High Dimensional Probability

Jacob Denson

April 21, 2019

Table Of Contents

Chapter 1

Introduction

In these notes, we study the problems and phenomena that arise when studying random phenomena in high dimensional spaces. These phenomena arise from numerous situations, including when studying large random graphs, large random matrices, or doing statistics with large data sizes. A few informal principles guide our exploration of the subject

- (Concentration): The law of large numbers gives the asymptotic result that if $\{X_k\}$ is a sequence of i.i.d random variables, then

$$\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbf{E}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) \to 0$$

almost surely. But in many cases in analysis we need non-asymptotic results, which replace this limit theorem with precise *deviation bounds* which provide upper bounds on

$$\mathbf{P}\left[\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbf{E}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) \geqslant t\right]$$

which decay fast with respect to t. For the general class of *subgaussian* random variables, one can obtain very fast decaying bounds on this limit process. Similar results hold if the X_k are only weakly dependant on one another. More generally, if $f : \mathbf{R}^n \to \mathbf{R}$ is not 'too sensitive' with respect to any of it's coordinates, then we should be able to obtain sharp bounds on

$$\mathbf{P}\left(|f(X_1,\ldots,X_n) - \mathbf{E}f(X_1,\ldots,X_n)| \geqslant t\right)$$

when the X_k are subgaussian. We note that concentration estimates the fluctuations of f, but not it's magnitude. We require other tools to compute $\mathbf{E}f(X_1,\dots,X_n)$. Though concentration holds for very general functions f, we cannot hope to find general methods of calculating $\mathbf{E}f(X_1,\dots,X_n)$ for general functions.

- (Controlling Suprema) It is often natural to control the expected magnitude of a family of random variables, i.e. we wish to control $\mathbf{E}\sup_{t\in T} X_t$, where t is some index set. A natural principle is that if the random field $\{X_t\}$ is 'sufficiently continuous', the magnitude is controlled by the 'complexity' of the index set T.

- (Universality) The central limit theorem says that for large n, if the X_i are independant then the CDF of the random variable

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(X_k - \mathbf{E}(X_k))$$

  behaves like the CDF of a Gaussian distribution. The fact that this is true irrespective of the distribution of the components X_k is known as *universality*. In general, universality refers to the features of the components of a distribution becoming irrelevant when n is large. Another way to state this is that if f is a 'sufficiently smooth function' and n is large, then $\mathbf{E}(f(X_1,\dots,X_n))$ is insensitive to the distributions of the X_k. This means that high dimensional phenomena we study are robust to the precise details of the model we approximate them with. Universality is very useful because it allows us to replace X_k with very well behaved distributions, i.e. Gaussian distributions. We note that universality is not necessarily related to a Gaussian distribution, but Gaussian distributions do tend to show up with high dimensional phenomena.

- (Sharp Transitions) The last understood principle is given by sharp transitions. In high dimensional models, as we vary parameters there tends to be abrupt changes in the qualitative phenomena. As an example, if $\{X_k\}$ is a sequence of $\{0,1\}$ valued Bernoulli random variables with parameter p, and Z_n is the majority function of $X_1,\dots,X_n$, then $\mathbf{E}(Z_n)\to 0$ if $p<1/2$, and $\mathbf{E}(Z_n)\to 1$ if $p>1/2$. As $n\to\infty$, there is an abrupt change in the behaviour of the Z_n as we vary p.

In some cases, this can be explained by concentration phenomena. But this occurs even in cases that cannot be explained using concentration. In general, if $f(E_1, \ldots, E_n)$ is 'sufficiently symmetric' and 'sufficiently monotone', with $\{E_k\}$ events depending on a probability p, then $f(E_1, \ldots, E_n)$ undergoes a 'sharp transition'.

Chapter 2

Concentration In High Dimensional Spaces

A basic instance of high dimensional probability occurs when studying random vectors $X \in \mathbf{R}^n$, where n is a very large number. The exponential increase in room in high dimensions leads to concentration of the vector in unlikely places. If X is a random standard Gaussian vector in $\mathbf{R}^n$, then

$$\mathbf{E}\,|X|^2 = \sum \mathbf{E} X_i^2 = n$$

Since $|X|$ is formed from n independant random variables, each having equal contribution to the magnitude of $|X|$, we could guess that the law of large numbers result would imply $|X|$ to be close to $\sqrt{n}$ if n is large. And this is certainly the case. Indeed, since $|X|^2$ is a sum of independant subexponential random variables, and it has mean n and standard deviation $O(\sqrt{n})$, then we should expect $|X|^2 = n + O(\sqrt{n})$ with high probability, and if this is true then

$$|X| = \sqrt{n + O\left(\sqrt{n}\right)} = \sqrt{n} + O(1).$$

Thus $|X|$ should deviate from $\sqrt{n}$ by a constant distance, independant of n. This is precisely the content of the next theorem.

Theorem 2.1. *Let X be a random vector in $\mathbf{R}^n$ with independant coordinates, and with $\|X_i\|_{\psi_2} \leqslant K$ and $\mathbf{E}(X_i^2) = 1$ for each i. Then $\||X| - \sqrt{n}\|_{\psi_2} \lesssim K^2$.*

Proof. In this proof we assume $K \geqslant 1$. Note that since

$$1 = \mathbf{E}(X_i^2) \lesssim \|X_i\|_{\psi_2}^2 \leqslant K^2,$$

we know $K \gtrsim 1$. Thus if $K \leqslant 1$, we can apply the theorem with $K = 1$ to obtain that

$$\| |X| - \sqrt{n} \|_{\psi_2} \lesssim 1 \lesssim K^2,$$

so the theorem is obtained for free in this case.

The random variables X_i^2 are subexponential, with

$$\|X_i^2\|_{\psi_1} = \|X_i\|_{\psi_2}^2 \leqslant K^2.$$

By centering, we know $\|X_i^2 - 1\|_{\psi_1} \lesssim \|X_i^2\|_{\psi_1}$. Thus we can apply Bernstein's inequality, finding a universal constant c such that

$$\begin{aligned}
\mathbf{P}\left(\left||X|^2 - n\right| \geqslant t\right) &\leqslant 2\exp\left(-c \cdot \min\left(\frac{t^2}{\sum \|X_i\|_{\psi_2}^4}, \frac{t}{\max \|X_i\|_{\psi_2}^2}\right)\right) \\
&\leqslant 2\exp\left(-c \cdot \min\left(t^2/K^4, t/K^2\right)\right) \\
&\leqslant 2\exp(-c/K^4 \cdot \min(t^2, t)).
\end{aligned}$$

Here we used the fact that $K \geqslant 1$, so that $1/K^2 \geqslant 1/K^4$. The inequality above is a good concentration bound for $|X|^2$, and we now need to turn it into a concentration bound for $|X|$.

If $u = \min(t, t^2)$, then $t = \max(u, u^{1/2})$. Thus we have shown

$$\mathbf{P}\left(\left||X|^2 - n\right| \geqslant \max(u, u^{1/2})\right) \leqslant 2\exp\left(-cu/K^4\right).$$

We note that

$$\begin{aligned}
\left||X|^2 - n\right| &= \left||X| - n^{1/2}\right| \left||X| + n^{1/2}\right| \\
&\geqslant \left||X| - n^{1/2}\right| \max\left(n^{1/2}, \left||X| - n^{1/2}\right|\right) \\
&\geqslant \max\left(\left||X| - n^{1/2}\right|, \left||X| - n^{1/2}\right|^2\right).
\end{aligned}$$

Thus

$$\mathbf{P}\left(\left||X| - n^{1/2}\right| \geqslant u\right) \leqslant \mathbf{P}\left(\left||X|^2 - n\right| \geqslant \max(u, u^2)\right) \leqslant 2\exp(-cu^2/K^4).$$

Since this holds for all $u \geqslant 0$, we have shown $\| |X| - n^{1/2} \|_{\psi_2} \lesssim K^2$. □

Corollary 2.2. *If X is as in Theorem 2.1, then*

$$\mathbf{E}\,|X| = n^{1/2} + O(K^2) \quad \textit{and} \quad \mathbf{V}\,|X| \lesssim K^4.$$

Proof. To prove the expectation bound, we first apply centering to the random variable $|X| - n^{1/2}$. Thus we know

$$\||X| - \mathbf{E}\,|X|\|_{\psi_2} \lesssim \||X| - n^{1/2}\|_{\psi_2} \lesssim K^2.$$

Thus the triangle inequality implies that

$$\|\mathbf{E}|X| - n^{1/2}\|_{\psi_2} \leqslant \|\mathbf{E}|X| - |X|\|_{\psi_2} + \||X| - n^{1/2}\|_{\psi_2} \lesssim K^2,$$

and the left hand side is proportional to $|\mathbf{E}|X| - n^{1/2}|$. The variance bound then follows easily, because

$$\begin{aligned}\mathbf{V}\,|X| &= \mathbf{V}(|X| - n^{1/2}) \\ &= \mathbf{E}((|X| - n^{1/2})^2) - \mathbf{E}(|X| - n^{1/2})^2 \\ &\lesssim \||X| - n^{1/2}\|_{\psi_2}^2 - O(K^4) \lesssim K^4.\end{aligned}$$

□

2.1 Isotropic Vectors

We recall some basic facts about random vectors. Given a centrally distributed random vector X in $\mathbf{R}^n$, we define the $n \times n$ covariance matrix $\Sigma(X)$ by the formula $\Sigma(X)_{ij} = \mathbf{E}(X_i X_j)$. This is a symmetric, positive semi-definite matrix, since

$$x^T \Sigma(X) x = \mathbf{E}((x \cdot X)^2) \geqslant 0.$$

The spectral decomposition theorem implies that there is a basis of normalized eigenvectors $u_1, \ldots, u_n$ for $\Sigma(X)$, with non-negative eigenvalues $\lambda_1, \ldots, \lambda_n$. We assume that they have been arrange so that the eigenvalues are placed in decreasing order. If $Y_i = u_i \cdot X$, then this means $\mathbf{E}(Y_i^2) = \lambda_i$, and $\mathbf{E}(Y_i Y_j) = 0$ if $i \neq j$. Thus we can always rotate a distribution so it's coordinates are independant 'up to second order'.

We say a random vector X is *isotropic* if $\Sigma(X)$ is the identity matrix. This is equivalent to saying that for each vector x,

$$\mathbf{E}(X \cdot x)^2 = x^T \sigma(X) x = |x|^2.$$

Thus the vector X is extended evenly in all directions. We can often reduce the analysis of random vectors to the centred, isotropic setting. If a random vector X has mean μ, and invertible covariance matrix Σ, then $\Sigma^{-1/2}(X-\mu)$ is a centered, isotropic random variable. If Σ is degenerate, then X almost surely lies on a lower dimensional subspace, and then we can reduce our analysis to this lower dimensional subspace.

Lemma 2.3. *If X is isotropic, then $\mathbf{E}|X|^2 = n$. More generally, if X and Y are independant and isotropic, then $\mathbf{E}(X \cdot Y)^2 = n$.*

Proof. We write

$$\begin{aligned}\mathbf{E}|X|^2 = \mathbf{E}X^T X &= \mathbf{E}\left(\operatorname{tr}(X^T X)\right) \\ &= \mathbf{E}\left(\operatorname{tr}(XX^T)\right) = \operatorname{tr}(\mathbf{E}(XX^T)) = \operatorname{tr}(I) = n\end{aligned}$$

Next, given Y, we find

$$\mathbf{E}((X \cdot Y)^2|Y) = \sum Y_i Y_j \mathbf{E}(X_i X_j|Y) = \sum Y_i Y_j \mathbf{E}(X_i X_j) = \sum Y_i^2 = |Y|^2$$

But this means that

$$\mathbf{E}((X \cdot Y)^2) = \mathbf{E}(\mathbf{E}((X \cdot Y)^2|Y)) = \mathbf{E}|Y|^2 = n. \qquad \square$$

Remark. Since

$$\mathbf{E}(X \cdot Y) = \sum \mathbf{E}(X_i)\,\mathbf{E}(Y_i) = 0 \quad \text{and} \quad \mathbf{V}(X \cdot Y) = \mathbf{E}((X \cdot Y)^2) = n,$$

we can expect that $X \cdot Y = O(n^{1/2})$ with high probability. But this means that with high probability

$$\frac{X \cdot Y}{|X||Y|} = \frac{O(n^{1/2})}{O(n^{1/2})O(n^{1/2})} = O\left(1/n^{1/2}\right).$$

Thus independant isotropic vectors in high dimensional spaces tend to lie almost at right angles to one another. This is very different from in two dimensions, where two independant unit vectors chosen uniformly at random on the unit circle are on average 45° from one another.

