High Dimensional Probability

Jacob Denson

November 4, 2022

Table Of Contents

1 Introduction 2

2 Concentration In High Dimensional Spaces 5
2.1 Isotropic Vectors . . . 8
2.2 Uniformly Subgaussian Vectors . . . 11
2.3 Concentration For Lipschitz Functions . . . 14
2.4 The Johnson-Lindenstrauss Lemma . . . 20

3 Useful Techniques 23
3.1 Decoupling . . . 23
3.2 Symmetrization . . . 30
3.3 Matrix Completion . . . 32
3.4 Contraction . . . 32

4 Chaining 35
4.1 Covering Numbers . . . 36
4.2 Chaining . . . 38
4.3 Matrix Concentration . . . 39
4.4 Empirical Processes . . . 47
4.5 Vapnik-Chervonenkis Dimension . . . 49
4.6 Generic Chaining . . . 57
4.7 Chevet's Inequality . . . 59

5 Gaussian Processes 60
5.1 Slepian Inequality . . . 61
5.2 Sudakov-Fernique Inequality . . . 64
5.3 Gaussian Width . . . 66
5.4 Stable Dimension . . . 70

5.5 Sudakov's Minorization Inequality 71
5.6 Two Sided Sudakov Inequality 72
5.7 Diameters of Random Projections 72

6 Deviations of Random Matrices 75

7 Applications of Random Processes 77

8 Applications to Random Graphs 80

9 Applications of High Dimensional Concentration 82
9.1 Community Detection 82
9.2 Covariance Estimation 84

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
>
> $$\mathbf{P}\left(\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbf{E}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) \geqslant t\right)$$
>
> decaying fast as t increases. If $\{X_k\}$ are independant and *subgaussian*, 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
>
> $$\mathbf{P}\left(|f(X_1,\dots,X_n) - \mathbf{E}f(X_1,\dots,X_n)| \geqslant t\right)$$

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, calculating $\mathbf{E}f(X_1,\dots,X_n)$ requires different techniques depending on the function f.

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 a family of random variables $\{X_1,\dots,X_n\}$ 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 least 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 a very simple example, if $\{X_1,\dots,X_n\}$ 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. For instance, we know that if $f(E_1, \dots, E_n)$ is 'sufficiently symmetric' and 'sufficiently monotone', with $\{E_k\}$ events depending on a probability p, then $f(E_1, \dots, E_n)$ undergoes a 'sharp transition'.

We often use asymptotic notation. We write $A \lesssim B$, for two positive quantities A and B which depend on various parameters, if $A \leqslant C \cdot B$ for some positive constant C. All constants involved in our results are *effective*, in the sense that one can feasibly calculate them by following through our arguments with some patience, and none are of exceedingly large magnitude. We also refer the reader to the Section on concentration bounds in my notes on probability theory, which will prove very useful in the sequel.

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 when n is large, $|X|$ is close to $\sqrt{n}$ with high probability. 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, with $\|X_i\|_{\psi_2} \leqslant K$ and with $\mathbf{E}(X_i^2) = 1$ for each $i \in \{1, \dots, n\}$. Then*

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

where the implicit constant is universal, independant of n.

Proof. Without loss of generality, we assume $K \geqslant 1$. In general, since for each i,

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

we know $K \gtrsim 1$. Thus if $K \leqslant 1$, we can apply our current proof 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. This gives 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 $|X|$. Given any $t \geqslant 0$, if $u = \min(t, t^2)^{1/2}$, then $t = \max(u, u^2)$. Thus we have shown

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

If u is fixed, and $||X| - n^{1/2}| \geqslant u$, then we conclude

$$||X|^2 - n| = ||X| - n^{1/2}|||X| + n^{1/2}| \geqslant u \cdot (|X| + n^{1/2}).$$

We either have $|X| \geqslant n^{1/2} + u$, or $|X| \leqslant n^{1/2} - u$. The former case implies

$$||X|^2 - n| \geqslant u \cdot (2n^{1/2} + u) \geqslant \max(u, u^2).$$

If the latter case holds, we must have $u \leqslant n^{1/2}$, so

$$||X|^2 - n| \geqslant u \cdot n^{1/2} \geqslant \max(u, u^2).$$

In both cases, $||X|^2 - n| \geqslant \max(u, u^2)$. Thus for any $u \geqslant 0$,

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

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

This theorem contradicts our intuitions from low dimensional probability. We would expect a standard normal distribution in $\mathbf{R}^n$ to lie near the origin, since that is where the density function is largest. But the volume near the origin is negligible in high dimensions, which means that in fact, a normal distribution is not likely to lie near the origin at all, and for most purposes acts distributionally the same as a uniformly random vector chosen on the sphere of radius $n^{1/2}$.

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

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

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

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

Applying the triangle inequality now shows 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 $|\mathbf{E}|X| - n^{1/2}|$, which gives the result. The variance bound then follows easily, because

$$\mathbf{V}\,|X| = \mathbf{V}(|X| - n^{1/2}) \leqslant \mathbf{E}[(|X| - n^{1/2})^2] \lesssim \||X| - n^{1/2}\|_{\psi_2}^2 \lesssim K^4. \qquad \square$$

Later on, we will show that if A is an $m \times n$ matrix, and X is as in Theorem 2.1, then

$$\||AX| - \|A\|_F\|_{\psi_2} \lesssim K^2 \cdot \|A\|$$

where $\|A\|_F$ is the Frobenius norm of A, and $\|A\|$ it's operator norm. Taking A as the identity transformation yields Theorem 2.1 as a special case.

2.1 Isotropic Vectors

We say a random vector X in $\mathbf{R}^n$ is **isotropic** if it's second moment matrix $\Sigma(X) = \mathbf{E}(XX^T)$ is the identity matrix. Note that for any $x \in \mathbf{R}^n$,

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

Thus being isotropic is equivalent to saying $\mathbf{E}((X \cdot x)^2) = |x|^2$. Thus the vector X is on average extended evenly in all directions.

It is often natural to assume random vectors under analysis are centered, and isotropic. And often one can reduce to this case. For any random vector X with mean μ, the random vector $Y = \text{Cov}(X)^{-1/2}(X - \mu)$ is centered and isotropic. To see this, we may assume without loss of generality that $\mu = 0$. Then $\text{Cov}(X)^{-1/2} = \Sigma(X)^{-1/2}$, and so for any $y \in \mathbf{R}^n$,

$$\begin{aligned}\Sigma(Y) = \mathbf{E}(YY^T) &= \Sigma(X)^{-1/2} \cdot \mathbf{E}(XX^T) \cdot \Sigma(X)^{-1/2} \\ &= \Sigma(X)^{-1/2} \cdot \Sigma(X) \cdot \Sigma(X)^{-1/2} = I_n.\end{aligned}$$

If $\Sigma(X)$ is degenerate, then X almost surely lies on a lower dimensional subspace of $\mathbf{R}^n$, and we can then reduce our analysis to this lower dimensional subspace, where the corresponding covariance matrix is non-degenerate.

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 use the cycle property of the trace to write

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

Next, given Y, we find

$$\mathbf{E}((X \cdot Y)^2 | Y) = \sum_{i,j} Y_i Y_j \mathbf{E}(X_i X_j | Y) = \sum_{i,j} Y_i Y_j \mathbf{E}(X_i X_j) = \sum_i Y_i^2 = |Y|^2.$$

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. Lemma 2.3 implies that for any two independant isotropic random vectors X and Y,

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

we can expect that $X \cdot Y = O(n^{1/2})$ with high probability. But combined with the fact that $|X|$ and $|Y|$ are $O(n^{1/2})$ with high probability, this means that with high probability

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

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

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 $\mathbf{E}|X_i|^2 = \mathbf{E}|X_j|^2$ for $i \neq j$, this implies $\mathbf{E}|X_i|^2 = 1$ for each i. Thus X is isotropic.*

It is good to remember that the coordinates of an isotropic vector need not be independant. In the example above, X is isotropic, but it's coordinates are certainly not independant, since the coordinates must satisfy the equation $X_1^2 + \cdots + X_n^2 = 1$.

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.*

The next example shows that isotropic random variable need not be centered either.

Example. *Pick a random vector X uniformly at random from the set*

$$\{n^{1/2} \cdot e_1, \dots, n^{1/2} \cdot e_n\}.$$

Then for each index i, X_i^2 is $\{0, n\}$ 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** in $\mathbf{R}^n$ is a set $\{v_1, \dots, v_m\}$ for which there are positive constants A and B for which the approximate Parseval's identity $A|x|^2 \leqslant \sum (v_i \cdot x)^2 \leqslant B|x|^2$ holds for all vectors $x \in \mathbf{R}^n$. if $A = B$, the frame is called **tight**. A frame is tight if and only if $\sum v_i^T v_i = AI_n$, and more generally, it is a frame with constants A and B if and only if $AI_n \preceq \sum v_i^T v_i \preceq BI_n$.

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

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

then

$$v_1^T v_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad v_2^T v_2 = \begin{pmatrix} 1/4 & -\sqrt{3}/4 \\ -\sqrt{3}/4 & 3/4 \end{pmatrix}$$

$$\textit{and} \quad v_3^T v_3 = \begin{pmatrix} 1/4 & \sqrt{3}/4 \\ \sqrt{3}/4 & 3/4 \end{pmatrix}.$$

This means $v_1^T v_1 + v_2^T v_2 + v_3^T v_3 = 3/2 \cdot I_2$, so the frame is tight.

Theorem 2.4. *If $\{u_1, \dots, u_m\}$ is a tight frame, 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 $\sqrt{p_i} \cdot u_i$ is a tight frame with $A = 1$.*

Proof. If $X = (m/A)^{1/2} \cdot \text{Uniform}(u_1, \dots, u_n)$, 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 Uniformly 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 if $|x| = 1$, then

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

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*

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

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. *There exists a universal constant c such that if X is a discrete, isotopic random vector in $\mathbf{R}^n$ with support S, and $\|X\|_{\psi_2} \leqslant 1$, then $|S| \geqslant \exp(cn)$.*

Proof. Suppose that $|s| \leqslant A \cdot n^{1/2}$ for all $s \in S$, for some constant A. Note that

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

Since $\|X \cdot s\|_{\psi_2} \leqslant An^{1/2}$, we can use the expectation bound on the supremum of random variables to conclude

$$n = \mathbf{E}|X|^2 \leqslant \mathbf{E}\left(\sup_{s \in S} |X \cdot s|\right) \lesssim \left\| \sup_{s \in S} |X \cdot s| \right\|_{\psi_2} \lesssim A \cdot \sqrt{n \log |S|}.$$

Thus there exists a constant c such that $|S| \geqslant \exp(cn/A^2)$.

