try to show that

$$\frac{\|X\|_k^k}{k!} \le \inf_t \frac{\mathbb{P}\exp(tX)}{t^k}$$

[3] Let X and Y be identically distributed random variables with $\mathbb{P}X = \mathbb{P}Y = 0$.

(i) Let H be a convex function. [Any other regularity conditions?] Show that $\mathbb{P}H(X) = \mathbb{P}H(X - \mathbb{P}Y) \le \mathbb{P}H(X-Y)$.

(ii) Show that $\|X\|_\psi \le \|X-Y\|_\psi \le 2\|X\|_\psi$ for each Orlicz function ψ.

(iii) Generalize the result from Problem [2]: Show that the moment characterization of subgaussianity still holds if replace the symmetry assumption on X by the assumption that $\mathbb{P}X = 0$.

4. Notes

Acknowledge Ledoux & Talagrand (1991) for several of the ideas used in this Chapter, including Example <11> Cite Aad van der Vaart (personal communication, or van der Vaart & Wellner 1996) for improvement on the method used in Pollard (1990, Section 3).

Who first got the characterization in Problems [2] and [3]? I got it from a sharper result in Lugosi (2003, Section 2), but it must be older.

Give some history of earlier work: Dudley, Pisier?

References

Ledoux, M. & Talagrand, M. (1991), *Probability in Banach Spaces: Isoperimetry and Processes*, Springer, New York.

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

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, Vol. 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*, Institute of Mathematical Statistics, Hayward, CA.

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

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