Here are several examples of isotropic random vectors.

Example. *Let X be a vector chosen uniformly at random on the sphere of radius $\sqrt{n}$ in $\mathbf{R}^n$. For $i \neq j$, (X_i, X_j) is identically distributed to $(X_i, -X_j)$, so $\mathbf{E}(X_i X_j) = -\mathbf{E}(X_i X_j)$. This implies $\mathbf{E}(X_i X_j) = 0$. Since $\mathbf{E}|X|^2 = n$, and the $\mathbf{E}|X_i|^2$ are independant of i, this implies $\mathbf{E}|X_i|^2 = 1$.*

It is good to remember that the coordinates of an isotropic vector need not be independant. A uniformly random point X on the radius $\sqrt{n}$ sphere need not be independant, because the points must satisfy $X_1^2 + \cdots + X_n^2 = 1$.

Example. *Let X be a random vector with independent, symmetric Bernoulli distributions as coordinates. Then $\mathbf{E}(X_i X_j) = \mathbf{E}(X_i)\mathbf{E}(X_j) = 0$ for $i \neq j$, and $\mathbf{E}(X_i^2) = 1$. More generally, any random vector with independant, mean zero, unit variance coordinates are isotropic. This includes the example of a random vector $X \sim N(0, I_n)$ with the standard, normal distribution.*

Example. *For the most extreme example of a discrete isotropic distribution, we can take a vector uniformly randomly from $\{\sqrt{n}e_1, \ldots, \sqrt{n}e_n\}$. Then X_i^2 is $\{0, 1\}$ valued, with $\mathbf{P}(X_i^2 = n) = 1/n$. This gives $\mathbf{E}(X_i^2) = 1$. On the other hand, $X_i X_j = 0$ for $i \neq j$, so $\mathbf{E}(X_i X_j) = 0$. Note that the mean of $\mathbf{E}(X_i)$ is non-zero; it is actually equal to $1/\sqrt{n}$.*

We obtain a family of discrete isotropic random vectors by considering uniformly distributions over discrete families of vectors used most notably in signal processing, known as **frames**. A **frame** is a set $\{v_1, \ldots, v_m\}$ which obeys the approximate $A|x| \leqslant \sum (v_i \cdot x)^2 \leqslant B|x|^2$ for all vectors x. if $A = B$, the frame is called **tight**. A frame is tight if and only if $\sum v_i^T v_i = AI_n$.

Example. *An example of a tight frame which isn't an orthonormal basis is the 'Mercedes Benz' frame, three uniformly separated points on a unit circle. If*

$$v_1 = (0, 1) \quad v_2 = \left(-\frac{1+\sqrt{3}}{2\sqrt{2}}, -\frac{\sqrt{3}-1}{2\sqrt{2}}\right) \quad v_3 = \left(\frac{1+\sqrt{3}}{2\sqrt{2}}, -\frac{\sqrt{3}-1}{2\sqrt{2}}\right)$$

then

$$v_1^T v_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad v_2^T v_2 = \begin{pmatrix} \frac{2+\sqrt{3}}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{2-\sqrt{3}}{4} \end{pmatrix} \quad v_3^T v_3 = \begin{pmatrix} \frac{2+\sqrt{3}}{4} & \frac{-1}{4} \\ \frac{-1}{4} & \frac{2-\sqrt{3}}{4} \end{pmatrix}$$

$$v_1 = (0, 1) \quad v_2 = (-\sqrt{3/4}, -\sqrt{1/4}) \quad v_3 = (\sqrt{3/4}, -\sqrt{1/4})$$

and this means that

$$(x,v_1)^2+(x,v_2)^2+(x,v_3)^2 = x_2^2+(1/4)(\sqrt{3}x_1+x_2)^2+(1/4)(\sqrt{3}x_1-x_2)^2$$
$$= 3/2|x|^2$$

So the frame is tight with $A = B = 3/2$.

Theorem 2.4. *If $\{u_1,\dots,u_m\}$ is tight, then a uniformly random choice of a frame element, scaled by $(m/A)^{1/2}$ is isotropic. Conversely, if X is isotropic, and takes only finitely many values $\{u_1,\dots,u_n\}$ with $\mathbf{P}(X = u_i) = p_i$, then $p_i^{1/2}u_i$ is a tight frame with $A = 1$.*

Proof. If $X = (m/A)^{1/2}\cdot \mathrm{Uni}(u_1,\dots,u_n)$, then

$$\mathbf{E}(X\cdot x)^2 = \frac{(m/A)\sum(u_i\cdot x)^2}{n} = (1/A)A|x|^2 = |x|^2$$

which means precisely that the frame is isotropic. To prove the converse, we find that

$$\sum(p_i^{1/2}u_i\cdot x)^2 = \sum p_i(u_i\cdot x)^2 = \mathbf{E}((X\cdot x)^2) = |x|^2$$

which is precisely the definition of a tight frame. □

Example. *Let K be a convex set with nonempty interior in $\mathbf{R}^n$. If X is a uniformly chosen point in K, which by translation we may assume to have mean zero, and covariance matrix Σ, then Σ is positive definite, because if Σ has a zero eigenvalue, there would be a vector a such that X is almost surely orthogonal to a, which is impossible since K has non-empty interior. Thus $\Sigma^{-1/2}X$ is isotropic, and thus $\Sigma^{-1/2}K$ can be seen as a convex set 'uniformly extended in each direction'. This is often useful as a preprocessing step before applying algorithms on convex sets.*

2.2 Subgaussian Vectors

We say a random vector $X \in \mathbf{R}^n$ is *subgaussian* if $X\cdot x$ is subgaussian for all x, or equivalently, if all coordinates of X are subgaussian. Then there is a smallest value $\|X\|_{\psi_2} < \infty$ such that $\|X\cdot x\|_{\psi_2} \leqslant \|X\|_{\psi_2}|x|$. We think of $\|X\|_{\psi_2}$ as a generalization of the subgaussian norm to random vectors.

Example. *Suppose X has independant, subgaussian coordinate $X_1, \dots, X_n$. If $x = (x_1, \dots, x_n)$ satisfies $|x| = 1$, then*

$$\|X \cdot x\|_{\psi_2}^2 = \left\| \sum x_i X_i \right\|_{\psi_2}^2 \lesssim \sum x_i^2 \|X_i\|_{\psi_2}^2 \leqslant \max \|X_i\|_{\psi_2}^2$$

Thus $\|X\|_{\psi_2} \lesssim \max \|X_i\|_{\psi_2}$. On the other hand, we know $\|X\|_{\psi_2} \geqslant \max \|X_i\|_{\psi_2}$, so in this case the subgaussian norm of the vector is essentially equal to the maximum subgaussian norm of it's coordinates. On the other hand, even if the coordinates of a random vector X are individually subgaussian, if independence is not satisfied then we may have $\|X\|_{\psi_2} \gg \max \|X_i\|_{\psi_2}$. For instance, if $X_i = X_j$ for all $i = j$, then

$$\|X\|_{\psi_2} = \sqrt{n} \cdot \max \|X_i\|_{\psi_2}$$

This is the maximal difference, since

$$\left\| \sum x_i X_i \right\|_{\psi_2} \leqslant \sum |x_i| \|X_i\|_{\psi_2} \leqslant \left(\sum |x_i| \right) \cdot \max \|X_i\|_{\psi_2} \leqslant \sqrt{n} \cdot \max \|X_i\|_{\psi_2}$$

Since we often want bounds which are independant of dimension, this is not a useful bound in practice.

Example. *The isotropic random vector X chosen uniformly randomly from $\{\sqrt{n} \cdot e_k\}$ is subgaussian, but not quantitatively subgaussian. Since*

$$\exp\left(\exp(X_k^2/t^2)\right) = \exp(n/t^2)/n,$$

we find

$$\|X\|_{\psi_2} \geqslant \|X_k\|_{\psi_2} = \left(\frac{n}{\log 2n} \right)^{1/2} \gtrsim \left(\frac{n}{\log n} \right)^{1/2}.$$

This large norm makes the subgaussian property fairly useless in practice.

In fact, if $X \in \mathbf{R}^n$ is an isotropic, discrete random vector with $\|X\|_{\psi_2} \leqslant 1$, then it must be supported on at least e^{cn} points. Thus subgaussian random vectors are not quantifiably discrete.

Theorem 2.5. *If X is a discrete, isotopic random vector with support S, and $\|X\|_{\psi_2} \leqslant 1$, then $|S| \geqslant \exp(cn)$.*

Proof. Suppose that $|X| \leqslant A \cdot n^{1/2}$ for some constant A. Note that

$$|X|^2 \leqslant \sup_{s \in S} |X \cdot s|$$

Note that since $\|X \cdot s\|_{\psi_2} \leqslant An^{1/2}$, we have

$$n = \mathbf{E}\,|X|^2 \leqslant \mathbf{E} \sup_{s \in S} |X \cdot s| \leqslant A \cdot \sqrt{n \log |S|}.$$

Thus $|S| \geqslant \exp(n/A^2)$.

It suffices to reduce the general case to the last case with a constant A independant of n. Because $\|X\|_{\psi_2} \leqslant 1$, there is a universal constant C such that for any x, $\mathbf{E}(X \cdot x)^4 \leqslant C$. Let

$$Y = X \cdot \mathbf{I}(|X|^2 \leqslant 4Cn) \quad \text{and} \quad Y' = X \cdot \mathbf{I}(|X|^2 > 4Cn)$$

By Cauchy-Schwartz,

$$\mathbf{E}((Y' \cdot x)^2) \leqslant (\mathbf{E}((X,x)^4)\,\mathbf{P}(|X|^2 > 4C \cdot n))^{1/2} \leqslant 1/2$$

Thus $|x|^2/2 \leqslant \mathbf{E}((Y \cdot x)^2) \leqslant |x|^2$, and so $I_n/2 \preceq \Sigma(Y) \preceq I_n$. In particular, this means that

$$I_n \preceq \Sigma(Y)^{-1} \preceq 2 \cdot I_n$$

The vector $\Sigma(Y)^{-1/2}Y$ is isotropic, and $|\Sigma(Y)^{-1/2}Y|^2$ is upper bounded by $8C \cdot n$. Thus we can set $A = \sqrt{8C}$. □

The uniform distribution on the sphere is an example of a well behaved subgaussian random variable for which the coordinates are not independant of one another.

Theorem 2.6. *If X is chosen uniformly at random on S^{n-1}, $\|X\|_{\psi_2} \lesssim n^{-1/2}$.*

Proof. By rotation invariance, to bound $\|x \cdot X\|_{\psi_2}$, it suffices to bound $\|X_1\|_{\psi_2}$. We also only need tail bounds for X_1 if $t < 1$, for they are trivial for $t \geqslant 1$. If we let Z be a Gaussian vector, then $Z/|Z|$ is identically distributed to X. We know

$$\mathbf{P}(||Z| - \sqrt{n}| \geqslant t/2) \leqslant 2\exp(-Ct^2),$$

so

$$\begin{aligned}\mathbf{P}(X_1 \geqslant t/\sqrt{n}) &= \mathbf{P}(Z_1/|Z| \geqslant t/\sqrt{n}) \\ &\leqslant 2\exp(-Ct^2) + \mathbf{P}(Z_1 \geqslant t(1 - t/2\sqrt{n})) \\ &\leqslant 2\exp(-Ct^2) + \mathbf{P}(Z_1 \geqslant t) \\ &= 2\exp(-Ct^2) + 2\exp(-Ct^2) = 4\exp(-Ct^2).\end{aligned}$$

This gives the subgaussian bound required. □

Note that the uniform distribution on the sphere is not isotropic. But if we scale by a factor of $n^{1/2}$, it becomes an isotropic distribution, and the resulting distribution has a subgaussian norm independant of n.

Corollary 2.7. *If X is a uniformly chosen vector on the sphere of radius $n^{1/2}$ in $\mathbf{R}^n$, then $\|X\|_{\psi_2} \lesssim 1$.*

If $x_1, \ldots, x_m$ are values with $x_1^2 + \cdots + x_m^2 = 1$, and X is uniformly distribution on the sphere of radius $n^{1/2}$ in $\mathbf{R}^n$, then $x_1X_1 + \cdots + x_nX_n$ looks like a $N(0,1)$ distribution when $n \gg m$. This observation is known as the *projective* central limit theorem. The last theorem shows the aspect of this result related to the tails of the distribution.