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 by the concentration of norm that there exists a small universal constant c such that

$$\mathbf{P}(||Z| - \sqrt{n}| \geqslant t/2) \leqslant 2\exp(-ct^2) \quad \text{and} \quad \mathbf{P}(Z_1 \geqslant t) \leqslant \exp(-ct^2).$$

Applying a union bound shows that

$$\begin{aligned}\mathbf{P}(X_1 \geqslant t/\sqrt{n}) &= \mathbf{P}(Z_1/|Z| \geqslant t/\sqrt{n})\\ &\leqslant \mathbf{P}(||Z| - \sqrt{n}| \geqslant t/2) + \mathbf{P}(Z_1/|Z| \geqslant t/\sqrt{n}, ||Z| - \sqrt{n}| \leqslant t/2)\\ &\leqslant \mathbf{P}(||Z| - \sqrt{n}| \geqslant t/2) + \mathbf{P}(Z_1 \geqslant t)\\ &= 2\exp(-ct^2) + \exp(-ct^2) = 3\exp(-ct^2).\end{aligned}$$

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 uniformly chosen on $n^{1/2} \cdot S^{n-1}$, then $\|X\|_{\psi_2} \lesssim 1$.*

If $x_1, \dots, x_m$ are values with $x_1^2 + \dots + x_m^2 = 1$, and X is uniformly distribution on the sphere of radius $n^{1/2}$ in $\mathbf{R}^n$, then $x_1 X_1 + \dots + x_n X_n$ looks like a $N(0,1)$ distribution when $n \gg m$. This observation is known as the *projective* central limit theorem. Corollary 2.7 gives the tail decay aspect of this theorem.

One may conjecture that a uniformly random vector lying on an isotropic convex body is subgaussian, independant of the body and the dimension n. But this not be the case.

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 $K = -K$, $\mathbf{E}(X_i X_j) = 0$ 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 from a result known as 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 consider the situation when the values of X lie in some metric space M satisfying an 'isoperimetric blowup' phenomenon with respect to the distribution of X. In such spaces, our result gives concentration bounds for Lipschitz functions $f : M \to \mathbf{R}$. Our main result is for Lipschitz functions on the sphere, but we also indicate concentration results using this method on other spaces.

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 phenomenon for the thickenings of sets on spheres.

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 implies that $\sigma(E_t) \geqslant \sigma(H_t)$. Thus we have reduced lower bounding the surface area of E_t to lower bounding the surface area of H_t.

Consider $x \in S^{n-1}$ with $x_1 \leqslant 2^{-1/2} \cdot t$. Set $a = (x_2^2 + \cdots + x_n^2)^{1/2}$. Then $\sqrt{1 - t^2/2} \leqslant a \leqslant 1$. If we set $x' = (0, x_2/a, \ldots, x_n/a) \in H$, then

$$\begin{aligned} |x - x'|^2 = |x_1|^2 + (1-a)^2 &\leqslant t^2/2 + \left(1 - (1 - t^2/2)^{1/2}\right)^2 \\ &= 2\left(1 - (1 - t^2/2)^{1/2}\right) = \frac{2t^2}{1 + (1 - t^2/2)^{1/2}} \leqslant t^2. \end{aligned}$$

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 there exists a small constant c such that

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

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. □

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 distributed on S^{n-1}, and $f : S^{n-1} \to \mathbf{R}$, then*

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

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

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

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

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

This means that

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

Similarily, we can show

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

A union bounds then shows

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

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

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

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

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

Thus $\|X - M\|_{\psi_2}$ and $\|X - \mathbf{E}X\|_{\psi_2}$ are comparable to one another.

Example. *For a geometric application of this claim, 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 a set of exponentially many almost orthogonal vectors inductively. Consider unit vectors $e_1, \dots, e_N$, 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 $N < 0.5 \cdot \exp(cn\varepsilon^2)$, 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 continue adding almost orthogonal vectors up to the point where $N \geqslant 0.5 \cdot \exp(cn\varepsilon^2)$.*

There is nothing really special to the sphere here. Given any other measure space with a metric, we can consider the minimizers of volume expansion, and thus achieve isoperimetric inequalities in this domain. If the minimizers of the isoperimetry problem have a mass blow up, we can obtain the same result.

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 distributions 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 $x, y \in \{-1,1\}^n$, $d(x,y) = \#\{i : x_i \neq y_i\}$ in this domain are balls with respect to the Hamming distance. We can conclude from this that*

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

for Lipschitz functions on this domain.

Example. *If we consider the Hamming distance on S_n, i.e. for two permutations π and η, we let $d(\pi, \eta) = \#\{i : \pi(i) \neq \eta(i)\}$, and consider the uniform distribution on S_n, then the minimizers of volume expansion in this domain are given by balls, and so we also conclude that*

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

for Lipschitz functions on this domain.

Using other tools, we can establish semigroup results in other spaces.

Example. *If M is a Riemannian manifold, we can consider the arclength distance, as well as a natural 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 semigroup tools enable us to establish 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. Here, we also obtain that

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

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 Haar measure induced by the action of $SO(n)$ on the space, i.e.*

$$A \cdot V = A \cdot V \cdot A^{-1}$$

Since this action is transitive, we can actually realize G_{nm} as a quotient of $SO(n)$. The stabilizer of the action is $SO(n) \times SO(n-m)$, and we find

$$G_{nm} \equiv \frac{SO(n)}{SO(m) \times SO(n-m)}.$$

Using this, we can also deduce concentration bounds on the Grassmanian from concentration bounds on the orthogonal group.

Example. *Let $\Phi(x)$ denote the cumulative distribution function of a normal distribution. If $Z \sim N(0, I_n)$, then $\phi(Z) = (\Phi(Z_1), \dots, \Phi(Z_n))$ 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* $\|f(X) - \mathbf{E}(f(X))\|_{\psi_2} \lesssim \|f\|_{Lip}$.

There is some results which are distribution dependant, but slightly more stringent conditions need to be satisfied.

Theorem 2.11 (Talagrand's Concentration Inequality)**.** *Let X be a random vector in $\mathbf{R}^n$, with independant coordinates $\{X_i\}$, with $|X_i| \leqslant 1$ almost surely. Then for any convex Lipschitz function $f : [-1,1]^n \to \mathbf{R}$,*

$$\|f(X) - \mathbf{E} f(X)\|_{\psi_2} \lesssim \|f\|_{Lip}.$$

2.4 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 2.12. *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 2.13. *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. $\square$

Remark. It is an important, and incredible 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 $1/m^{1/2}$.

The Johnson Lindenstrauss lemma is tight for general data. For instance, if X is an orthonormal basis in $\mathbf{R}^n$, and $Q : \mathbf{R}^n \to \mathbf{R}^m$ is an almost isometry on this set of points, then for each $x, y \in X$,

$$\sqrt{2} \cdot (1-\varepsilon) \leqslant |Qx - Qy| \leqslant \sqrt{2} \cdot (1+\varepsilon)$$

Thus $\{Q(x) : x \in X\}$ is a $\sqrt{2}\cdot(1-\varepsilon)$ packing in the ball B of radius $\sqrt{2}\cdot(1+\varepsilon)$ in $\mathbf{R}^n$. A volumetric argument shows that

$$\begin{aligned} n \leqslant P(Q(X), \sqrt{2}\cdot(1-\varepsilon)) &\leqslant P\left(B, \sqrt{2}\cdot(1-\varepsilon)\right) \\ &\leqslant \frac{(1+\varepsilon)^m[2^{1/2}+1/2^{1/2}]^m}{2^{-m/2}\cdot(1-\varepsilon)^m} = \left(1+2\cdot\frac{\varepsilon}{1-\varepsilon}\right)3^m, \end{aligned}$$

where $P(T,\varepsilon)$ is the ε *packing number* of T, discussed later on in these notes. We conclude $m \gtrsim_\varepsilon \log n$.

Chapter 3

Useful Techniques

Here we discuss some very useful techniques for reducing the analysis of certain distributions to other situations. Decoupling enables us to reduce the analysis of a random quadratic form to a random bilinear form, which is much easier to understand. Symmetrization enables us to reduce the study of a certain distribution to the study of a random distribution.

3.1 Decoupling

In this section, we study the distribution of quadratic forms

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

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

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

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 $X^T A X$ with $X^T A X'$, where X' is an independant copy of X. Since bilinear forms are much easier to analyze than quadratic forms, this is often a useful reduction.

Lemma 3.1. *Let Y and Z be independant random variables with $\mathbf{E} Z = 0$. Then for any convex function $F : \mathbf{R}^n \to \mathbf{R}$, $\mathbf{E} F(Y) \leqslant \mathbf{E} F(Y + Z)$.*

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

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

We now use this lemma to establish a decoupling inequality.

Theorem 3.2. *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 $\delta_1, \dots, \delta_n \in \{0,1\}$ be independant symmetric Bernoulli random variables, and define a random subset $I = \{k : \delta_k = 1\}$ of $\{1, \dots, n\}$. Since $\mathbf{E}(\delta_i(1-\delta_j)) = 1/4$ for $i \neq j$,

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

We now apply the function F to both sides, calculating

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

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 $\sum_{(i,j)\in I} A_{ij}X_iX_j$ is identically distributed to $\sum_{(i,j)\in I} A_{ij}X_iX_j'$, 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* $\{X_i\}_{i\in I^c}$ and $\{X_j'\}_{j\in I}$. 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 $Y + Z_1 + Z_2 = \sum A_{ij}X_iX_j$. $\square$

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. This is quite often a useful technique when upper bounding certain monotonic quantities which depend on subgaussian variables.

Lemma 3.3. *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 3.4. *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

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

We can now prove a concentration inequality for quadratic forms, which should be viewed as analogoue to Bernstein's inequality in the linear case.

Theorem 3.5 (Hanson-Wright). *Let X be a random vector with independant, mean zero, subgaussian coordinates. Then for $t \geqslant 0$,*

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

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

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

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 $S = \sum_{i\neq j} a_{ij}X_iX_j$, then our decoupling bound implies that provided $\lambda \leqslant c/\|A\|$.

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

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

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

As a consequence, we can obtain concentration bounds for the norms of more general subgaussian distributions.

Theorem 3.6. *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 $A = B^TB$. Then $X^TAX = |BX|^2$, and $\mathbf{E}\,X^TAX = \|B\|_F^2$. Note that $\|A\| = \|B\|^2$, and

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

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

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

Example. *Let E be a subspace of $\mathbf{R}^n$ of dimension m. let X be a random vector in $\mathbf{R}^n$ with independant, mean zero, unit variance, sub-gaussian coordinates. If we let P denote orthgononal projection onto E, then $d(X,E) = |X - P(X)|$. Since the singular value decomposition of $I - P$ consists of $n - m$ copies of 1, and m copies of 0, we conclude $\|I - P\| = 1$, and $\|I - P\|_F = \sqrt{n-m}$. Thus the Hardy-Wright inequality implies $\|d(X,E) - \sqrt{n-m}\|_{\psi_2} \lesssim K^2$, where $K = \max \|X_i\|_{\psi_2}$.*

We can also use decoupling to obtain tail bounds on the norms of subgaussian vectors which may not necessarily have independant coordinates.

Lemma 3.7. *There exists a universal constant C and c such that if X is a mean zero subgaussian vector, with $\|X\|_{\psi_2} \leqslant K$, then*

$$\mathbf{E}\exp(\lambda^2|BX|^2) \leqslant \exp(CK^2\lambda^2\|B\|_F^2) \quad \textit{provided} \quad |\lambda| \leqslant \frac{c}{K\|B\|}.$$

Proof. For each $y \in \mathbf{R}^m$, $y \cdot BX = B^T y \cdot X$. Thus we find $\|y \cdot BX\|_{\psi_2} \leqslant K|B^T y|$, and so there exists a universal constant C' such that

$$\mathbf{E}(\exp(\lambda y \cdot BX)) \leqslant \exp(C'\lambda^2 K^2 |B^T y|^2).$$

If $g \sim N(0, I_m)$ is independant of X, then

$$\mathbf{E}_X(\exp(\lambda g \cdot BX)) \leqslant \exp(C'\lambda^2 K^2 |B^T g|^2).$$

Taking expectations on both sides yields that

$$\mathbf{E}(\exp(\lambda g \cdot BX)) \leqslant \mathbf{E}(\exp(C'\lambda^2 K^2 |B^T g|^2)).$$

On the other hand,

$$\mathbf{E}_g(\exp(\lambda g \cdot BX)) = \exp(\lambda^2 |BX|^2/2)$$

Taking expectations on both sides of this equality, and combining it with the previous inequality yields that

$$\mathbf{E}\exp(\lambda^2|BX|^2) \leqslant \mathbf{E}(2C'\lambda^2K^2|B^Tg|^2)$$

Thus we have reduced our analysis to the understanding of a normal distribution.

Now consider the singular value decomposition of B^T, i.e. we can write $B^T = \sum s_i u_i v_i^T$. Thus for any $\gamma > 0$,

$$\begin{aligned}\mathbf{E}(\exp(\gamma^2|B^Tg|^2)) &= \prod \mathbf{E}(\exp(\gamma^2 s_i^2 (g \cdot v_i)^2)) \\ &= \prod (1 - 2\gamma^2 s_i^2)^{-1/2}\end{aligned}$$

Provided $\gamma^2 s_i^2 < 1/2$ for all i, which is equivalent to saying $\gamma \leqslant 2^{-1/2}\|B\|^{-1}$. Since $\log(1-x) \geqslant -2x$ if $x \leqslant 1/2$, we conclude that if $\gamma^2 s_i^2 < 1/4$,

$$(1-2\gamma^2 s_i^2)^{-1/2} = \exp(-\log(1-2\gamma^2 s_i^2)/2) \leqslant \exp(2\gamma^2 s_i^2)$$

and so

$$\mathbf{E}(\exp(\gamma^2|B^Tg|^2)) \leqslant \prod \exp(2\gamma^2 s_i^2) = \prod \exp(2\gamma^2 \sum s_i^2) = \exp(2\gamma^2\|B\|_F^2).$$

This completes the proof. □

Theorem 3.8. *Let B be an $m \times n$ matrix, and X a mean zero sub-gaussian random variable with $\|X\|_{\psi_2} \leqslant K$. Then there exists a universal constant C such that*

$$\mathbf{P}\left(|BX| \geqslant CK\|B\|_F + t\right) \leqslant \exp\left(\frac{-ct^2}{K^2\|B\|^2}\right).$$

Proof. We know that, if C and c are defined as in the last lemma, then

$$\mathbf{E}\left(\exp\left(\frac{c^2|BX|^2}{K^2\|B\|^2}\right)\right) \leqslant \exp\left(\frac{c^2CK^2\|B\|_F^2}{K^2\|B\|^2}\right).$$

Thus

$$\mathbf{P}(|BX|^2 - CK^2\|B\|_F^2 \geqslant t^2) \leqslant \exp\left(\frac{-c^2t^2}{K^2\|B\|^2}\right).$$

But now we find

$$\mathbf{P}(|BX| - C^{1/2}K\|B\|_F \geqslant t) \leqslant \mathbf{P}(|BX|^2 - CK^2\|B\|_F^2 \geqslant t^2) \leqslant \exp\left(\frac{-c^2t^2}{K^2\|B\|^2}\right). \square$$

Remark. This is roughly as good as we can get without assuming the coordinates of X are independant. Consider the random vector $X = 2^{1/2} \cdot \phi g$, where $g \sim N(0, I_n)$ is Gaussian, and $\phi \in \{0,1\}$ is a Bernoulli distribution, and these distributions are both independant of one another. Then it is easy to verify that X is isotropic, mean zero, with $\|X\|_{\psi_2} \lesssim 1$, yet $\mathbf{P}(|X| = 0) = 1/2$ and

$$\mathbf{P}\left(|X| \geqslant \sqrt{2n}\right) = \mathbf{P}(|g|^2 \geqslant n)/2 \geqslant 1/2 - o(1).$$

Thus $|X|$ does not concentrate near $\sqrt{n}$, as suggested by the independant coordinate result.

3.2 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 3.9. *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. The proof is exactly the same as above, combined with an application of Jensen's inequality.

Theorem 3.10. *Let $F : [0,\infty) \to \mathbf{R}$ be an increasing, convex function. Then*

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

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

Lemma 3.11. *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*

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

Proof. The maximum of a set of numbers is a convex function, so if T is finite, this claim is already proved. And if we interpret the expected supremum as the supremum of the expected supremum of finite subsets of indices, this inequality remains true under taking a limit. □

We can also remove the mean zero assumption in some circumstances.

Theorem 3.12. *Let $X_1, \dots, X_n$ be independant, mean zero random vectors in a normed space. Then*

$$\mathbf{E}\left\|\sum X_i - \mathbf{E} X_i\right\| \leqslant 4\,\mathbf{E}\left\|\sum \varepsilon_i X_i\right\|,$$

where $\{\varepsilon_i\}$ are i.i.d and symmetric Bernoulli.

Proof. Since $\sum X_i' - \mathbf{E} X_i'$ has mean zero, where $\{X_i\}$ and $\{X_i'\}$ are independant and identically distributed,

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

We can now apply the standard symmetrization lemma to $X_i - X_i'$ to yield

$$\mathbf{E}\left\|\sum X_i - X_i'\right\| \leqslant 2\,\mathbf{E}\left\|\sum \varepsilon_i X_i - \varepsilon_i X_i'\right\| \leqslant 4\,\mathbf{E}\left\|\sum \varepsilon_i X_i\right\|.$$

TODO: Can we improve the 4 to a 2? □

3.3 Matrix Completion

3.4 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 3.13. *Let $x_1, \dots, x_n$ be vectors in some normed space, and let $a_1, \dots, a_n$ be real numbers. Then*

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

Proof. Without loss of generality, assume $\|a\|_\infty = 1$. The map $f(a) = \mathbf{E}\left\|\sum a_i \varepsilon_i x_i\right\|$ is a convex function for $\mathbf{R}^n$ to $\mathbf{R}$. The maximum principle for convex functions tells us the maximum of $f(a)$ for all a with $\|a\|_\infty \leqslant 1$ is attained at an extreme point of the set, i.e. at a value of a with $a_i = \pm 1$ for all i. For this point, $a_i \varepsilon_i$ has the same distribution as ε_i, and so

$$f(a) = \mathbf{E} \left\| \sum \varepsilon_i x_i \right\|$$

which completes the proof. □

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

We can also contract on general distributions.

Theorem 3.14. *Let $X_1, \dots, X_n$ be mean zero random vectors in a normed space, and let $a \in \mathbf{R}^n$. Then*

$$\mathbf{E} \left\| \sum a_i X_i \right\| \leqslant 4 \|a\|_\infty \mathbf{E} \left\| \sum X_i \right\|$$

Proof. We apply symmetrization together with the last contraction principle to conclude

$$\mathbf{E} \left\| \sum a_i X_i \right\| \leqslant 2\, \mathbf{E} \left\| \sum a_i \varepsilon_i X_i \right\| \leqslant 2 \|a\|_\infty \mathbf{E} \left\| \sum \varepsilon_i X_i \right\| \leqslant 4 \|a\|_\infty \mathbf{E} \, \|X_i\|. \qquad \square$$

As an application, we can prove a version of symmetrization which introduces Gaussians into a random sum, rather than symmetric Bernoulli distributions.