One may conjecture that for most families of isotropic convex bodies, the subgaussian bound is uniform in the dimension of the space the bodies lie in. But this need not be the case even for nice convex bodies.

Example. *Let K be the ball of radius t with respect to the l^1 norm in $\mathbf{R}^n$, i.e.*

$$K = \{x \in \mathbf{R}^n : |x_1| + \cdots + |x_n| \leqslant 1\}.$$

Since the l^1 ball has volume $2^n/n!$, and the intersection of K with any plane $\{x_1 = s\}$ is equal to an l^1 ball of radius $1 - s$, if $t < 0.1$,

$$\begin{aligned}\mathbf{P}(X_1 \geqslant t) &= \frac{1}{|K|}\frac{2^{n-1}}{(n-1)!}\int_t^1 (1-s)^{n-1}\,ds \\ &= \frac{(1-t)^n}{2} = \frac{\exp(n\log(1-t))}{2} \geqslant \exp(-nt)/2.\end{aligned}$$

Thus $\|X\|_{\psi_2} \gtrsim 1$. Since $K = -K$, $\mathbf{E}(X_iX_j) = 0$ if $i \neq j$. But the coordinates of X are all identically distributed, some scalar multiple of X is isotropic. One can calculate that $\mathbf{V}(X_i^2) = 2/(n+1)(n+2)$, so tX is isotropic, where $t = [(n+1)(n+2)/2]^{1/2}$. But now tX is certainly not uniformly subgaussian in n, because $\|tX\|_{\psi_2}$ is proportional to n.

Nonetheless, it is possible to prove that if K is an arbitrary isotropic convex body, and X is uniformly distributed on K, then X is uniformly *subexponential*, i.e. $\|X\|_{\psi_1} \lesssim 1$, uniformly in n. This follows C. Borell's lemma.

2.3 Concentration For Lipschitz Functions

Let X be a subgaussian vector, and f a real valued function. A natural question to ask is when $f(X)$ concentrates about it's mean $\mathbf{E}f(X)$. For linear functions f, this question is easy. And if f does not oscillate too much under small pertubations of the input, the theorem remains true. Here we establish a result for Lipschitz functions f.

The core technique to our proof is to utilize an isoperimetric inequality for the probability measure, as well as a related 'blowup' phenomenon. Our main result is fo Lipschitz functions on the unit sphere, but the techniques can be applied to Lipschitz functions on any domain with similar phenomena, which we indicate at the end of this section.

First, we state, without proof, the isoperimetry phenomenon for the sphere. Recall that if E is a subset of a metric space, we let

$$E_\delta = \{x : d(x, E) < \delta\}$$

denote the δ thickening of E. We let σ denote the normalized surface area measure on S^{n-1}. Now let C denote a spherical cap with $\sigma(C) = A$. Isoperimetry says that if $E \subset S^{n-1}$ is *any* set with $\sigma(E) = A$, then $\sigma(E_\delta) \geqslant \sigma(C_\delta)$. In other words, spherical caps minimize volume expansion on the sphere. A simple corollary is a blow-up phenomena.

Lemma 2.8. *Let $E \subset S^{n-1}$. There exists a universal constant c such that if $\sigma(E) \geqslant 1/2$, then for any $t \geqslant 0$, $\sigma(E_t) \geqslant 1 - 2\exp(-cnt^2)$.*

Proof. Let H denote the lower hemisphere of S^{n-1}, i.e.

$$H = \{x \in S^{n-1} : x_1 \leqslant 0\}.$$

By assumption, $\sigma(E) \geqslant 1/2 = \sigma(H)$. Thus the isoperimetric inequality impies that $\sigma(E_t) \geqslant \sigma(H_t)$. Consider $x \in S^{n-1}$ with $x_1 \leqslant 2^{-1/2}t$. Set $C = (x_2^2 + \cdots + x_n^2)^{1/2}$. Then $\sqrt{1 - t^2/2} \leqslant C \leqslant 1$. If we set

$$x' = (0, x_2/C, \ldots, x_n/C) \in H$$

Then

$$|x - x'|^2 = |x_1|^2 + (1 - C)^2 \leqslant t^2/2 + \left(1 - \sqrt{1 - t^2/2}\right)^2$$
$$= 2\left(1 - \sqrt{1 - t^2/2}\right) = \frac{t^2}{1 + \sqrt{1 - t^2/2}} \leqslant t^2.$$

So $|x - x'| \leqslant t$. In particular, this means we have proved

$$H_t \supset \left\{x \in S^{n-1} : x_1 \leqslant t2^{-1/2}\right\}$$

If X is uniformly distributed on the unit sphere, then $\|X\|_{\psi_2} \lesssim 1$, which means

$$\sigma(H_t) \geqslant 1 - \mathbf{P}\left(X_1 \geqslant t2^{-1/2}\right) \geqslant 1 - 2\exp(-cnt^2). \qquad \square$$

We can use this fact to obtain blow up phenomena even for exponentially small sets.

Lemma 2.9. *Let E be a subset of S^{n-1} with $\sigma(E) > 2\exp(-cns^2)$. Then for any $t \geqslant s$, $\sigma(E_{2t}) \geqslant 1 - \exp(cnt^2)$.*

Proof. First, we argue that $\sigma(E_s) > 1/2$. If not, then $\sigma(E_s^c) \geqslant 1/2$. So we can then apply the last lemma to conclude that for any t,

$$\sigma((E_s^c)_t) \geqslant 1 - 2\exp(-cnt^2)$$

In particular, we can select some $t < s$ such that $\sigma((E_s^c)_t) + \sigma(E) > 1$, so $(E_s^c)_t \cap E$ is non-empty. But this mean that $d(E, E_s^c) \leqslant t < s$, which is impossible. Thus $\sigma(E_s) > 1/2$, so we can apply the last lemma to E_s to yield the required inequality. $\square$

Using isoperimetry and blow-up, we can now prove a concentration result for Lipschitz functions on the sphere. Given a Lipschitz function f, we let $\|f\|_{\text{Lip}}$ denote the minimum value with $|f(x) - f(y)| \leqslant \|f\|_{\text{Lip}}|x - y|$.

Theorem 2.10. *If X is uniformly distribution on S^{n-1}, and $f : S^{n-1} \to \mathbf{R}$ is Lipschitz, then $\|f(X) - \mathbf{E}f(X)\|_{\psi_2} \lesssim \|f\|_{Lip} n^{-1/2}$.*

Proof. Let M be a medium for $f(X)$, i.e. a value such that

$$\mathbf{P}(f(X) \geqslant M) \geqslant 1/2 \quad \mathbf{P}(f(X) \leqslant M) \geqslant 1/2$$

Let $E = \{f(X) \leqslant M\}$ denote a level set of f. Then

$$E_t \subset \{f(X) \leqslant M + \|f\|_{\mathrm{Lip}} \cdot t\}.$$

This means that

$$\mathbf{P}(f(X) \leqslant M + \|f\|_{\mathrm{Lip}} \cdot t) \geqslant \mathbf{P}(E_t) \geqslant 1 - \exp(-cnt^2)$$

Similarily, we can show

$$\mathbf{P}(f(X) \geqslant M - \|f\|_{\mathrm{Lip}} \cdot t) \geqslant 1 - \exp(-cnt^2)$$

and then a union bound shows

$$\mathbf{P}(|f(X) - M| \geqslant \|f\|_{\mathrm{Lip}} \cdot t) \leqslant 2\exp(-cnt^2)$$

This gives that $\|f(X) - M\|_{\psi_2} \lesssim \|f\|_{\mathrm{Lip}} n^{-1/2}$. But we can now apply centering to show

$$\|f(X) - \mathbf{E}\, f(X)\|_{\psi_2} \lesssim \|f(X) - M\|_{\psi_2} \lesssim \|f\|_{\mathrm{Lip}} \cdot n^{-1/2}. \qquad \square$$

Remark. Concentration around the expectation and concentration around the medium are essentially equivalent facts. Centering tells us that if M is a median, then for any random variable X, $\|X - \mathbf{E}\, X\|_{\psi_2} \lesssim \|X - M\|_{\psi_2}$. On the other hand,

$$\mathbf{P}(X \geqslant \mathbf{E}\, X + t) \leqslant 2\exp\left(-\frac{ct^2}{\|X - \mathbf{E}\, X\|_{\psi_2}^2}\right)$$

In particular, if $C = (\log(4))^{1/2}/c$, and $t \geqslant C\|X - \mathbf{E}\, X\|_{\psi_2}$, then $\mathbf{P}(|X - \mathbf{E}\, X| \geqslant t) \leqslant 1/2$, which means that $|M - \mathbf{E}\, X| \leqslant C\|X - \mathbf{E}\, X\|_{\psi_2}$, and so

$$\|X - M\|_{\psi_2} \lesssim |\mathbf{E}\, X - M| + \|X - \mathbf{E}\, X\|_{\psi_2} \leqslant (1 + C)\|X - \mathbf{E}\, X\|_{\psi_2}$$

In particular, since C is an absolute constant, we conclude $\|X - M\|_{\psi_2}$ and $\|X - \mathbf{E}\, X\|_{\psi_2}$ are comparable to one another.

Example. *For a geometric application, we show that while there are at most n orthogonal vectors in* $\mathbf{R}^n$*, we can have exponentially many almost orthogonal vectors. Two unit vectors x and y are almost orthogonal if* $|x \cdot y| \leqslant \varepsilon$*. We construct such a set inductively. Consider unit vectors* $e_1, \ldots, e_N$*, which are almost orthogonal. For each k, we can consider* $E_k = \{x \in S^{n-1} : |(x \cdot e_k)| \leqslant \varepsilon\}$*. Then* $\sigma(E_k) \geqslant 1 - 2\exp(-cn\varepsilon^2)$*, and so* $\sigma(E_1 \cap \cdots \cap E_N) \geqslant 1 - 2N\exp(-cn\varepsilon^2)$*. If* $N < \exp(cn\varepsilon^2)/2$*, this is positive, so there certainly exists a unit vector simultaneously orthogonal to all other vectors. Adding this to the list and continuing, we can work up to the point where* $N \geqslant \exp(cn\varepsilon^2)/2$*.*

There is nothing really special to the sphere here. Given any other measure space with a metric, we can consider isoperimetric inequalities. If the minimizers of the isoperimetry problem have a mass blow up, we can obtain the same result. Here we also consider the examples of concentration of Gaussian measures, and concentration of mass on the Hamming cube.

Example. *Consider* $\mathbf{R}^n$ *equipped with the Gaussian measure, which has the Gaussian distribution as a density function. It is non-obvious, but the minimizers of measure expansion are achieved by half planes. From this, we can calculate the precise constants of the blow up phenomenon, and then deduce that if* $X \in \mathbf{R}^n$ *is Gaussian, and f is Lipschitz, then* $\|f(X) - \mathbf{E}f(X)\|_{\psi_2} \lesssim \|f\|_{Lip}$*. We should expect the Gaussian result to look essentially the same as the result based on the uniform distribution on spheres, because in high dimensions, the two results are essentially the same.*

Example. *A similar phenomenon is obtained over* $\{-1,1\}^n$*, where the measure is the uniform probability distribution, and the metric is the Hamming distance, i.e. for* $x, y \in \{-1,1\}^n$*, the Hamming distance gives the number of indices i upon which* $x_i \neq y_i$*. The minimizers for the isoperimetry problem are balls with respect to the Hamming distance. We can conclude from this that* $\|f(X) - \mathbf{E}f(X)\|_{\psi_2} \lesssim \|f\|_{Lip} n^{-1/2}$*. Similar techniques work for the Hamming distance on the symmetric group, and give the same equation.*

Example. *If M is a Riemannian manifold, we can consider the arclength distance, as well as normalized volume of M inducing a probability distribution X chosen uniformly at random on M. If* $c(M)$ *denotes the infinum of the Ricci curvature tensor over all points, and* $c(M) > 0$*, then we have a concentration bound*

$$\|f(X) - \mathbf{E}f(X)\|_{\psi_2} \lesssim \frac{\|f\|_{Lip}}{c(M)^{1/2}}.$$

For instance, $c(S^n) = n$, which gives the concentration inequality for the sphere. Other examples include the matrix group $SO(n)$, with the metric induced by the Frobenius norm, which gives

$$\|f(X) - \mathbf{E}f(X)\|_{\psi_2} \lesssim \frac{\|f\|_{Lip}}{n^{1/2}}$$

Another important example is the Grassmanian space G_{nm} consisting of m dimensional subspaces of $\mathbf{R}^n$, with the distance metric between two vectors spaces V and W given by operator norm of $\|P_V - P_W\|$, where P_V and P_W are orthogonal projections. We obtain the same concentration as for $SO(n)$. We note that the measure given in $SO(n)$ is the Haar measure, and the measure on G_{nm} is the translation invariant measure given by the action of $SO(n)$ on the space.