Theorem 3.15. *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. We first perform normal symmetrization, so that

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

But now, since $\mathbf{E}|g_i| = (2/\pi)^{1/2}$, we can apply Jensen's inequality to conclude

$$\begin{aligned} \mathbf{E} \left\| \sum \varepsilon_i X_i \right\| &\leqslant (\pi/2)^{1/2} \mathbf{E}_X \left\| \mathbf{E}_g \sum \varepsilon_i |g_i| X_i \right\| \\ &\leqslant (\pi/2)^{1/2} \mathbf{E} \left\| \sum \varepsilon_i |g_i| X_i \right\| \\ &= (\pi/2)^{1/2} \mathbf{E} \left\| \sum g_i X_i \right\|. \end{aligned}$$

where the last equality followed because $\varepsilon_i|g_i|$ is distributed identically to g_i. Putting this together with the first inequality gives the upper bound.

To prove the lower bound, we apply contraction and symmetrization to conclude

$$\begin{aligned}\mathbf{E}\left\|\sum g_i X_i\right\| &= \mathbf{E}\left\|\sum \varepsilon_i g_i X_i\right\| \\ &= \mathbf{E}\,\|g\|_\infty \cdot \mathbf{E}_\varepsilon \left\|\sum \varepsilon_i X_i\right\| \\ &\leqslant 2\,\mathbf{E}\,\|g\|_\infty \cdot \mathbf{E}_X \left\|\sum X_i\right\| \\ &\lesssim 2(\log n)^{1/2} \cdot \mathbf{E}\left\|\sum X_i\right\|. \qquad \square\end{aligned}$$

We can also prove a version of contraction which works for more general functions, which are *contraction* maps.

Theorem 3.16 (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*

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

Proof. TODO. □

Chapter 4

Chaining

In this chapter we try and simultaneously control the supremum of a family of random variables $\{X_t : t \in T\}$, where T is an arbitrary infinite index set. Of course, if the family of random variables are independant and identically distributed with common distribution X, this is easy, since $\sup X_t = \|X\|_{L^\infty}$ almost surely. To get interesting results, we need to assume some kind of dependence among the random variables. We will find that if the random variables $\{X_t\}$ are 'sufficiently continuous', then controlling $\sup_t X_t$ relies solely on studying the geometry of the index set T.

Example. *Given a random matrix A. It is an important problem to bound the operator norm*

$$\mathbf{E}\,\|A\| = \mathbf{E} \sup_{|x|=1} |Ax|.$$

Here the index set $T = \{x : |x| = 1\}$ is the unit sphere. Note that if the rows of A are independant and uniformly subgaussian, with subgaussian norm K, then $\|Ax\|_{\psi_2} \leqslant K|x|$, so by linearity, given $x, y \in T$,

$$\||Ax| - |Ay|\|_{\psi_2} \leqslant \||A(x-y)|\|_{\psi_2} \leqslant K|x-y|.$$

Thus if x and y are close, then Ax and Ay are very close with high probability. This is what we mean by the process being 'sufficiently continuous'.

We make a technical note that though we do assume that our index sets are infinite, we really concentrate on obtaining results for the suprema of finite, or countable index stochastic processes, which are *independant* of the cardinality of the index set. For instance, $\sup X_t$, need not even be measurable if we take the supremum over uncountably many indices t. For instance, when we refer to $\mathbf{E}(\sup X_t)$, what we really study is

$$\sup\left\{\mathbf{E}\left(\sup_{t\in T_0} X_t\right) : T_0 \subset T, T_0 \text{ finite}\right\}.$$

We find this agrees with the standard expectation if the index set is countable, by the monotone convergence theorem. For uncountable index sets that are of interest, such as variants of Brownian motion, or the random matrix process described above, one can use certain continuity properties of the process to ensure that the supremum is a measurable function, and that the standard expectation agrees with the finitary expectation defined above.

4.1 Covering Numbers

We recall some notation from metric space theory. If T is a metric space, a ε **net** S is a subset of T such that for any $t \in T$, there is $s \in S$ such that $d(t,s) < \varepsilon$. A ε **packing** of T is a ε separated family of points in T. The ε **covering number** of T is the smallest cardinality of a ε net, denoted $N(T,\varepsilon)$. On the other hand, the maximum cardinality of a ε packing is denoted by $P(T,\varepsilon)$.

Lemma 4.1. *For any metric space T and $\varepsilon > 0$,*

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

Proof. If S is a 2ε packing, and U a ε net, then each set in U can cover at most one element of S, so $|S| \leqslant |U|$. This gives that $P(T,2\varepsilon) \leqslant N(T,\varepsilon)$. On the other hand, if S is a maximal ε packing of T, then it is automatically a ε net, and so $|S| \geqslant N(T,\varepsilon)$. Since S was arbitrary, $P(T,\varepsilon) \geqslant N(T,\varepsilon)$. □

It is often useful to apply a 'volumetric argument' to calculate covering numbers. We suppose T is a subset of some ambient metric space S, which is equipped with a Radon measure μ. We consider the quantities

$$\mu(B(\varepsilon)) = \sup_{t\in S} \mu(B(t,\varepsilon)).$$

Given a set $U \subset S$, we let

$$U(\varepsilon) = \{s \in S : \text{There is } u \in U \text{ such that } d(s,u) < \varepsilon\}$$

denote the ε *thickening* of U.

Lemma 4.2. *For any $\varepsilon > 0$,*

$$\frac{\mu(T)}{\mu(B(\varepsilon))} \leqslant N(T,\varepsilon) \leqslant P(T,\varepsilon) \leqslant \frac{\mu(T(\varepsilon/2))}{\mu(B(\varepsilon/2))},$$

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

Proof. To prove the lower bound, let U denote a ε net of T. Then the ε balls around each point in U cover T, so a union bound gives

$$\mu(T) \leqslant |U| \cdot \mu(B(\varepsilon)).$$

Since U was arbitrary, taking U of minimum cardinality gives that $\mu(T) \leqslant N(T,\varepsilon)\mu(B(\varepsilon))$.

To prove the upper bound, let U denote a packing of T. Then the balls of radius $\varepsilon/2$ around the points of U are disjoint from one another, and are contained in $T(\varepsilon/2)$. Thus

$$|U| \cdot \mu(B(\varepsilon/2)) \leqslant \mu(T(\varepsilon/2)).$$

Since U was arbitrary, taking U of maximum cardinality gives that $P(T,\varepsilon)\cdot \mu(B(\varepsilon/2)) \leqslant \mu(T(\varepsilon/2))$. □

Example. *Consider the unit ball B in $\mathbf{R}^n$, equipped with the Lebesgue measure $|\cdot|$. Because the Lebesgue measure is homogenous, we know that $|B(\varepsilon)| = \varepsilon^n \cdot |B|$ for all $\varepsilon > 0$. Furthermore, $B(\varepsilon/2)$ is the ball of radius $1 + \varepsilon/2$. Thus the volumetric argument implies that*

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

If $S \subset B$ is the unit sphere, then $S(\varepsilon/2)$ is also contained in the ball of radius $1 + \varepsilon/2$, which gives that

$$N(S,\varepsilon) \leqslant P(S,\varepsilon) \leqslant \frac{(2+\varepsilon)^n}{\varepsilon^n}.$$

Example. *Consider the Hamming cube $H = \{0,1\}^n$, with the Hamming distance metric, defined for two strings $x, y \in H$ by setting*

$$d(x,y) = \#\{i : x_i \neq y_i\}.$$

For any $x \in H$, and any integer $k \leqslant n$, the ball of radius k around x contains $\sum_{i<k} \binom{n}{i}$ points. Thus applying the volumetric argument, we find that

$$\frac{2^n}{\sum_{i<k} \binom{n}{i}} \leqslant N(H,k) \leqslant P(H,k) \leqslant \frac{2^n}{\sum_{i<k/2} \binom{n}{i}} \ .$$

If $T_0 \subset T_1$, it is not necessarily true that $N(T_0, \varepsilon) \leqslant N(T_1, \varepsilon)$. But there is a remedy to this situation. We define an **exterior** ε **cover** of T as a subset U of S such that for any $t \in T$, there is $u \in U$ such that $d(t,u) < \varepsilon$. The **exterior** ε covering number $N(T,S,\varepsilon)$ of T with respect to S is the minimum cardinality of an exterior ε net.

Lemma 4.3. *For any set S, and $\varepsilon > 0$, $N(T,\varepsilon) \leqslant N(T,S,\varepsilon/2)$.*

Proof. If U is any exterior $\varepsilon/2$ net of T, define a set U_0, by picking, for each $u \in U$, a point $u_0 \in T$ such that $d(u,u_0) < \varepsilon/2$ (for any minimal $\varepsilon/2$ net, such a point u_0 must exist). Then U_0 is a ε net of T by the triangle inequality. □

The exterior covering number is obviously monotone, and the previous lemma implies that if $T_0 \subset T_1$, then

$$N(T_0,\varepsilon) \leqslant N(T_0,T_1,\varepsilon/2) \leqslant N(T_1,T_1,\varepsilon/2) = N(T_1,\varepsilon/2).$$

Thus $N(T_0,\varepsilon) \leqslant N(T_1,\varepsilon/2)$, which normally implies $N(T_1,\varepsilon)$ upper bounds $N(T_0,\varepsilon)$, up to a constant.

4.2 Chaining

We begin by studying a central method for bounding the suprema of random processes. Given a random process $\{X_t : t \in T\}$ indexed by a metric space T, we consider a ε net S of T with cardinality $N(T,\varepsilon)$. Given any $t \in T$, there is $\pi(t) \in S$ such that $d(t,\pi(t)) < \varepsilon$. Thus

$$\sup_{t \in T} X_t = \sup_{t \in T} X_{\pi(t)} + (X_t - X_{\pi(t)}) \leqslant \sup_{s \in S} X_s + \sup_{t \in T} (X_t - X_{\pi(t)}).$$

We have a quantative bound on the number of random variables in the sub-process $\{X_s : s \in S\}$ so it is often easy to bound this supremum. And the fact that $d(t, \pi(t)) < \varepsilon$ for each t often enables one to obtain bounds on $X_t - X_{\pi(t)}$, given that $X_t - X_{\pi(t)}$ is small if $d(t, \pi(t))$ is small. This is already enough to obtain certain results.

Example. *Let $T = \{x \in \mathbf{R}^n : |x| = 1\}$. Then if S is a ε, by linearity, the above equation gives*

$$\|A\| \leqslant \sup_{x \in S} |Ax| + \varepsilon \|A\|.$$

This inequality can be rearranged to read

$$\|A\| \leqslant \frac{1}{1-\varepsilon} \sup_{x \in S} |Ax|$$

This removes the problem of bounding the differences completely.

Example. *Let A be a $m \times n$ matrix. If $x \in \mathbf{R}^n$, and $y \in \mathbf{R}^m$ are given, with $|x - x_0| = \varepsilon$ and $|y - y_0| = \varepsilon$, then*

$$|y^T Ax - y_0^T Ax_0| \leqslant |y^T A(x - x_0)| + |(y - y_0)^T Ax_0| \leqslant 2\|A\|\varepsilon.$$

Thus if $T_n = \{x \in \mathbf{R}^n : |x| = 1\}$ and $T_m = \{x \in \mathbf{R}^m : |x| = 1\}$, then for any pair of ε net $S_n \subset T_n$ and $S_m \subset T_m$,

$$\begin{aligned} \|A\| &= \sup_{x \in T_n, y \in T_m} |y^T Ax| \\ &\leqslant \sup_{x \in S_n, y \in S_m} |y^T Ax| + 2\|A\|\varepsilon. \end{aligned}$$

This can be rearranged to read

$$\|A\| \leqslant \frac{1}{1 + 2\varepsilon} \sup_{x \in S_n, y \in S_m} |y^T Ax|.$$

4.3 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.

Lemma 4.4. *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. □

Applying this trick, together with a covering argument and a union bound, easily enables us to bound the norm of a matrix with high probability.

Theorem 4.5. *Let A be an $m \times n$ matrix with independant, mean zero subgaussian entries. There exists a universal constant C such that if $K = \max \|A_{ij}\|_{\psi_2}$, and $t > 0$,*

$$\|A\| \leqslant CK(\sqrt{m} + \sqrt{n} + t),$$

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

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 $|M| \leqslant 9^n$. The last lemma implies

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

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

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

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

Working slightly harder, using Bernstein's inequality instead of Hoeffding's ienquality, we can both upper bound and lower bound the behaviour of the operator with high probability; our proof 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 4.7. *Let A be an $m \times n$ matrix whose rows are independant, mean zero, subgaussian isotropic random vectors. Then there exists a universal constant C such that if K bounds the maximum subgaussian norm of the rows of A, then with probability $1 - 2\exp(-t^2)$, for any $x \in \mathbf{R}^n$,*

$$||Ax| - m^{1/2}|x|| \leqslant CK^2 m^{1/2}(n^{1/2} + t)|x|.$$

Proof. We know that $K \gtrsim 1$, and without loss of generality, we can assume that $K \geqslant 1$. We prove the stronger conclusion that with probability $1 - 2\exp(-t^2)$,

$$\|(A^T A)/m - 1\| \leqslant K^2 \max(\delta, \delta^2) \quad \text{where} \quad \delta = C\left((n/m)^{1/2} + t/m^{1/2}\right).$$

If this inequality is true, then for any x,

$$\begin{aligned} ||Ax|^2 - m|x|^2| &= |(Ax \cdot Ax) - m(x \cdot x)| = |(A^T A - m)x \cdot x| \\ &\leqslant \|A^T A - m\| \cdot |x|^2 \leqslant \max(mK^2\delta, (mK^2\delta)^2) \cdot |x|^2. \end{aligned}$$

Recalling the calculation we made for subgaussian vectors, this implies

$$||Ax| - m^{1/2}|x|| \leqslant K^2 \delta m|x| \leqslant CK^2 m^{1/2}(n^{1/2} + t)|x|.$$

To prove the identity, 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 there exists a universal constant c such that

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

We can then apply a union bound to conclude

$$\begin{aligned}\mathbf{P}\left(\max_{x\in N}|Ax|^2/m-1\geqslant\varepsilon/2\right)&\leqslant 2|N|\exp(-c\cdot C^2(n+t^2))\\&\leqslant 2\cdot 9^n\exp(-c\cdot C^2(n+t^2))\\&\leqslant 2\exp((\log 9-c\cdot C^2)n-c\cdot C^2t^2).\end{aligned}$$

Choosing $C^2\geqslant\log 9/c$ gives the required inequality. □

In other applications of covering, the value $\sup X_t-X_{\pi(t)}$ is not able to be completely discarded, and we must come up with a more careful method to bound the quantity. We can do this by applying a 'multi-scale' covering argument, rather than just a single scale covering. This technique is known as *chaining*. Given a random process $\{X_t : t\in T\}$, where T is a metric space, we define the Subgaussian norm

$$\|X\|_{\psi_2}=\sup_{t,s\in T}\frac{\|X_t-X_s\|_{\psi_2}}{d(t,s)}.$$

If a random process is subgaussian, then chaining enables us to obtain a bound on the expectation of it's supremum. If $\|X\|_{\psi_2}<\infty$, we say X has **subgaussian increments**.

Theorem 4.8. *Let $\{X_t : t\in T\}$ be a random process. Then*

$$\mathbf{E}\left(\sup_{t\in T}X_t\right)\lesssim\|X\|_{\psi_2}\int_0^\infty\sqrt{\log N(T,\varepsilon)}\,d\varepsilon.$$

Proof. Without loss of generality, we may assume that T is finite. For each k, we consider a $1/2^k$ net S_k with $|S_k|=N(T,2^k)$. If $1/2^M<\min(d(t,t') : t,t'\in T)$, then $S_M=T$, and if $1/2^N>\operatorname{diam}(T)$, $|S_N|=1$. If for each $t\in T$, we let $\pi_k(t)$ be a point in S_k with $d(t,\pi_k(t))<1/2^k$, then π_N is constant, and π_M is the identity function. Thus for any t,

$$X_t=\sum_{k=-\infty}^{\infty}X_{\pi_{k+1}(t)}-X_{\pi_k(t)}.$$

In particular,

$$\sup_{t\in T}X_t\leqslant\sum_{k=-\infty}^{\infty}\sup_{t\in T}X_{\pi_{k+1}(t)}-X_{\pi_k(t)}.$$

Notice that the latter supremum is really over pairs of indices in S_k and S_{k+1}. Thus it is a supremum over at most $N(T,1/2^{k+1})^2$ random variables, and each has subgaussian norm at most $K/2^{k-1}$. Thus we conclude that

$$\mathbf{E}\left(\sup_{t\in T} X_{\pi_{k+1}(t)} - X_{\pi_k(t)}\right) \lesssim (K/2^{k-1})\sqrt{\log N(T,1/2^{k+1})^2}$$
$$\lesssim (K/2^k)\sqrt{\log N(T,1/2^{k+1})}.$$

Thus

$$\mathbf{E}\left(\sup_{t\in T} X_t\right) \lesssim K\sum_{k=-\infty}^{\infty} 2^k\sqrt{\log N(T,1/2^k)} \approx K\int_0^\infty \sqrt{\log N(T,\varepsilon)}d\varepsilon. \quad \square$$

If we work slightly harder, then we can obtain a tail bound version of Dudley's inequality, which is often useful.

Theorem 4.9. *With probability* $1-2\exp(-u^2)$,

$$\sup_{t\in T} X_t \lesssim \|X\|_{\psi_2}\left(\int_0^\infty \sqrt{\log N(T,\varepsilon)}\, d\varepsilon + u\cdot diam(T)\right).$$

Proof. A simple union bound argument shows that there exists a constant C such that

$$\sup_{t\in T}(X_{\pi_{k+1}(t)} - X_{\pi_k(t)}) \leqslant (C\cdot K/2^k)\left(\sqrt{\log|S_{k+1}|}+u\right)$$

with probability $1-2\exp(-u^2)$. Next, we note that if $1/2^N \leqslant \mathrm{diam}(T)$, then $S_N = T$. If M is the greatest integer with $1/2^M \geqslant \mathrm{diam}(T)$, then a simple union bound gives, for any sequence $\{u_k\}$,

$$\sup_{t\in T} X_t \leqslant CK\sum_{k=M}^{\infty}(1/2^k)\cdot\left(\sqrt{\log|S_k|}+u_k\right)$$
$$\leqslant K\int_0^\infty \sqrt{\log N(T,\varepsilon)}\, d\varepsilon + CK\sum_{k=M}^{\infty}(u_k/2^k).$$

with probability $1-2\sum_{k=M}^\infty \exp(-u_k^2)$. If $u_k = u+k+10$, then

$$\sup_{t\in T} X_t \lesssim K\int_0^\infty \sqrt{\log N(T,\varepsilon)}\, d\varepsilon + u\,\mathrm{diam}(T) + \mathrm{diam}(T)$$

with probability greater than

$$1 - 2\exp(-u^2)\exp(-100)\sum_{k=1}^{\infty}\exp(-k^2) \geqslant 1 - 2\exp(-u^2)$$

Note that if $\varepsilon \leqslant \text{diam}(T)/2$, then $N(T,\varepsilon) \geqslant 2$, so

$$\int_0^{\infty}\sqrt{\log N(T,\varepsilon)} \gtrsim \text{diam}(T).$$

In particular, this means that with probability greater than $1-2\exp(-u^2)$,

$$\sup_{t\in T} X_t \lesssim K\int_0^{\infty}\sqrt{\log N(T,\varepsilon)}\, d\varepsilon + u\,\text{diam}(T). \qquad \square$$

Here is a fundamental example of the kinds of random processes we study. Given $T \subset \mathbf{R}^n$, and a Gaussian vector $g \sim N(0,I_n)$, we consider the supremum

$$\sup_{t\in T} g\cdot t.$$

The expectation of this supremum is known as the **Gaussian width** of T, denoted $w(T)$. Note that if $X_t = g\cdot t$, then

$$\|X_t - X_s\|_{\psi_2} = \|g\cdot(t-s)\|_{\psi_2} \lesssim |t-s|.$$

Thus $\|X\|_{\psi_2} \lesssim 1$. Thus Dudley's inequality shows that

$$w(T) \lesssim \int_0^{\infty}\sqrt{\log N(T,\varepsilon)},$$

where we use the standard metric on $\mathbf{R}^n$ to define the covering numbers.