Example. *Let $\Phi(x)$ denote the cumulative distribution function of a normal distribution. If $Z \sim N(0, I_n)$, then $\phi(Z) = (\Phi(Z_1), \ldots, \Phi(Z_n))$ is uniformly distributed on $[0,1]^n$. To see why, it suffices to show $\Phi(Z_1)$ is uniformly distributed on $[0,1]$, and we calculate that for $t \in [0,1]$,*

$$\mathbf{P}(\Phi(Z_1) \leqslant t) = \mathbf{P}(Z_1 \leqslant \Phi^{-1}(t)) = \Phi(\Phi^{-1}(t)) = t.$$

Given a Lipschitz function $f : [0,1]^n \to \mathbf{R}$, consider $f \circ \phi : \mathbf{R}^n \to \mathbf{R}$. Then $\|f \circ \phi\|_{Lip} \leqslant \|f\|_{Lip}\|\phi\|_{Lip}$. Since

$$|\nabla \Phi(x)| = \frac{|x| e^{-|x|^2/2}}{(2\pi)^{n/2}} \lesssim 1$$

Thus $\|\Phi\|_{Lip} \lesssim 1$, and so

$$|\phi(x-y)| \leqslant \sqrt{\sum \Phi(x_i - y_i)^2} \lesssim \sqrt{\sum |x_i - y_i|^2} = |x - y|$$

which implies $\|\phi\|_{Lip} \lesssim 1$. Thus we can apply concentration in Gaussian space to conclude that if $X = \phi(Z)$, then $\|f(X) - \mathbf{E}(f(X))\|_{\psi_2} \lesssim \|f\|_{Lip}$. Thus we have Lipschitz concentration for the uniform distribution on $[0,1]^n$.

Example. *If the density of a random vector X in $\mathbf{R}^n$ is of the form $\exp(-U(x))$, where $U \to \mathbf{R}^n \to \mathbf{R}$. Assume there is κ such that $H(U) \geq \kappa$. Then for any Lipschitz function $f : \mathbf{R}^n \to \mathbf{R}$, $\|f(X) - \mathbf{E}(f(X))\|_{\psi_2} \lesssim \|f\|_{Lip} \cdot \kappa^{-1/2}$.*

Example. *Let X be a random vector whose coordinates are independant and $|X_i| \leqslant 1$ almost surely. Then* Talagrand's concentration inequality *implies $\|f(X) - \mathbf{E}(f(X))\|_{\psi_2} \lesssim \|f\|_{Lip}$.*

2.4 Matrix Concentration

Let A be an $m \times n$ matrix. Recall the operator norm $\|A\|$, which is the smallest value such that $|Ax| \leqslant \|A\||x|$ for all $x \in \mathbf{R}^n$. We can use a *covering argument* to establish concentration results for the operator norm of random matrices. One difficulty in bounding the operator norm is that we must bound the random quantities $|Ax|$ for *infinitely many* values x. To do this, we rely on a *covering argument*.

First, we need to recall some notation. If X is a metric space, a ε *net* N is a subset of X such that for any $x \in X$, there is $y \in N$ such that $d(x, y) < \varepsilon$. The *covering number* of X is the smallest cardinality of a ε net. On the other hand, a ε packing of X is a ε separated family of points in X. The cardinality of a maximal ε separated family is denoted by $P(X, \varepsilon)$. We find that

$$P(X, 2\varepsilon) \leqslant N(X, \varepsilon) \leqslant P(X, \varepsilon).$$

This is justified by the fact that each element of a ε net can only cover a single element of a ε packing at once, whereas every point in X is contained within 2ε of a maximal ε packing.

Lemma 2.11. *If $X \subset \mathbf{R}^n$,*

$$\frac{|X|}{\varepsilon^n |B|} \leqslant N(X, \varepsilon) \leqslant P(X, \varepsilon) \leqslant 2^n \frac{|X + (\varepsilon/2)B|}{\varepsilon^n |B|},$$

where B is the unit ball in $\mathbf{R}^n$.

Proof. We utilize a *volumetric argument*. To prove the lower bound, if we can cover X by N balls of radius ε, then a union bound gives $|X| \leqslant N(\varepsilon^n |B|)$. On the other hand, to prove the upper bound, consider a packing of X consisting of N points. Then the balls of radius $\varepsilon/2$ around each point are disjoint from one another, and each ball is contained in $X + (\varepsilon/2)B$, so we conclude $|X + (\varepsilon/2)B| \geqslant N(\varepsilon/2)^n |B|$. □

Example. *By convexity, if B is the unit ball, then $B + (\varepsilon/2)B = (1 + \varepsilon/2)B$. Thus the bound above gives*

$$1/\varepsilon^n \leqslant N(B, \varepsilon) \leqslant P(B, \varepsilon) \leqslant 2^n \frac{(1 + \varepsilon/2)^n}{\varepsilon^n} = (1 + 2/\varepsilon)^n$$

The same bound is true for the unit sphere S^{n-1}, and for large n, we should expect this bound also to be tight for the unit sphere, since the majority of the mass of B is concentrated near S^{n-1}.

Lemma 2.12. *Let A be an $m \times n$ matrix, and N an ε net on S^{n-1}. Then*

$$\sup\{|Ax| : x \in N\} \leqslant \|A\| \leqslant (1-\varepsilon)^{-1} \sup\{|Ax| : x \in N\}.$$

If, additionally, M is a ε net on S^{n-1}, then

$$\sup\{(Ax) \cdot y : x \in N, y \in M\} \leqslant \|A\| \leqslant \frac{1}{1-2\varepsilon} \cdot \sup\{(Ax) \cdot y : x \in N, y \in M\}.$$

Proof. We begin with the first equation. The lower bound is obvious. Given $x_0 \in S^{n-1}$, consider $x \in N$ with $|x - x_0| \leqslant \varepsilon$. Then

$$|Ax_0| \leqslant |Ax| + |A \cdot (x_0 - x)| \leqslant |Ax| + \varepsilon\|A\| \leqslant \sup\{|Ax| : x \in N\} + \varepsilon\|A\|.$$

Taking suprema on both sides for all $x_0 \in S^{n-1}$ gives

$$\|A\| \leqslant \sup\{|Ax| : x \in N\} + \varepsilon\|A\|.$$

And rearranging gives the upper bound.

To prove the second equation, we note that for any $y \in \mathbf{R}^m$,

$$|y| = \sup\{z \cdot y : |z| = 1\}$$

This gives the lower bound. To prove the upper bound, for each $x_0 \in S^{n-1}$ and $y_0 \in S^{m-1}$, consider $x \in N$, $y \in M$ with $|x - x_0| \leqslant \varepsilon$ and $|y - y_0| \leqslant \varepsilon$. Thus

$$\begin{aligned}(Ax_0) \cdot y_0 &= Ax \cdot y + A(x_0 - x) \cdot y + Ax_0 \cdot (y_0 - y) \\ &\leqslant Ax \cdot y + 2\varepsilon\|A\| \\ &\leqslant \sup\{(Ax \cdot y) : x \in N, y \in M\} + 2\varepsilon\|A\|\end{aligned}$$

We then just take suprema over x_0 and y_0, and rearrange. □

Theorem 2.13. *Let A be an $m \times n$ matrix with independant, mean zero subgaussian entries. Then if $K = \max \|A_{ij}\|_{\psi_2}$, and $t > 0$*

$$\|A\| \lesssim K(\sqrt{m} + \sqrt{n} + t)$$

with probability at least $1 - 2\exp(-t^2)$, i.e. with overwhelming probability.

Proof. Fix $\varepsilon = 1/4$. Then consider a ε net N of S^{n-1} with $|N| \leqslant 9^n$, and a ε net M of S^{n-1} with $|M| \leqslant 9^n$. Then we know

$$\|A\| \leqslant 2 \sup \{(Ax) \cdot y : x \in N, y \in M\}$$

For each $x \in N$ and $y \in M$, we calculate

$$(Ax) \cdot y = A_{ij} x_i y_j$$

This is the sum of nm subgaussian independant random variables. Thus

$$\|(Ax) \cdot y\|_{\psi_2}^2 \lesssim \sum \|A_{ij} x_i y_j\|_{\psi_2}^2 \leqslant K^2 \sum x_i^2 y_j^2 = K^2.$$

Thus

$$\mathbf{P}(|(Ax) \cdot y| \geqslant u) \leqslant 2 \exp(-cu^2/K^2)$$

Applying a union bound, we find that

$$\begin{aligned} \mathbf{P}(\|A\| \leqslant 2u) &\geqslant \mathbf{P}(\forall x \in N, y \in M : |(Ax) \cdot y| \leqslant u) \\ &\geqslant 1 - 2 \cdot 9^{n+m} \exp(-cu^2/K^2), \end{aligned}$$

setting $u = (C/c)^{-1/2} K(\sqrt{m} + \sqrt{n} + t)$, where C is a large constant, we find that since $(\sqrt{m} + \sqrt{n} + t)^2 \geqslant (m + n + t^2)$,

$$\begin{aligned} \mathbf{P}(\|A\| \leqslant CK(\sqrt{m} + \sqrt{n} + t)) &\geqslant 1 - 2 \cdot 9^{n+m} \exp(-C(\sqrt{m} + \sqrt{n} + t)^2) \\ &\geqslant 1 - 2 \cdot 9^{n+m} \exp(-C(m+n) - Ct^2)). \end{aligned}$$

For $C \geqslant \log 9$, this is greater than $1 - 2\exp(-Ct^2) \geqslant 1 - 2\exp(-t^2)$. □

Corollary 2.14. $\mathbf{E}\|A\| \lesssim K(\sqrt{m} + \sqrt{n})$.

Remark. The expectation bound is essentially tight. If the entries $\{A_{ij}\}$ have unit variances, then

$$\begin{aligned} \mathbf{E}\|A\| &\geqslant \frac{1}{\min(m,n)^{1/2}} \mathbf{E}\|A\|_F \geqslant \frac{1}{\min(m,n)^{1/2}} \left(\sum \mathbf{E} A_{ij}^2 \right)^{1/2} \\ &= \max(n,m)^{1/2} \gtrsim n^{1/2} + m^{1/2}. \end{aligned}$$

By relying on Bernstein's inequality rather than Hoeffding's inequality, we can get a much tighter estimate, which also shows that A is with high probability close to a isometry scaled by $m^{1/2}$; this should be directly compared to the fact that a random vector is with high probability close to the sphere with radius $n^{1/2}$.

Theorem 2.15. *Let A be an $m \times n$ matrix whose rows are independant, mean zero, subgaussian isotropic random vectors. Then if $s_1, \ldots, s_n$ are the singular values, and $K = \max \|A_i\|_{\psi_2}$, then*

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leqslant s_n \leqslant s_1 \leqslant \sqrt{m} + CK^2(\sqrt{n} + t)$$

with probability $1 - 2\exp(-t^2)$.

Proof. We prove the stronger conclusion that

$$|(A^T A)/m - I_n| \leqslant K^2 \max(\delta, \delta^2) \quad \text{where} \quad \delta = C\left(\sqrt{n/m} + t/\sqrt{m}\right).$$

To compute this, it suffices to prove that for a $1/4$-net N, with $|N| \leqslant 9^n$, with the required probability we have

$$\max_{x \in N} \left| |Ax|^2/m - 1 \right| \leqslant \varepsilon/2.$$

Fix $x \in N$. If we set $X_i = A_i \cdot x$, then $|Ax|^2 = \sum X_i^2$. By assumption, the A_i are independant, isotropic, subgaussian random vectors with $\|A_i\|_{\psi_2} \leqslant K$. Thus the X_i are independant subgaussian random vectors with $\mathbf{E} X_i^2 = 1$ and $\|X_i\|_{\psi_2} \leqslant K$. Therefore $X_i^2 - 1$ are independant, mean zero, subexponential random variables with $\|X_i^2 - 1\| \lesssim K^2$, and we can apply Bernstein's inequality to conclude that

$$\begin{aligned}
\mathbf{P}(|Ax|^2/m - 1| \geqslant \varepsilon/2) &= \mathbf{P}(|(1/m)\sum X_i^2 - 1| \geqslant \varepsilon/2) \\
&\leqslant 2\exp(-cm\min(\varepsilon^2/K^4, \varepsilon/K^2)) \\
&= 2\exp(-c\delta^2 m) \\
&\leqslant 2\exp(-cC^2(n + t^2))
\end{aligned}$$

If $C \gg 0$, we can apply a union bound to finish the proof. □

By applying some matrix calculus, we can obtain variants of Hoeffding and Bernstein's inequality for matrices. We recall that if A is an $n \times n$ symmetric matrix, then it has a diagonalization $A = \sum \lambda_i u_i u_i^T$, where $\lambda_i \in \mathbf{R}$ and the collection of vectors $\{u_i\}$ is an orthogonal basis for $\mathbf{R}^n$. Given a real-valued function f defined on a neighbourhood of the eigenvalues of A, we set $f(A) = \sum f(\lambda_i) u_i u_i^T$. If f is a polynomial, i.e. $f(x) = a_0 +$

$a_1 x + \cdots + a_m x^m$, then $f(A) = a_0 + a_1 A + \cdots + a_m A^m$. Given two symmetric matrices A, B, we say $A \preceq B$ if $B - A$ is positive definite. Our proof of the matrix Hoeffding and Bernstein's inequality will be based on the moment generating proofs in the scalar case. But if A and B are independant, it is no longer necessarily true that $\mathbf{E}(e^{A+B}) = \mathbf{E}(e^A)\,\mathbf{E}(e^B)$. We circumvent this by applying a trace estimate, konwn as Lieb's inequality, which we don't prove.

Theorem 2.16 (Lieb's Inequality)**.** *Let H be a symmetric $n \times n$ matrix. Then the function $f(A) = tr[\exp(H + \log A)]$ is convex on the space of positive-definite $n \times n$ symmetric matrices.*

Applying Lieb's inequality to e^Z for some symmetric random matrix A, and applying Jensen's inequality yields the following corollary.

Corollary 2.17. *If H is a symmetric $n \times n$ matrix, and A is a random symmetric matrix, then $\mathbf{E}(tr(e^{H+A})) \leqslant tr(e^{H + \log \mathbf{E} e^Z})$.*

Now we can prove Bernsteins' inequality.

Theorem 2.18. *Let $A_1, \ldots, A_n$ be independant, mean zero, $m \times m$ symmetric random matrices with $\|A_i\| \leqslant K$. Then*

$$\mathbf{P}\left(\left\|\sum A_i\right\| \geqslant t\right) \leqslant 2m \exp\left(\frac{-t^2/2}{\sigma^2 + Kt/3}\right).$$

Proof. It suffices to control the largest eigenvalue of $S = \sum A_i$. To do this, we apply a Chernoff bound. Thus for any $\lambda > 0$,

$$\mathbf{P}\left(\|S\| \geqslant t\right) \leqslant e^{-\lambda t}\,\mathbf{E}\left(e^{\lambda \|S\|}\right) = e^{-\lambda t}\,\mathbf{E}\left(\|e^{\lambda S}\|\right).$$

where the last equality follows because the largest eigenvalue of $e^{\lambda S}$ is equal to the exponential of the largest eigenvalue of S. Since all eigenvalues of $e^{\lambda S}$ are positive, $\|e^{\lambda S}\| \leqslant \operatorname{tr}(e^{\lambda S})$. Applying Lieb's inequality iteratively, we conclude

$$\mathbf{E}(\operatorname{tr}(e^{\lambda S})) \leqslant \operatorname{tr}\left(e^{\sum \log \mathbf{E}(e^{\lambda A_i})}\right)$$

All that remains is to bound $\mathbf{E}(e^{\lambda A_i})$. Note that if $|z| < 3$, then

$$e^z \leqslant 1 + z + \frac{1}{1 - |z|/3}\frac{z^2}{2}$$

If $z = \lambda x$, then if $|x| \leqslant K$ and $|\lambda| < 3/K$ this ineuality implies that

$$e^{\lambda x} \leqslant 1 + \lambda x + g(\lambda)x^2$$

where $g(\lambda) = (\lambda^2/2)/(1 - |\lambda|K/3)$. Applying this inequality to symmetric matrices yields

$$e^{\lambda A_i} \preceq 1 + \lambda A_i + g(\lambda)A_i^2$$

Taking expectations on both sides gives that

$$\mathbf{E}(e^{\lambda A_i}) \preceq 1 + g(\lambda)A_i^2 \preceq e^{g(\lambda)A_i^2}.$$

where we used the fact that $1 + g(\lambda)x^2 \leqslant e^{g(\lambda)x^2}$. Thus

$$\operatorname{tr}\left(\exp\left(\sum \log \mathbf{E}\, e^{\lambda X_i}\right)\right) \leqslant \operatorname{tr}\left(\exp(g(\lambda)Z)\right)$$

where $Z = \sum \mathbf{E}(X_i^2)$. But now

$$\operatorname{tr}\left(\exp(g(\lambda)Z)\right) \leqslant n\exp(g(\lambda)\|Z\|) \leqslant n\exp(g(\lambda)\sigma^2)$$

Putting this inequality back to the original, and setting $\lambda = t/(\sigma^2 + Kt/3)$ completes the proof. □

Remark. To make this inequality look closer to the classical Bernstein's inequality, we note that it implies there is a universal constant c such that

$$\mathbf{P}\left(\left\|\sum A_i\right\| \geqslant t\right) \leqslant 2m\exp\left(-c\min(t^2/\sigma^2, t/K)\right)$$

Corollary 2.19. *Given the A_i as in the last proof,*

$$\mathbf{E}\left\|\sum A_i\right\| \lesssim \left\|\sum \mathbf{E}(A_i^2)\right\|^{1/2} (\log m)^{1/2} + K\log m$$

Proof. TODO □

Similar techniques yield further concentration inequalities for matrices.

Theorem 2.20 (Hoeffding)**.** *Let $\varepsilon_1, \ldots, \varepsilon_n$ be independant symmetric Bernoulli random variables and let $A_1, \ldots, A_n$ by deterministic symmetric $m \times m$ matrices. Then*

$$\mathbf{P}\left(\left\|\sum \varepsilon_i A_i\right\| \geqslant t\right) \leqslant 2m\exp(-t^2/2\sigma^2),$$

where $\sigma^2 = \left\|\sum A_i^2\right\|$.

Proof. TODO □

TODO: LIST OTHER MATRIX CONCENTRATION TECHNIQUES IN VERSHYNIN'S BOOK.

Chapter 3

Applications of High Dimensional Concentration

3.1 Community Detection

For each positive even integer n, and $p, q \in [0,1]$, we construct a random graph $G(n,p,q)$ by dividing n vertices into two sets of $n/2$ vertices, which we call communities. We connect two vertices in a common community independantly with probability p, and two vertices in separate communities with probability q. We assume $p > q$ here so vertices in a common community are more likely to be connected. A natural problem, given such a graph, is to be able to partition the vertices into two communities given no prior knowledge about the graph.

To obtain such an algorithm for this process, we apply our results about matrix concentration. Let A denote the *random* adjacency matrix for $G(n,p,q)$. We can write $A = D + R$, where $D = \mathbf{E}(A)$ is the deterministic part of the adjacency matrix, and R is the random part. It is easy to so the matrix D has rank two. For illustration, if $n = 4$, then after reordering the vertices, we find

$$D = \begin{pmatrix} p & p & q & q \\ p & p & q & q \\ q & q & p & p \\ q & q & p & p \end{pmatrix}.$$

It therefore has two non-zero eigenvalues

$$\lambda_1 = \left(\frac{p+q}{2}\right) \cdot n \quad \text{and} \quad \lambda_2 = \left(\frac{p-q}{2}\right) \cdot n.$$

Corresponding to the two eigenvectors u_1 and u_2. Note that $u_{1i} = 1$ for all i, and gives no useful information. But $u_{2i} = 1$ if i is in the first community, and $u_{2i} = -1$ if i is in the second community. If we could identify u_2, we could identify the communities precisely.

We do not have access to D, but we have access to $D + R$, and we can certainly diagonalize this matrix. Matrix concentration tells us that with probability $1 - 4e^{-n}$, we have $\|R\| \leqslant C \cdot n^{1/2}$ for some universal constant C. Thus if $p - q > 0$, for large n $\|R\|$ is much smaller than $\|D\|$, which is proportional to n. Now this means that all the eigenvalues of A differ from D by at most $C \cdot n^{1/2}$. Furthermore, the Davis-Kahan theorem says the eigenvectors corresponding to these eigenvalues do not differ much from each other either.

Theorem 3.1 (Davis-Kahan)**.** *Let S and T by symmetric $n \times n$ matrices, and let $\lambda_i(S), \lambda_i(T)$, $v_i(S)$ and $v_i(T)$ denote the i'th largest eigenvalues and unit eigenvectors of the matrices. If $|\lambda_i(S) - \lambda_j(S)| \geqslant \delta$ for all $j \neq i$, then*

$$v_i(S) \cdot v_i(T) \geqslant \left(1 - \frac{4\|S-T\|^2}{\delta^2}\right)^{1/2}.$$

This means there exists a sign $\theta \in \{-1, 1\}$ such that

$$|v_i(S) - \theta v_i(T)| \leqslant \frac{2^{3/2}\|S-T\|}{\delta}.$$

In particular, since $\|A - D\| \leqslant C \cdot n^{1/2}$, if we set

$$\delta = \min\left(\frac{p-q}{2}, \left(\frac{p+q}{2} - \frac{p-q}{2}\right)\right) \cdot n = \min(0.5 \cdot (p-q), q) \cdot n.$$

Then we find there is $\theta \in \{-1, 1\}$ such that

$$|v_i(A) - \theta v_i(D)| \lesssim \frac{n^{-1/2}}{\min(q, p-q)}$$

Thus the signs of most of the coefficients of A and D must agree. The number of disagreeing signs between $v_2(A)$ and $v_2(D)$ is bounded up to a constant by $\min(q, p-q)^{-1}$. Thus by finding the second largest eigenvector to A, and clustering the two communities by the sign of the vector, we will be correct with high probability, with few errors. This is known as a *spectral clustering* method. This is efficiently computable even when n is large.

3.2 Covariance Estimation

Suppose we are analyzing data in high dimensions, represented as points $X_1, \ldots, X_m$ sampled from a distribution in $\mathbf{R}^n$. One of the standard tools for studying such data is principal component analysis. Given a distribution X, the distribution can be understood by computing the spectral decomposition of $\Sigma(X)$. The direction corresponding to the largest eigenvalue is known as the *first principal direction*. This explains most of the variability in the data. In some cases, only of the few of the eigenvalues of $\Sigma(X)$ are large, and projection onto the eigenspaces corresponding to these eigenvalues represents the information of the data in a low dimensional space. If only three eigenvalues are significant, this even makes the data visualizable.

In practice, $\Sigma(X)$ cannot be calculated exactly. But we can calculate the sample covariance matrix

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} X_i X_i^T$$

We certainly have $\mathbf{E}(\Sigma) = \Sigma(X)$, and the law of large numbers implies $\Sigma \to \Sigma(X)$ almost surely as $m \to \infty$. But we want a non asymptotic result. Of course, dimensional considerations mean we need at least $m = \Omega(n)$ in order for Σ to be close to $\Sigma(X)$. In fact, $O(n)$ results suffice.

Theorem 3.2. *Suppose X is a random vector such that for any $x \in \mathbf{R}^n$,*

$$\|X \cdot x\|_{\psi_2} \leqslant K \|X \cdot x\|_{L^2(\Omega)}$$

Then

$$\mathbf{E} \|\Sigma - \Sigma(X)\| \lesssim K^2 \left((n/m)^{1/2} + (n/m) \right) \|\Sigma(X)\|$$

In particular, if $m \geqslant n$, then $\mathbf{E} \|\Sigma - \Sigma(X)\| \lesssim K^2 (n/m)^{1/2}$.

Proof. Consider the isotropic random vectors Z and $Z_1, \dots, Z_m$ such that $X = \Sigma(X)^{1/2} Z$ and $X_i = \Sigma(X)^{1/2} Z_i$. Then the subgaussian assumption implies $\|Z\|_{\psi_2} \leqslant K$ and $\|Z_i\|_{\psi_2} \leqslant K$. If we set $\Pi = m^{-1} \sum Z_i Z_i^T - I_n$, then

$$\|\Sigma - \Sigma(X)\| = \|\Sigma^{1/2}(X) \Pi \Sigma^{1/2}(X)\| \leqslant \|\Pi\| \|\Sigma(X)\|$$

But we know from our matrix concentration results tha

$$\mathbf{E}\,\|\Pi\| \lesssim K^2 \left((n/m)^{1/2} + (n/m) \right).$$

which gives the result for free. □

Consider the following application of this theorem. We consider two normal distributions $N(\mu, I_n)$ and $N(-\mu, I_n)$, with means μ and $-\mu$. We then pick m points $X_1, \dots, X_m$, which each have a 50/50 chance of being picked by one of the distributions. A natural goal is to cluster these vectors into whether they were picked from one distribution or the other. Just as in community detection, we can use a spectral clustering algorithm.