Example. *Let T be the unit ball in $\mathbf{R}^n$. Then the volumetric argument we considered early on in this section let us show $N(T,\varepsilon) \leqslant (3/\varepsilon)^n$, for $\varepsilon \leqslant 1$, and $N(T,\varepsilon) = 1$ for $\varepsilon \geqslant 1$. Thus*

$$w(T) \lesssim \int_0^1\sqrt{n\log(3/\varepsilon)}\, d\varepsilon \lesssim \sqrt{n}.$$

There are situations where Dudley's inequality is not sharp. But we will later provide lower bounds on the expectation of certain random processes which show in most cases, Dudley's inequality is sharp up to a $\log(n)$ factor.

Example. *Consider the Gaussian width of*

$$T = \left\{ \frac{e_k}{\sqrt{1+\log k}} : k = 1, \dots, n \right\}$$

We claim that $w(T)$ is bounded independantly of n. Obtaining this result is equivalent to bounding

$$\mathbf{E}\left(\max_{1 \leqslant k \leqslant n} \frac{g_k}{\sqrt{1+\log k}} \right),$$

independantly of n, where $g_1, \dots, g_n$ are independant $N(0,1)$ random variables. Applying a union bound, we find there exists a small constant c such that

$$\begin{aligned} \mathbf{P}\left(\max_{1 \leqslant k \leqslant n} \frac{g_k}{\sqrt{1+\log k}} \geqslant t \right) &\leqslant \sum_{k=1}^n \mathbf{P}\left(g_k \geqslant t\sqrt{1+\log k} \right) \\ &\leqslant \sum_{k=1}^n \exp(-c \cdot (1+\log k) t^2) \end{aligned}$$

If $t^2 \geqslant 4/c$, we find

$$\begin{aligned} \sum_{k=1}^n \exp(-c \cdot (1+\log k)t^2) &= \exp(-(c/2) \cdot t^2) \cdot \sum_{k=1}^n \exp((c/2 - c(1+\log k)) \cdot t^2) \\ &\leqslant \exp(-(c/2) \cdot t^2) \sum_{k=1}^\infty k^{-2} \lesssim \exp(-(c/2) \cdot t^2). \end{aligned}$$

Thus

$$\mathbf{E}\left(\max_{1 \leqslant k \leqslant n} \frac{g_k}{\sqrt{1+\log k}} \right) \lesssim \sqrt{4/c} + \int_{(4/c)^{1/2}}^\infty \exp(-(c/2) \cdot t^2) \lesssim 1.$$

This gives a bound independant on n, so $w(T) \lesssim 1$. On the other hand, we calculate

$$\left| \frac{g_i}{\sqrt{1+\log i}} - \frac{g_j}{\sqrt{1+\log j}} \right|^2 \leqslant \frac{2}{1+\min(\log i, \log j)}$$

Thus for each m, the values

$$\left\{g_{n-m}/\sqrt{1+\log(n-m)},\dots,g_n/\sqrt{1+\log m}\right\}$$

give a $2/\sqrt{1+\log(n-m)}$ *packing. Thus for each k, if*

$$\frac{2}{\sqrt{1+\log(k+1)}} \leqslant \varepsilon \leqslant \frac{2}{1+\log(k)},$$

then $N(T,\varepsilon) \geqslant k$. *In particular, this means*

$$\begin{aligned}\int_0^\infty N(T,\varepsilon) &\geqslant \sum_{k=1}^n k\left(\frac{2}{\sqrt{1+\log(k)}} - \frac{2}{\sqrt{1+\log(k+1)}}\right)\\ &\geqslant \sum_{k=2}^n k \cdot \frac{\log(1+1/k)}{\log(k)^{3/2}}\end{aligned}$$

and this value becomes unbounded as $n \to \infty$.

Gaussian width is very important to the application of the concepts of high dimensional probability to statistics.

4.4 Empirical Processes

An *empirical process* is a type of random process whose index set consists of *functions*. The motivating example, given i.i.d random variables $X_1,\dots,X_n$ with distribution given by some probability measure μ, to consider the functional

$$f \mapsto \frac{1}{n}\sum_{k=1}^n f(X_k).$$

A natural question is how close this functional is to the functional

$$f \mapsto \int f\, d\mu.$$

Approximating this integral by random samples is known as the *Monte Carlo method*. The law of large numbers implies that this functional does converge *pointwise* for each function f. But quantitative estimates on the error are much more useful. We now show that these functional do converge uniformly, over the class $\mathcal{F} = \{f : [0,1] \to \mathbf{R} : \|f\|_{\mathrm{Lip}} \leqslant L\}$, for some $L < \infty$.

For a fixed function f, we can calculate that

$$\mathbf{E}\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbf{E}f(X)\right| \leqslant \mathbf{V}\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k)\right)^{1/2} = O(1/\sqrt{n}).$$

The next theorem shows we can guarantee this expectation bound *uniformly* across an infinite family of functions. Notice that in the theorem, the supremum occurs inside the expectation, which makes the consequence that much more powerful.

Theorem 4.10. *Let $X_1,\dots,X_n$ be i.i.d random variables taking values in $[0,1]$. Then*

$$\mathbf{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbf{E}f(X)\right| \lesssim L\cdot n^{-1/2}.$$

Proof. Let X denote the common distribution of the $\{X_k\}$. For each $f \in \mathcal{F}$, set

$$X_f = \frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbf{E}f(X).$$

Without loss of generality, we assume $L = 1$. Furthermore, because X_f is translation invariant, we may assume we are working with the class of all 1 Lipschitz functions f with $f(0) = 0$. In particular, this implies that $\|f\|_\infty \leqslant 1$.

We are interested in bounding $\mathbf{E}\sup_{f\in\mathcal{F}}|X_f|$. We may then apply independence and centering to conclude

$$\begin{aligned}\|X_f - X_g\|_{\psi_2}^2 &= n^{-2}\sum\|[f(X_k) - g(X_k)] - \mathbf{E}(f-g)(X)\|_{\psi_2}^2\\ &\leqslant n^{-2}\sum\|f(X_k) - g(X_k)\|_{\psi_2}\\ &\lesssim n^{-1}\|f-g\|_{L^\infty(\mathbf{R}^n)}^2.\end{aligned}$$

Thus $\|X_f - X_g\|_{\psi_2} \lesssim n^{-1/2}\|f-g\|_{L^\infty(\mathbf{R}^n)}$, so the process is subgaussian with respect to the L^∞ norm on $\mathcal{F}$.

We wish to apply Dudley's inequality, so we must calculate the covering number of $\mathcal{F}$. For each m, form a grid of $1/m \times 1/m$ squares partitioning $[0,1]\times[-1,1]$. Then any function $f : [0,1] \to [0,1]$, we can approximate the graph of f by travelling along squares, up to a ε or error. Thus we obtain a $1/m$ net by taking the class of all 1 Lipschitz functions obtained by travelling along the net. The number of such functions is bounded by 3^m. We now apply Dudley's inequality. Since $\|f-g\|_{L^\infty[0,1]} \leqslant 2$, for each $f,g \in \mathcal{F}$ with $f(0) = g(0) = 0$,

$$\begin{aligned} \mathbf{E} \sup_{f \in \mathcal{F}} |X_f| &\lesssim n^{-1/2} \sum_{k=-10} 2^{-k} (\log(3^{2^k}))^{1/2} \\ &\lesssim n^{-1/2} \sum_{k=0}^{\infty} 2^{-k/2} \lesssim n^{-1/2}. \end{aligned}$$ □

Remark. For any two measures μ and η, the *Wasserstein distance* is defined to be

$$\sup_{f \in \mathcal{F}} \left| \int f \, d\mu - \int f \, d\eta \right|.$$

The theorem we just proved shows that if we draw i.i.d random variables $X_1, \dots, X_n$ with some distribution μ, and define $\mu_n = \sum X_i \delta_{X_i}$ to be the sum of points masses at the same point, then the Wasserstein distance between μ and μ_n is, in expectation, $O(n^{-1/2})$.

4.5 Vapnik-Chervonenkis Dimension

We now discuss Vapnik-Chervonenkis dimension, which is a notion of complexity for families of Boolean functions $\{0,1\}^\Omega$ where Ω is some arbitrary set. This notion of complexity is related to covering numbers, and thus we can relate certain learning bounds with Dudley's inequality.

Given a family $\mathcal{F}$ of Boolean functions on some domain Ω, we say a set $\Lambda \subset \Omega$ is *shattered* by $\mathcal{F}$ if the restrict map $F : \{0,1\}^\Omega \to \{0,1\}^\Lambda$ is surjective. The **VC Dimension** of $\mathcal{F}$, denoted by $\mathrm{vc}(\mathcal{F})$ is the largest cardinality of a set shattered by $\mathcal{F}$.

Example. *Let $\mathcal{F} = \{\chi_{[a,b]} : a < b\}$ be the class of all indicator functions of intervals on $\mathbf{R}$. The family $\mathcal{F}$ shatters any two point set, but no three point set is shattered; for instance, if $t < s < u$, then there is no $f \in \mathcal{F}$ with $f(t) = 1$, $f(s) = 0$, and $f(u) = 1$. Thus $vc(\mathcal{F}) = 2$.*

Example. *Let $\mathcal{F} = \{\chi_{\mathbf{H}} : \mathbf{H}$ is a half plane in $\mathbf{R}^2\}$. Then $\mathcal{F}$ is a family of Boolean functions in $\mathbf{R}^2$. The family $\mathcal{F}$ shatters any three points in general position. But $\mathcal{F}$ doesn't shatter any four point set, so $vc(\mathcal{F}) = 3$.*

Example. *If $\Omega = \{1,2,3\}$, then elements of $\{0,1\}^\Omega$ can be identified with length three binary strings. Consider*

$$\mathcal{F} = \{001, 010, 100, 111\}.$$

Then $\mathcal{F}$ shatters any two point set. But it doesn't shatter Ω, since $\mathcal{F} \neq \{0,1\}^\Omega$, so $vc(\mathcal{F}) = 2$.

Example. *Let*

$$\mathcal{F} = \{\chi_{[a,b]\cup[c,d]} : a,b,c,d \in \mathbf{R}\}$$

be the indicator functions of unions of two intervals. This class can even shatter four point sets. But it cannot shatter five point sets, because if $s < t < u < v < w$, there is no $f \in \mathcal{F}$ with $f(s) = f(u) = f(w) = 1$, and $f(t) = f(v) = 0$. Thus $vc(\mathcal{F}) = 4$.