Note that if $X = \theta \mu + g$ is identically distributed to $X_1, \dots, X_m$, where θ is a symmetric Bernoulli random variable and g is Gaussian. Note that X is not isotropic. Instead, $\Sigma(X) = I_n + \mu \mu^T$. Note that μ is the only eigenvector, corresponding to the eigenvalue $1 + |\mu|^2$. This makes sense, since the only interesting non-noise related features of the data correspond to whether the data comes from $N(\mu, I_n)$ or $N(-\mu, I_n)$. If $m \sim \varepsilon^{-2} n$, then our results about covariance estimation imply that if we define $\Sigma = m^{-1} \sum X_i X_i^T$, then $\mathbf{E}\,\|\Sigma - \Sigma(X)\| \leqslant \varepsilon(1 + |\mu|^2)$. In the case that $\|\Sigma - \Sigma(X)\| \lesssim \varepsilon(1 + |\mu|^2)$, we can apply the Davis-Kahan theorem to show that there is $\theta' \in \{-1, 1\}$ such that if v is the principal eigenvector of Σ, then

$$|v - \theta' \mu| \lesssim \varepsilon$$

Note that if X_i belongs to $N(\mu, I_n)$, then $X_i \cdot \mu > 0$ with high probability. Thus if we partition $X_1, \dots, X_m$ depending on whether $X_i \cdot v > 0$ or $X_i \cdot v < 0$, then we cluster the data effectively. TODO: Fill in details here.

3.3 The Johnson-Lindenstrauss Lemma

Suppose we have N data points in $\mathbf{R}^n$, where n is very large. We would like to reduce the dimension of the data, while still preserving the geometric

properties of the data points. The simplest data reduction is to project the data points onto a lower dimensional subspace. A natural question is the smallest dimension we can project the points, while still approximately preserving the distance between points. The Johnson-Lindenstrauss lemma says the distances will be approximately preserved when projecting into a space with dimension $\log N$. Given $V \in G_{nm}$, let $Q_V = (n/m)^{1/2} \cdot P_V$.

Lemma 3.3. *Let V be a randomly chosen projection onto an m dimensional subspace of $G_{n,m}$. If $z \in \mathbf{R}^n$ is fixed, and $\varepsilon > 0$, then $\mathbf{E}|Q_V(z)|^2 \leqslant |z|^2$, and with probability greater than $1 - 2\exp(-c\varepsilon^2 m)$,*

$$(1-\varepsilon)|z| \leqslant |Q_V(z)| \leqslant (1+\varepsilon)|z|$$

Proof. Without loss of generality, assume that $|z| = 1$. Then, instead of considering a random subspace V, we can consider a fixed space V acting on a random unit vector z, since the distribution of $Q_V(z)$ will be the same. Using rotation invariance, we may assume that P_V is the projection onto the first m coordinates. Since $\mathbf{E}(z_i^2) = 1/n$ for each i,

$$\mathbf{E}|Q_V(z)|^2 = (n/m) \cdot \sum_{i=1}^{m} \mathbf{E} z_i^2 = 1.$$

Thus the first part of the lemma is proven. Next, we apply the concentration result for Lipschitz functions on a sphere. if $f(x) = |Q_V(x)|$, then $\|f\|_{\text{Lip}} = (n/m)^{1/2}$. Thus

$$\|Q_V(X) - (n/m)^{1/2}\|_{\psi_2} \lesssim 1/m^{1/2},$$

so

$$\mathbf{P}\left(||Q_V(z)| - 1| \geqslant t\right) \leqslant 2\exp(-cmt^2). \qquad \square$$

Theorem 3.4. *Let $V \in G_{nm}$ be uniformly chosen. Then there exists constants c and C such that if X is a set of N points in $\mathbf{R}^n$, $\varepsilon > 0$, and $m \geqslant C \log N/\varepsilon^2$, then with probability $1 - 2\exp(-cm\varepsilon^2)$, the projection Q_V of X onto E satisfies*

$$(1-\varepsilon)|x-y| \leqslant |Q_V(x) - Q_V(y)| \leqslant (1+\varepsilon)|x-y|$$

for all $x, y \in X$.

Proof. We can apply the last lemma, for each $x, y \in X$, to $z = x - y$ and then take a union bound over all possible N^2 pairs of points. Combined with the fact that there is a constant C with $N \leqslant \exp(C\varepsilon^2 m)$, this gives that the inequality is satisfied for all x, y with probability

$$1 - 2N^2 \exp(-c\varepsilon^2 m) \geqslant 1 - 2\exp((2C - c)\varepsilon^2 m)$$

if C is sufficiently small, depending on c, this is bounded by $1-2\exp(-c\varepsilon^2 m)$ for a slightly smaller constant c. □

Remark. It is an important fact that the random choice of projections depends in no way on the incoming data. Furthermore, the dimension n of the ambient space is not featured in the lemma at all. We also remark that the theorem remains true if we consider a random matrix whose rows are independant, mean zero, subgaussian random vectors, and we normalize by $1/m^{1/2}$.

Chapter 4

Techniques for High Dimensional Probability

4.1 Decoupling

In this chapter, we study concentration bounds for quadratic forms

$$X^T AX = \sum A_{ij} X_i X_j$$

where $\{X_i\}$ are independant random variables, and A_{ij} are arbitrary constants. The expectation is easy to describe. If X_i has variance σ_i^2, then

$$\mathbf{E}(X^T AX) = \sum A_{ii} \sigma_i^2$$

But establishing concentration bounds is much harder – one cannot use a Lipschitz bound here unless that variables $\{X_i\}$ are bounded, and this probably won't give a good concentration bound regardless. Decoupling is a technique to replace the random variable $X^T AX$ with $X^T AX'$, where X' is an independant copy of X.

Lemma 4.1. *Let Y and Z be independant random variables such that $\mathbf{E}\, Z = 0$. Then for each convex function F, $\mathbf{E}\, F(Y) \leqslant \mathbf{E}\, F(Y+Z)$.*

Proof. We apply Jensen's inequality. Since $\mathbf{E}\, Z = 0$,

$$\begin{aligned} \mathbf{E}\, F(Y) &= \mathbf{E}\, F(Y + \mathbf{E}(Z)) = \mathbf{E}\, F(\mathbf{E}(Y+Z|Y)) \\ &\leqslant \mathbf{E}(\mathbf{E}(F(Y+Z)|Y)) = \mathbf{E}(F(Y+Z)). \end{aligned}$$

□

Theorem 4.2. *Let A be a diagonal free matrix. Let X be a random vector with independant, mean zero coordinates. Then for any convex function F, $\mathbf{E}(F(X^TAX)) \leqslant \mathbf{E}(F(4X^TAX'))$, where X' is an independant copy of X.*

Proof. Let $\delta_1, \ldots, \delta_n \in \{0,1\}$ be independant symmetric Bernoulli random variables, and define $I = \{k : \delta_k = 1\}$ be a random subset of $\{1, \ldots, n\}$. Since $\mathbf{E}(\delta_i(1-\delta_j)) = 1/4$,

$$\mathbf{E}\left(\sum_{(i,j)\in I\times I^c} A_{ij}X_iX_j\right) = \mathbf{E}\sum_{ij} A_{ij}\delta_i(1-\delta_j)X_iX_j = (1/4)\,\mathbf{E}(X^TAX)$$

We now apply the function F to both sides, calculating

$$\mathbf{E}(X^TAX) \leqslant 4\,\mathbf{E}\left(F\left(\sum_{(i,j)\in I\times I^c} A_{ij}X_iX'_j\right)\right)$$

In particular, this means that we may fix a *non random* choice of I for which this equation still remains true, which we do for the remainder of the proof. Note that $\sum A_{ij}X_iX_j$ is identically distributed to $\sum A_{ij}X_iX'_j$, where X' is an independant copy of X. Write

$$Y = \sum_{(i,j)\in I\times I^c} A_{ij}X_iX'_j \quad Z_1 = \sum_{(i,j)\in I\times I} A_{ij}X_iX'_j \quad \text{and} \quad Z_2 = \sum_{(i,j)\in I^c\times[n]} A_{ij}X_iX'_j.$$

Let $\mathbf{E}'$ denote conditional expectations with respect to all random variables *except* $\{X_i\}_{i\in I^c}$, $\{X'_j\}_{j\in I}$. Then $\mathbf{E}'(Y) = Y$, and $\mathbf{E}'(Z_1) = \mathbf{E}'(Z_2) = 0$. If we apply the last lemma, we conclude

$$F(4Y) \leqslant \mathbf{E}'(F(4Y + 4Z_1 + 4Z_2))$$

Taking expectations on both sides of this inequality concludes the argument, since $Y + Z_1 + Z_2 = \sum A_{ij}X_iX_j$. □

We can use this fact to get bounds on moment generating functions of quadratic forms, which yields deviation inequalities. We first show how to replace the question of random variables X, X' with arbitrary distributions with Gaussian distributions.

Lemma 4.3. *If $X, X' \in \mathbf{R}^n$ are mean zero independant subgaussian random vectors, with $\|X\|_{\psi_2}, \|X'\|_{\psi_2} \leqslant K$. If $g, g' \sim N(0, I_n)$ are independant normal random vectors, and A is an $n \times n$ matrix, then*

$$\mathbf{E} \exp(\lambda X^T A X') \leqslant \mathbf{E} \exp(CK^2 \lambda g^T A g').$$

Proof. We let $\mathbf{E}_X$ denote conditioning with respect to X', averaging over X. When X' is fixed, $X^T A X' = X \cdot AX'$ is subgaussian, with $\|X^T A X'\|_{\psi_2} \leqslant K|AX'|$. Thus

$$\mathbf{E}_X \exp(\lambda X^T A X') \leqslant \exp(C\lambda^2 K^2 |AX'|^2)$$

Note that if $\mathbf{E}_g$ is obtained by averaging over g,

$$\mathbf{E}_g \exp(\gamma g^T A X') = \exp(\gamma^2 |AX'|^2/2)$$

If $\gamma^2 = 2C\lambda^2 K^2$, then we conclude

$$\mathbf{E}_X \exp(\lambda X^T A X') \leqslant \mathbf{E}_g \exp\left((2C)^{1/2} \lambda K g^T A X'\right)$$

Taking expectations on both sides shows that we can replace X with g with the cost of $(2C)^{1/2}K$. A similar argument replaces X' with g' at the cost of an additional $(2C)^{1/2}K$ factor. □

Lemma 4.4. *Let X, X' be mean zero subgaussian random vectors. Then*

$$\mathbf{E} \exp(\lambda X^T A X') \leqslant \exp(CK^4 \lambda^2 \|A\|_F^2)$$

for all λ satisfying $|\lambda| \leqslant c/\|A\|$.

Proof. Consider the singular value decomposition of A, i.e. write

$$A = \sum s_i u_i v_i^T.$$

Consider first the case of two Gaussian random vectors g, g'. Then $g^T A g' = \sum s_i (g \cdot u_i)(g' \cdot v_i)$. Since the u_i and v_i are orthonormal, $\sum s_i (g \cdot u_i)(g' \cdot v_i)$ is identically distributed to $\sum s_i g_i g_i'$. By independence,

$$\mathbf{E}(\exp(\lambda g^T A g')) = \prod \mathbf{E}(\exp(\lambda s_i g_i g_i'))$$

and if $\lambda^2 s_i^2 \leqslant c$,

$$\begin{aligned} \mathbf{E}(\exp(\lambda s_i g_i g_i')) &= \mathbf{E}(\mathbf{E}(\exp(\lambda s_i g_i g_i' | g_i))) \\ &\leqslant \mathbf{E}(\exp(\lambda^2 s_i^2 g_i^2/2)) \leqslant \exp(C\lambda^2 s_i^2) \end{aligned}$$

where we used the fact that g_i^2 is subexponential. This means that provided $\lambda^2 \leqslant c/\max s_i = c/\|A\|$,

$$\mathbf{E}(\exp(\lambda g^T A g')) \leqslant \exp\left(C\lambda^2 \sum s_i^2\right) = \exp(C\lambda^2 \|A\|_F).$$

In general, we apply the comparison inequality. If $\|X\|_{\psi_2}, \|X'\|_{\psi_2} \leqslant K$, then

$$\mathbf{E}(\exp(\lambda X^T A X')) \leqslant \mathbf{E}\exp(CK^2\lambda g^T A g') \leqslant \exp(CK^4\lambda^2\|A\|_F). \qquad \square$$

Theorem 4.5 (Hanson-Wright)**.** *Let X be a random vector with independant, mean zero, subgaussian coordinates. Then for $t \geqslant 0$,*

$$\mathbf{P}\left(|X^T A X - \mathbf{E}\, X^T A X| \geqslant t\right) \leqslant 2\exp\left(-c\min\left(t^2/K^4\|A\|_F^2, t/K^2\|A\|\right)\right).$$

Proof. Without loss of generality, assume $\|X\|_{\psi_2}, \|X'\|_{\psi_2} \lesssim 1$. We note that $\mathbf{E}\, X^T A X = \sum a_{ii}\, \mathbf{E}(X_i^2)$. Thus

$$\begin{aligned}
&\mathbf{P}(X^T A X - \mathbf{E}\, X^T A X \geqslant t) \\
&\qquad \leqslant \mathbf{P}\left(\sum a_{ii}(X_i^2 - \mathbf{E}\, X_i^2) \geqslant t/2\right) + \mathbf{P}\left(\sum_{i\neq j} a_{ij} X_i X_j \geqslant t/2\right)
\end{aligned}$$

We note that

$$\|X_i^2 - \mathbf{E}\, X_i^2\|_{\psi_1} \lesssim \|X_i^2\|_{\psi_1} \lesssim \|X_i\|_{\psi_2}^2 \lesssim 1$$

So Bernstein's inequality implies

$$\begin{aligned}
\mathbf{P}\left(\sum a_{ii}(X_i^2 - \mathbf{E}\, X_i^2) \geqslant t/2\right) &\leqslant \exp\left(-c\min\left(t^2/\sum a_{ii}^2, t/\max|a_{ii}|\right)\right) \\
&\leqslant \exp\left(-c\min\left(t^2/\|A\|_F^2, t/\|A\|\right)\right)
\end{aligned}$$

We use our moment generating bound for the non-diagonal elements. We note that if $S = \sum_{i\neq j} a_{ij} X_i X_j$, then our decoupling bound implies that provided $\lambda \leqslant c/\|A\|$.

$$\begin{aligned}
\mathbf{P}(S \geqslant t/2) &\leqslant \exp(-\lambda t/2)\, \mathbf{E}\exp(\lambda S) \\
&\leqslant \exp(\lambda t/2)\exp(C\lambda^2\|A\|_F^2)
\end{aligned}$$

Optimizing the choice of λ, we conclude

$$\mathbf{P}(S \geqslant t/2) \leqslant \exp(-c\min(t^2/\|A\|_F, t/\|A\|)).$$

Putting the non-diagonal bound with the diagonal bound together, we conclude that

$$\mathbf{P}(X^T AX - \mathbf{E}\, X^T AX \geqslant t) \leqslant 2\exp\left(-c\min\left(t^2/\|A\|_F, t/\|A\|\right)\right). \qquad \square$$

As a consequence, we can obtain concentration bounds for *anisotropic vectors*.

Theorem 4.6. *Let B be an $n \times n$ matrix, and X a random variable with independance, mean zero, unit variance sub-gaussian coordinates. Then if $K = \max \|X_i\|_{\psi_2}$,*

$$\||BX| - \|B\|_F\|_{\psi_2} \lesssim K^2\|B\|.$$

Proof. We apply the Hanson-Wright inequality for $A = B^T B$. Then $X^T AX = |BX|^2$, and $\mathbf{E}\, X^T AX = \|B\|_F^2$. Note that $\|A\| = \|B\|^2$, and

$$\|B^T B\|_F \leqslant \|B^T\|\|B\|_F = \|B\|\|B\|_F$$

Thus, since $K^4 \geqslant K^2$,

$$\mathbf{P}(||BX|^2 - \|B\|_F^2| \geqslant u) \leqslant 2\exp(-(c/K^4)\cdot\min(u^2/\|B\|^2\|B\|_F, u/\|B\|^2)).$$

Setting $u = \varepsilon\|B\|_F^2$, we conclude

$$\mathbf{P}(||BX|^2 - \|B\|_F^2| \geqslant \varepsilon\|B\|_F^2) \leqslant 2\exp(-c\min(\varepsilon, \varepsilon^2)\|B\|_F^2/K^4\|B\|^2).$$

Observe that if $\varepsilon = \max(\delta, \delta^2)$, i.e. $\delta^2 = \min(\varepsilon, \varepsilon^2)$, then

$$\mathbf{P}(||BX| - \|B\|_F| \geqslant \delta\|B\|_F) \leqslant 2\exp(-c\delta^2\|B\|_F^2/K^4\|B\|^2),$$

But this means that

$$\mathbf{P}(|BX| - \|B\|_F| \geqslant t) \leqslant 2\exp(-ct^2/K^4\|B\|^2). \qquad \square$$

4.2 Symmetrization

Another technique often used in high dimensional probability is to replace random variables with symmetric random variables, i.e. variables X for which X is distributed identically to $-X$. For instance, mean zero normal random variables are symmetric, as are symmetric Bernoulli random

variables. To obtain a symmetric version of any random variable X, we can take an independant Bernoulli random variable ε, and consider εX, or we can take an independant copy X' of X, and we then consider $X - X'$. Throughout this section, we let ε_i stand for a family of Bernoulli random variables, independant of each other and any other random variable considered in the argument.

Lemma 4.7. *Let $X_1, \dots, X_n$ be independant mean zero random variables in some normed space. Then*

$$0.5\,\mathbf{E}\left\|\sum \varepsilon_i X_i\right\| \leqslant \mathbf{E}\left\|\sum X_i\right\| \leqslant 2\,\mathbf{E}\left\|\sum \varepsilon_i X_i\right\|.$$

Proof. If $X_1', \dots, X_n'$ are independant copies of $X_1, \dots, X_n$, then since $\mathbf{E}(X_i') = 0$,

$$\mathbf{E}\left\|\sum X_i\right\| \leqslant \mathbf{E}\left\|\sum X_i - \sum X_i'\right\| = \mathbf{E}\left\|\sum (X_i - X_i')\right\|$$

Now note that since $X_i - X_i'$ is symmetric, it has the same distribution as $\varepsilon_i(X_i - X_i')$. Thus

$$\begin{aligned}\mathbf{E}\left\|\sum (X_i - X_i')\right\| &= \mathbf{E}\left\|\sum \varepsilon_i (X_i - X_i')\right\| \\ &\leqslant \mathbf{E}\left\|\sum \varepsilon_i X_i\right\| + \mathbf{E}\left\|\sum \varepsilon_i X_i'\right\| \\ &\leqslant 2\,\mathbf{E}\left\|\sum \varepsilon_i X_i\right\|.\end{aligned}$$

Conversely,

$$\begin{aligned}\mathbf{E}\left\|\sum \varepsilon_i X_i\right\| &\leqslant \mathbf{E}\left\|\sum \varepsilon_i (X_i - X_i')\right\| = \mathbf{E}\left\|\sum (X_i - X_i')\right\| \\ &\leqslant \mathbf{E}\left\|\sum X_i\right\| + \mathbf{E}\left\|\sum X_i'\right\| \leqslant 2\,\mathbf{E}\left\|\sum X_i\right\|.\end{aligned}$$

□

A common use of the symmetrization technique is obtained by introducing the random variables ε_i, then conditioning on X_i. This reduces the problem to a statement purely about Bernoulli random variables, which are often simpler to reason about. More generally, we can prove variants of this technique for convex functions, which in particular means we can apply symmetrization when using moment generating function bounds.

Theorem 4.8. *Let $F : [0, \infty) \to \mathbf{R}$ be an increasing, convex function. Then*

$$\mathbf{E}\,F\left(0.5 \cdot \left\|\sum \varepsilon_i X_i\right\|\right) \leqslant \mathbf{E}\,F\left(\left\|\sum X_i\right\|\right) \leqslant \mathbf{E}\,F\left(2 \cdot \left\|\sum \varepsilon_i X_i\right\|\right)$$

Proof. The argument is the same symmetrization technique as before, since sup is a convex function. □

Later on, we discuss suprema of random processes. We can use symmetrization in this setting as well.

Lemma 4.9. *Let $X_1(t),\dots,X_n(t)$ be independant, mean zero random processes indexed by points $t \in T$. Let $\varepsilon_1,\dots,\varepsilon_n$ be independant, mean zero, symmetric Bernoulli random processes. Then*

$$0.5 \cdot \mathbf{E}\left(\sup_{t\in T}\sum \varepsilon_i X_i(t)\right) \leqslant \mathbf{E}\left(\sup \sum X_i(t)\right) \leqslant 2\,\mathbf{E}\left(\sum \varepsilon_i X_i(t)\right)$$

4.3 Contraction

There is one more useful inequality we discuss in this chapter, known as *contraction*. It works kind of like an l^1, l^∞ bound.

Theorem 4.10. *Let $x_1,\dots,x_n$ be vectors in some normed space, and let $a_1,\dots,a_n$ be real numbers. Then*

$$\mathbf{E}\left\|\sum a_i \varepsilon_i x_i\right\| \leqslant \|a\|_\infty \left\|\sum \varepsilon_i x_i\right\|.$$

Remark. We can also prove this identity for convex functions of the norm.

As an application, we can prove a version of symmetrization for Gaussian vectors.

Theorem 4.11. *Let $X_1,\dots,X_n$ be independant, mean zero random vectors in a norm space. Let $g_1,\dots,g_n \sim N(0,1)$ be independant vectors. Then*

$$\frac{1}{(\log n)^{1/2}}\,\mathbf{E}\left\|\sum g_i X_i\right\| \lesssim \mathbf{E}\left\|\sum X_i\right\| \leqslant 3\,\mathbf{E}\left\|\sum g_i X_i\right\|$$

Proof. s □

We can also prove a version of contraction which can be used in the study of random processes.

Theorem 4.12 (Talagrand's Contraction Principle)**.** *Let T be a bounded subset of $\mathbf{R}^n$, and let $\varepsilon_1,\dots,\varepsilon_n$ be independant symmetric Bernoulli random variables. Let $\phi : \mathbf{R} \to \mathbf{R}$ be contraction maps, i.e. Lipschitz functions with $|\phi(x-y)| \leqslant |x-y|$. Then*

$$\mathbf{E}\left(\sup_t\left(\sum \varepsilon_i \phi_i(t_i)\right)\right) \leqslant \mathbf{E}\left(\sup_t \sum \varepsilon_i t_i\right)$$

Chapter 5

Suprema of Random Processes

In this chapter we try and simultaneously control a family of random variables $\{X_t : t \in T\}$, where T is an arbitrary index set, allowed to be of infinite cardinality. An important example is bounding $\{X_t = g \cdot t : t \in T\}$, where T is a subset of $\mathbf{R}^n$, and $g \sim N(0,1)$. We will find that if the random variables $\{X_t\}$ are 'sufficiently continuous', then controlling $\sup_t X_t$ essentially only relies on studying the geometry of the index set T. Our main focus will be on *Gausssian processes*, i.e. processes $\{X_t\}$ such that all finite dimensional subdistributions are normal, or equivalently, if $\sum a_i X_{t_i}$ is normally distributed for any finite sum of $t_i \in T$ and $a_i \in \mathbf{R}$.

In our analysis, we make the simplifying assumption that the random process we study is centered, so $\mathbf{E}(X_t) = 0$ for all $t \in T$. Then we can define the covariance function $\Sigma(t,s) = \mathbf{E}(X_t X_s)$. The *increments* of the random process are defined as $d(t,s) = \|X_t - X_s\|_2$. These incremeents naturally give T the structure of a metric space, with $d(t,s) = 0$ if and only if $X_t = X_s$.

Example. *Let $\{X_t : t \in RR\}$ be a Brownian motion. The metric induced on $\mathbf{R}$ is given by $d(t,s) = \sqrt{t-s}$. Similarily, if we consider independant normal random variables $Z_1, Z_2, \ldots$ and set $S_n = Z_1 + \cdots + Z_n$, then $\{S_n\}$ is a process defined on $\mathbf{N}$, and $d(n,m) = \sqrt{n-m}$.*

The increments of a process and it's covariance function are tightly related. Indeed,

$$d(t,s)^2 = \mathbf{E}((X_t - X_s)^2) = \Sigma(t,t) + \Sigma(s,s) - 2\Sigma(t,s)$$

Conversely, if $X_0 = 0$ belongs to the process, then

$$\Sigma(t,s) = \frac{d(t,0)^2 + d(s,0)^2 - d(t,s)^2}{2}$$

Thus the two functions determine one another.