Example. *Let $\mathcal{F}$ be the class of indicator functions of circles in $\mathbf{R}^2$. Then $\mathcal{F}$ can shatter three point sets, but not four point sets. To prove the latter point, we may assume a four point set is in general position, because otherwise we can reduce to the case of indicator functions of intervals. If one of the points is in the convex hull of the other three, then $\mathcal{F}$ cannot shatter these points, because a circle cannot contain the three points, but not the fourth point. TODO.*

Example. *The class of indicator functions of axis aligned rectangles in $\mathbf{R}^2$ has VC dimension 4. Conversely, the indicators of axis aligned squares has VC dimension 3.*

Example. *Let $\mathcal{F}$ be the class of indicator functions of all convex polygons in $\mathbf{R}^2$. Then one can show $\mathcal{F}$ shatters sets of arbitrarily large cardinality. Thus $vc(\mathcal{F}) = \infty$.*

Lemma 4.11 (Pajor's Lemma)**.** *Given any class of Boolean functions $\mathcal{F}$ on Ω,*

$$|\mathcal{F}| \leqslant |\{\Lambda \subset \Omega : \mathcal{F} \text{ shatters } \Lambda\}|.$$

Proof. Let $S(\mathcal{F}) = \{\Lambda \subset \Omega : \mathcal{F} \text{ shatters } \Lambda\}$. We prove by induction on Ω. If $\Omega = \varnothing$, then $\{0,1\}^\Omega = \varnothing$, so $\mathcal{F} = \varnothing$, and $S(\varnothing) = \{\varnothing\}$, so the inequality in the lemma reads $0 \leqslant 1$, which is true.

Next, we suppose $\Omega = \Omega_0 \cup \{\omega\}$. Then $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$, where $\mathcal{F}_i = \{f \in \mathcal{F} : f(\omega) = i\}$. By induction, $|S(\mathcal{F}_i)| \geqslant |\mathcal{F}_i|$. To finish the proof, we need only check $|S(\mathcal{F}_0)| + |S(\mathcal{F}_1)| \leqslant |S(\mathcal{F})|$. We note that for each $\Lambda \in S(\mathcal{F}_0) \cap S(\mathcal{F}_1)$, $\mathcal{F}$ shatters Λ and $\Lambda \cup \{\omega\}$. Thus

$$S(\mathcal{F}) \supset S(\mathcal{F}_0) \cup S(\mathcal{F}_0) \cup (S(\mathcal{F}_0) \cap S(\mathcal{F}_1)) \times \{\omega\}$$

And this is sufficient to obtain the inequality. □

Theorem 4.12 (Sauer-Shelah Lemma). *Let $\mathcal{F}$ be a class of Boolean functions on an n point set Ω. Then*

$$|\mathcal{F}| \leqslant \sum_{k=0}^{vc(\mathcal{F})} \binom{n}{k} \leqslant \left(\frac{en}{vc(\mathcal{F})}\right)^{vc(\mathcal{F})}.$$

Proof. The combinatorial sum counts the number of subsets of Ω containing at most $\mathrm{vc}(\mathcal{F})$ points. A simple application of Pajor's lemma completes the proof. □

These lemmas are useful for analyzing finite collections of Boolean functions, but do not give any useful information if the class of functions is infinite. To analyze these functions, it turns out that the covering number of $\mathcal{F}$ becomes useful here.

Given any probability measure μ on Ω, we can discuss the L^2 norm, which for any two Boolean functions $f, g \in \{0,1\}^\Omega$,

$$\|f - g\|_{L^2(\mu)} = (\mathbf{E}|f(X) - g(X)|^2)^{1/2} = \mathbf{P}(f(X) \neq g(X))^{1/2}.$$

We can then discuss the covering numbers $N(\mathcal{F}, \mu, \varepsilon)$ with respect to this metric. First, we need to discuss a way to reduce the size of the set $\mathcal{F}$.

Lemma 4.13. *Let $\mathcal{F}$ consist of N functions on a probability space Ω with probability measure μ. If $\|f - g\|_{L^2(\mu)} > \varepsilon$ for any distinct $f, g \in \mathcal{F}$, then there exists $n \lesssim \log N/\varepsilon^4$ and an n point subset Λ of Ω such that if η is the uniform probability measure on Λ, then $\|f - g\|_{L^2(\eta)} > (\varepsilon/2)$ for distinct $f, g \in \mathcal{F}$.*

Proof. We rely on the probabilistic method. Let $X_1, \dots, X_n$ be n independent independent points drawn from the measure μ, and let η be the average of point masses on each X_i. Set $h = (f - g)^2$. We want to bound the deviation

$$\|f - g\|_{L^2(\eta)}^2 - \|f - g\|_{L^2(\mu)}^2 = \frac{1}{n}\sum h(X_k) - \mathbf{E}\,h(X).$$

We calculate that

$$\|h(X_k) - \mathbf{E}\, h(X)\|_{\psi_2} \lesssim \|h(X)\|_{\psi_2} \lesssim \|h(X)\|_\infty \lesssim 1.$$

Then Hoeffding's inequality implies that there is a constant c with

$$\mathbf{P}\left(\left|\|f-g\|_{L^2(\eta)}^2 - \|f-g\|_{L^2(\mu)}^2\right| > \varepsilon^2/4\right) \leqslant 2\exp(-cn\varepsilon^4).$$

And if this is true, then $\|f-g\|_{L^2(\eta)}^2 > 3\varepsilon^2/4$, so

$$\|f-g\|_{L^2(\eta)} > (3/4)^{1/2}\varepsilon \geqslant \varepsilon/2.$$

Taking a union bound over all $f,g \in \mathcal{F}$, we conclude that with probability at least $1 - 2|\mathcal{F}|^2\exp(-cn\varepsilon^4) = 1 - 2N^2\exp(-cn\varepsilon^4)$, the inequality above holds for all f,g. If

$$n \gtrsim \log(N)/\varepsilon^4,$$

this quantity is nonzero, and therefore such a choice of $X_1,\dots,X_n$ exists. Now some of the X_i may be repeated, but removing some of them only decreases the mass of η, so this isn't a problem. □

Theorem 4.14. *Let $\mathcal{F} \subset \{0,1\}^\Omega$. Then there is a universal constant C such that for every $\varepsilon \in (0,1)$,*

$$N(\mathcal{F},\varepsilon) \leqslant (2/\varepsilon)^{C\cdot vc(\mathcal{F})}.$$

Proof. If Ω is finite, we could employ the Sauer-Shelah lemma to conclude

$$N(\mathcal{F},\varepsilon) \leqslant \left(\frac{en}{\mathrm{vc}(\mathcal{F})}\right)^{\mathrm{vc}(\mathcal{F})}.$$

This is almost what we want. We apply the dimension-reduction lemma to apply something along the lines of this.

For a fixed ε, let $\mathcal{G} \subset \mathcal{F}$ be a ε packing consisting of at least $N(\mathcal{F},\mu,\varepsilon)$ elements. If we apply the reduction lemma, we obtain a set Λ with cardinality $O(\log N(\mathcal{F},\mu,\varepsilon)/\varepsilon^4)$ such that $\mathcal{G}$ is still a $\varepsilon/2$ packing with respect to the uniform measure on Λ. Applying Sauer-Shelah to $\mathcal{G}$, we conclude that if d is the VC dimension of $\mathcal{G}$ on Λ,

$$N(\mathcal{F},\mu,\varepsilon) \leqslant O\left(\frac{\log N(\mathcal{F},\mu,\varepsilon)}{d\varepsilon^4}\right)^d$$

But $\log N(\mathcal{F},\mu,\varepsilon)/2d = \log(N^{1/2d}) \leqslant N^{1/2d}$, so we can simplify to conclude

$$N(\mathcal{F},\mu,\varepsilon) \leqslant O(1/\varepsilon)^{8d}.$$

The proof is completed when we notice $d \leqslant \mathrm{vc}(\mathcal{F})$. □

Now we have connected covering numbers to VC dimension, we can now apply Dudley's inequality to upper bound the supremum of deviations of a class of Boolean functions.

Theorem 4.15. *Let $\mathcal{F}$ be a class of Boolean functions on a probability space Ω with VC dimension $vc(\mathcal{F})$. If $X_1,\dots,X_n$ are i.i.d samples taken from Ω, with common distribution X, then*

$$\mathbf{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum f(X_k) - \mathbf{E}f(X)\right| \lesssim \sqrt{\frac{vc(\mathcal{F})}{n}}.$$

Proof. A simple symmetrization argument shows that

$$\mathbf{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum f(X_k) - \mathbf{E}f(X)\right| \leqslant 2\,\mathbf{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum \varepsilon_k f(X_k)\right|,$$

where $\{\varepsilon_k\}$ are i.i.d symmetric Bernoulli variables. Next, we condition on the $\{X_k\}$, viewing $\{\varepsilon_k\}$ as the only random quantities. Set $Z_f = \sum \varepsilon_i f(X_i)$, so that we are now interested in bounding $\mathbf{E}\sup|Z_f|$. Notice that

$$\begin{aligned}\|Z_f - Z_g\|_{\psi_2} &= \left\|\sum \varepsilon_k (f-g)(X_k)\right\|_{\psi_2}\\ &\lesssim \sqrt{n}\cdot\left(\frac{1}{n}\sum (f-g)(X_k)^2\right)^{1/2} = \sqrt{n}\cdot\|f-g\|_{L^2(\eta)}.\end{aligned}$$

Thus $\|Z_f - Z_g\|_{\psi_2} \lesssim \sqrt{n}\cdot\|f-g\|_{L^2(\eta)}$, where η is the uniform distribution on the conditioned points $\{X_1,\dots,X_n\}$. If we assume $0\in\mathcal{F}$, then we can apply Dudley's inequality, combined with the last theorem, to conclude

$$\begin{aligned}\mathbf{E}\sup|Z_f| = \mathbf{E}\sup|Z_f - Z_0| &\lesssim \frac{1}{\sqrt{n}}\mathbf{E}_X\int_0^1 (\log N(\mathcal{F},\eta,\varepsilon))^{1/2}\,d\varepsilon\\ &\lesssim \frac{1}{\sqrt{n}}\mathbf{E}_X\int_0^1 (\mathrm{vc}(\mathcal{F})\log(2/\varepsilon))^{1/2}\,d\varepsilon\\ &\lesssim \left(\frac{\mathrm{vc}(\mathcal{F})}{n}\right)^{1/2}.\end{aligned}$$

In general, if $0 \notin \mathcal{F}$, then we can still apply the previous bound to obtain

$$\mathbf{E} \sup |Z_f| \lesssim \left(\frac{\mathrm{vc}(\mathcal{F} \cup \{0\})}{n} \right)^{1/2}.$$

But this really proves the result. If Λ is shattered by $\mathcal{F} \cup \{0\}$, then $\mathcal{F}$ shatters every proper subset of Λ, so $\mathrm{vc}(\mathcal{F} \cup \{0\}) \leqslant \mathrm{vc}(\mathcal{F}) + 1$. □

An important consequence of this result is the Glivenko-Cantelli theorem. How many samples does it take to estimate an arbitrary distribution given by a cumulative distribution function F? Given $X_1, \dots, X_n$, we might want to define an estimate

$$\widehat{F}(x) = \frac{\#\{i : X_i \leqslant x\}}{n}$$

The weak law of large numbers shows $\mathbf{E}\,|\widehat{F}(x) - F(x)| \lesssim 1/\sqrt{n}$ for every $x \in \mathbf{R}$. The Glivenko-Cantelli theorem shows that this expectation bound is achieved uniformly.