5.1 Slepian Inequality

A natural goal is to obtain a bound on $\mathbf{E}(\sup X_t)$. In all but the most basic process, this is a non-trivial task. The first bound we discuss enables us to replace the problem of bounding a process with bounding another process, whose supremum may be more easily calculated. Given two processes $\{X_t\}$ and $\{Y_t\}$, we say $\{Y_t\}$ *stochastically dominates* $\{X_t\}$ if for any $s \in \mathbf{R}$,

$$\mathbf{P}(\sup X_t \geqslant s) \leqslant \mathbf{P}(\sup Y_t \geqslant s)$$

The method we discuss is called *Slepian's inequality*, and gives conditions for a random process to be bound by another random process.

Our proof of the method will involve the method of *Gaussian interpolation*. Given two independant Gaussian vectors $X \sim N(0, \Sigma(X))$ and $Y \sim N(0, \Sigma(Y))$, we consider

$$Z_u = \sqrt{u} \cdot X + \sqrt{1-u} \cdot Y,$$

defined so that $\Sigma(Z_u) = u\Sigma(X) + (1-u)\Sigma(Y)$.

Lemma 5.1 (Gaussian Integration by Parts)**.** *Let $X \sim N(0,\Sigma)$. Then for any function $f : \mathbf{R}^n \to \mathbf{R}$, $\mathbf{E}(Xf(X)) = \Sigma \cdot \mathbf{E}(\nabla f(X))$.*

Proof. Assume first that f has bounded support. If $p(x)$ is the density function of X, then

$$\begin{aligned}
\Sigma \cdot \mathbf{E}(\nabla f(X)) &= \mathbf{E}(\Sigma \cdot \nabla f(X)) \\
&= \int \Sigma \cdot (\nabla f) \cdot p \, dx \\
&= -\int \Sigma \cdot \nabla p \cdot f(x)
\end{aligned}$$

Note that $(\nabla p)(x) = -p(x) \cdot \Sigma^{-1} \cdot x$. Thus

$$-\int \Sigma \cdot \nabla p \cdot f(x) = \int p(x) f(x) x = \mathbf{E}(Xf(X)). \qquad \square$$

Lemma 5.2. *Let $X \sim N(0, \Sigma(X))$ and $Y \sim N(0, \Sigma(Y))$ be two independant Gaussian vectors. Let $Z_u = \sqrt{u} \cdot X + \sqrt{1-u} \cdot Y$. Then for any twice differentiable function $f : \mathbf{R}^n \to \mathbf{R}$,*

$$\frac{d\,\mathbf{E}[f(Z_u)]}{du} = 0.5 \sum_{i,j} \left(\Sigma(X)_{ij} - \Sigma(Y)_{ij}\right) \mathbf{E}\left(\frac{\partial f}{\partial x_i\, \partial x_j}(Z_u)\right)$$

Proof. Using the chain rule, we find

$$\begin{aligned}\frac{d\,\mathbf{E}[f(Z_u)]}{du} &= \sum \mathbf{E}\left[\frac{\partial f}{\partial z_i}(Z_u)\frac{dZ_{u,i}}{du}\right] \\ &= \sum \mathbf{E}\left[\frac{\partial f}{\partial z_i}(Z_u)\left(\frac{X_i}{2\sqrt{u}} - \frac{Y_i}{2\sqrt{1-u}}\right)\right]\end{aligned}$$

If

$$g_i(X,Y) = \left(\frac{\partial f}{\partial z_i}\right)(\sqrt{u} \cdot X + \sqrt{1-u} \cdot Y) = \left(\frac{\partial f}{\partial z_i}\right)(Z_u),$$

then we can apply a Gaussian integration by parts to conclude

$$\begin{aligned}\mathbf{E}\left[X_i \left(\frac{\partial f}{\partial z_i}\right)(Z_u)\middle| Y\right] &= \mathbf{E}[X_i \cdot g_i(X,Y)|Y] \\ &= \sum_j \Sigma(X)_{ij}\, \mathbf{E}\left(\frac{\partial g_i}{\partial x_j}(X,Y)\middle| Y\right) \\ &= \sqrt{u} \sum_j \Sigma(X)_{ij}\, \mathbf{E}\left(\left(\frac{\partial^2 f}{\partial z_j\, \partial z_i}\right)(Z_u)\middle| Y\right).\end{aligned}$$

Then we can take expectations with respect to Y on both sides to remove the conditional expectation. Similarily, we calculate

$$\mathbf{E}\left[X_i \left(\frac{\partial f}{\partial z_i}\right)(Z_u)\right] = \sqrt{1-u} \sum_j \Sigma(Y)_{ij}\, \mathbf{E}\left(\left(\frac{\partial^2 f}{\partial z_j\, \partial z_i}\right)(Z_u)\right)$$

Putting these two terms together completes the calculation. □

Lemma 5.3. *If $f : \mathbf{R}^n \to \mathbf{R}$ is twice-differentiable and for all $i \neq j$,*

$$\frac{\partial^2 f}{\partial x_i\, \partial x_j} \geqslant 0.$$

Let X and Y be Gaussian vectors such that for all i, $\mathbf{E}\,X_i^2 = \mathbf{E}\,Y_i^2$, and for all indices i, j, $\mathbf{E}[(X_i - X_j)^2] \leqslant \mathbf{E}[(Y_i - Y_j)^2]$. Then $\mathbf{E}\,f(X) \geqslant \mathbf{E}\,f(Y)$.

Proof. Note that if $\Sigma(X)$ and $\Sigma(Y)$ are the covariance matrices of X and Y, then $\Sigma(X)_{ii} = \Sigma(Y)_{ii}$, and $\Sigma(X)_{ij} \geqslant \Sigma(Y)_{ij}$. If $\Pi(X,Y)_{ij} = \mathbf{E}(X_i Y_j)$, then the vector $Z = (X, Y)$ is Gaussian, and

$$\Sigma(Z) = \begin{pmatrix} \Sigma(X) & \Pi(X,Y) \\ \Pi(X,Y) & \Sigma(Z) \end{pmatrix}$$

We can assume that X and Y are independant, because the inequalities we need to prove only rely on the individual distributions of each random variable. Then the last lemma implies that

$$\frac{d\,\mathbf{E}[f(Z_u)]}{du} \geqslant 0,$$

so $\mathbf{E}[f(Z_u)]$ is increasing in u. But this means that $\mathbf{E}\,f(X) \geqslant \mathbf{E}\,f(Y)$. □

Theorem 5.4 (Slepian's Inequality)**.** *Let $\{X_t\}$ and $\{Y_t\}$ be two mean zero processes. Assume that for all t, s, $\mathbf{E}\,X_t^2 = \mathbf{E}\,X_s^2$ and $d_X(t,s) \leqslant d_Y(t,s)$ for all t and s. Then for any u,*

$$\mathbf{P}\,(\sup X_t \geqslant u) \leqslant \mathbf{P}\,(\sup Y_t \geqslant u)$$

and consequently, $\mathbf{E}(\sup X_t) \leqslant \mathbf{E}(\sup Y_t)$.

Proof. We use the techniques of *Gaussian interpolation*. Let $h : \mathbf{R} \to [0,1]$ be a twice-differentiable, non-increasing approximation of the indicator function $\mathbf{I}(x < s)$. Then the function $f : \mathbf{R}^n \to [0,1]$ defined by $f(x) = h(x_1)\ldots h(x_n)$ is an approximation of $\mathbf{I}(\max(x_1,\ldots,x_n) < s)$. Then for $i \neq j$

$$\frac{\partial^2 f}{\partial x_i\,\partial x_j} = h'(x_i)h'(x_j)\prod\prod\{h(x_k) : k \notin \{i,j\}\} \geqslant 0.$$

Thus the last lemma implies $\mathbf{E}\,h(X) \geqslant \mathbf{E}\,h(Y)$. But since h was essentially arbitrary, we conclude

$$\begin{aligned} \mathbf{P}(\max(X_1,\ldots,X_n) < s) &= \mathbf{E}\,\mathbf{I}(\max(X_1,\ldots,X_n) < s) \\ &\geqslant \mathbf{E}\,\mathbf{I}(\max(Y_1,\ldots,Y_n) < s) \\ &= \mathbf{P}(\max(Y_1,\ldots,Y_n) < s). \end{aligned}$$

Taking complements, we find

$$\mathbf{P}(\max(X_1,\ldots,X_n) \geqslant s) \leqslant \mathbf{P}(\max(Y_1,\ldots,Y_n) \geqslant s).$$

Both sides converge monotonely, so we conclude

$$\mathbf{P}(\sup X_t \geqslant s) \leqslant \mathbf{P}(\sup Y_t \geqslant s).$$

This gives the first part of Slepian's inequality. And now we find

$$\mathbf{E}(\sup X_t) = \int_0^\infty \mathbf{P}(\sup X_t \geqslant s)\, ds \leqslant \int_0^\infty \mathbf{P}(\sup Y_t \geqslant s)\, ds = \mathbf{E}(\sup Y_t). \quad \square$$

5.2 Sudakov-Fernique and Gordan's Inequalities

Sudakov-Fernique's theorem gives the expectation bound of Slepian's inequality, but works without the assumption of equality of variances.

Theorem 5.5 (Sudakov-Fernique)**.** *Let $\{X_t\}$ and $\{Y_t\}$ be mean zero Gaussian processes. If $d^X \leqslant d^Y$, then $\mathbf{E}\sup(X_t) \leqslant \mathbf{E}\sup(Y_t)$.*

Proof. It suffices to prove this theorem for Gaussian vectors X in $\mathbf{R}^n$, because the same limiting process as in Slepian's inequality proves the result in general. We also apply Gaussian interpolation. Given α

$$f_\alpha(x) = \frac{\log\left(\sum e^{\alpha x_i}\right)}{\alpha}$$

Then as $\alpha \to \infty$, $f_\alpha(x) \to \max(x_1,\ldots,x_n)$. Note that

$$\frac{d\,\mathbf{E}(f(Z_u))}{du} = \frac{1}{2}\sum_{i,j}(\Sigma(X)_{ij} - \Sigma(Y)_{ij})\,\mathbf{E}\left(\frac{\partial^2 f_\alpha}{\partial x_i\,\partial x_j}\right)$$

Let $p_i = \partial f/\partial x_i = e^{\alpha x_i}/\sum_k e^{\alpha x_k}$, so $p_1 + \cdots + p_n = 1$. We calculate

$$\frac{\partial^2 f_\alpha}{\partial x_i\,\partial x_j} = \alpha(\delta_{ij}p_i - p_i p_j)$$

where $\delta_{ij} = 1$ if $i = j$, and is zero otherwise. Now for any a_{ij},

$$\sum a_{ij}(\delta_{ij}p_i - p_i p_j) = \sum_{i\neq j}(a_{ii} + a_{jj} - 2a_{ij})p_i p_j.$$

Setting $a_{ij} = d^X(i,j)^2 - d^Y(i,j)^2$, we conclude

$$\frac{d\,\mathbf{E}(f_\alpha(Z_u))}{du} = -\sum_{i\neq j} a_{ij} p_i p_j \leqslant 0$$

Thus we find $\mathbf{E}(f_\alpha(X)) \leqslant \mathbf{E}(f_\alpha(Y))$. We note that

$$\max(x_1,\ldots,x_n) \leqslant f_\alpha(x) \leqslant \max(x_1,\ldots,x_n) + \frac{\log(n)}{\alpha},$$

so

$$\mathbf{E}\,|f_\alpha(X) - \max(X_1,\ldots,X_n)| = \mathbf{E}\,f_\alpha(X) - \max(X_1,\ldots,X_n) \leqslant \frac{\log n}{\alpha}.$$

Thus $f_\alpha(X) \to \max(X_1,\ldots,X_n)$ in L^1 norm. Similarily, $f_\alpha(Y) \to \max(Y_1,\ldots,Y_n)$ in the L^1 norm. So we find $\mathbf{E}(\max(X_1,\ldots,X_n)) \leqslant \mathbf{E}(\max(Y_1,\ldots,Y_n))$. And now we take limits on both sides to obtain the full theorem. □