Theorem 4.16 (Glivenko-Cantelli)**.** *Given a CDF F, samples $X_1, \dots, X_n$, and an induced estimate $\widehat{F}$,*

$$\mathbf{E}\,\|F - \widehat{F}\|_{L^\infty(\mathbf{R})} \lesssim 1/\sqrt{n}.$$

Proof. Let $\Omega = \mathbf{R}$, and $\mathcal{F}$ the indicators of $(-\infty, x]$. Then $\mathrm{vc}(\mathcal{F}) \leqslant 2$, which immediately implies the result. □

Remark. A class of functions $\mathcal{F}$ is called *uniformly Glivenko-Cantelli* if for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \sup_{\mu} \mathbf{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum f(X_k) - \mathbf{E} f(X) \right| > \varepsilon \right) = 0$$

where $\{X_k\}$ are i.i.d samples with respect to μ, and the supremum is taken over all probability measures μ. Markov's inequality combined with the main result of this section shows any class of Boolean functions with finite VC dimension is uniform Glivenko-Cantelli.

Theorem 4.17. *Any class of Boolean functions with infinite VC dimension is not uniform Glivenko-Cantelli.*

Proof. Let $\mathcal{F}$ be a collection of Boolean functions, and let Ω be an N point set shattered by $\mathcal{F}$. If μ is the uniform distribution on Ω, then for any $X_1, \dots, X_n$ drawn from μ, we can find f with $f(X_k) = 0$ for all k, but with $\mathbf{E} f(X) = 1 - n/N$. Thus

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum f(X_k) - \mathbf{E} f(X) \right| \geqslant 1 - n/N.$$

Thus for all n,

$$\sup_{\mu} \mathbf{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum f(X_k) - \mathbf{E} f(X) \right| \right) = 1.$$

□

An important application of the theory of complexity of Boolean functions is to classification in mathematical statistics. Given a hidden rule $T : \Omega \to \{0,1\}$, we consider samples $(X_1, Y_1), \dots, (X_n, Y_n)$ with $X_k \in \Omega$, and $T(X_k) = Y_k$, our goal is to learn T from the samples. An answer this problem would be a function $f : \Omega \to \{0,1\}$. The *risk* corresponding to this function is $R(f) = \mathbf{P}(f(X) \neq T(X))$, and our goal is to choose f such that $R(f)$ is as small as possible.

If T is arbitrary, it can be very difficult to learn T from samples, if not impossible. To restrict the complexity of T, we force our candidate functions f to lie in some hypothesis space $\mathcal{F}$. We must balance *fit* and *complexity*. Hopefully, some function in $\mathcal{F}$ approximates T accurately, but $\mathcal{F}$ itself has low complexity.

Our best answer to the question would be to compute $f^* = \text{argmin}_{f \in \mathcal{F}} R(f)$. But we are not able to compute $R(f)$, because we do not know the distribution our data came from. But we can always *estimate* using our training data. We define the *empirical risk* to be

$$\widehat{R}(f) = \frac{1}{n} \sum (f(X_k) - T(X_k))^2.$$

Let $\widehat{f^*}$ denote the minimizer of empirical risk. The main question is the difference between $R(\widehat{f^*})$ and $R(f^*)$. We can use the VC dimension to answer this question.

Theorem 4.18. *Let T be a Boolean function, and a hypothesis space $\mathcal{F}$. Then*

$$\mathbf{E}\left(R(\widehat{f^*})\right) \leqslant R(f^*) + O\left(\left(\frac{vc(\mathcal{F})}{n}\right)^{1/2}\right).$$

Proof. Let $\varepsilon = \sup_{f\in\mathcal{F}} |\widehat{R}(f) - R(f)|$. Then

$$R\left(\widehat{f^*}\right) \leqslant \widehat{R}\left(\widehat{f^*}\right) + \varepsilon \leqslant \widehat{R}(f^*) + \varepsilon \leqslant R(f^*) + \varepsilon.$$

Thus $R(\widehat{f^*}) - R(f^*) \leqslant 2\varepsilon$. It thus suffices to show that

$$\mathbf{E}(\varepsilon) = \mathbf{E}\left(\sup_{f\in\mathcal{F}} |\widehat{R}(f) - R(f)|\right) \lesssim \left(\frac{\mathrm{vc}(\mathcal{F})}{n}\right)^{1/2}.$$

If we set $\mathcal{L} = \{(f-T)^2 : f \in \mathcal{F}\}$, then

$$\varepsilon = \sup_{l\in\mathcal{L}} \left|\frac{1}{n}\sum l(X_k) - \mathbf{E}\, l(X)\right|$$

There is no easy way to relate $\mathrm{vc}(\mathcal{L})$ to $\mathrm{vc}(\mathcal{F})$, so we cannot just apply the normal uniform bound. But if η is the uniform distribution on $\{X_1, \dots, X_n\}$, then for $\varepsilon \in (0,1)$, we do have

$$N(\mathcal{L}, \eta, \varepsilon) \leqslant N(\mathcal{F}, \eta, \varepsilon/4).$$

This is because if $\mathcal{G}$ is a $\varepsilon/4$ net in $\mathcal{F}$, then $\{(g-T)^2 : g \in \mathcal{G}\}$ is a ε net in $\mathcal{L}$. And then we apply Dudley's inequality to yield the theorem. □

If we learn from data $X_1, \dots, X_n$, but we can take $n \to \infty$, then by Glivenko Cantelli we can approximate the CDF of the underlying distribution arbitrarily small, and therefore calculate R up to an arbitrarily small error as well. This means that we can find the risk minimizer f^*, and $R(f^*)$ is the 'risk' associated with choosing the hypothesis class $\mathcal{F}$. But if we cannot let n be arbitrarily large, we occur an additional risk quantity

$$R\left(\widehat{f^*}\right) - R(f^*),$$

known as *excess risk*. The last theorem says that the expected excess risk is on the order of $(\mathrm{vc}(\mathcal{F})/n)^{1/2}$. Thus if we want the expected excess risk to be smaller than ε, we need $n \sim \mathrm{vc}(\mathcal{F})/\varepsilon^2$ samples. To learn, we need more samples than the complexity of the hypothesis class.

Remark. VC dimension is not the end all description of the excess risk incurred by learning. If $\mathcal{F}$ is the class of functions $f : [0,1] \to [0,1]$ with $\|f\|_{\text{Lip}} \leqslant L$, then similar techniques to our Lipschitz bound on Monte Carlo integration show

$$\mathbf{E}\, R(\widehat{f^*}) - R(f^*) \lesssim L/\sqrt{n}.$$

On the other hand, $\mathcal{F}$ has infinite VC dimension.

4.6 Generic Chaining

Using a slightly more technical chaining argument, we can obtain much tight bounds on the expectation of the supremum of a random process. To do this, we avoid using covering numbers at all. The idea is that, rather than fixing a number ε, and considering the minimum cardinality of a ε net, we instead consider a fixed cardinality N, and try and find the smallest ε such that there exists a ε net of cardinality N.

We say a sequence of sets $\{T_k\}$ is **admissible** if $|T_0| = 1$, and $|T_k| \leqslant 2^{2^k}$. This sequence of sets induces a sequence of values $\varepsilon_k = \sup_{t \in T} d(t, T_k)$. The chaining technique of the last section then shows

$$\mathbf{E} \sup_{t \in T} X_t \lesssim \sum_{k=0}^{\infty} 2^{k/2} \sup_{t \in T} d(t, T_k)$$

The important step is to take the supremum outside the sum, to obtain a single, universal quantity. The γ_2 **functional** is defined as

$$\gamma_2(T) = \inf_{\{T_k\}} \sup_{t \in T} \sum_{k=0}^{\infty} 2^{k/2} d(t, T_k).$$

Taking the supremum outside of the sum means that γ_2 is smaller than the bound obtained by Dudley's inequality. It can be minor, but in some cases it is a big difference.

Example. *Consider the example*

$$T = \left\{ \frac{e_k}{\sqrt{1 + \log k}} : 1 \leqslant k \leqslant n \right\}.$$

If we set T_k equal to the first 2^{2^k} vectors in T, then $T_k = T$ for $k \geqslant \lg\lg n$, and so

$$\begin{aligned}\gamma_2(T) &\leqslant \sup_{t\in T} \sum_{k=0}^{\lg\lg n} 2^{k/2} d(t, T_k) \\ &\lesssim \sup_{1\leqslant m\leqslant n} \sum_{k=2}^{\lg\lg m} \frac{2^{k/2}}{\sqrt{\log k}} \lesssim 1.\end{aligned}$$

On the other hand, we have already seen that Dudley's inequality gives a value which tends to ∞ as $n \to \infty$.

We now improve Dudley's inequality to give a result involving $\gamma_2(T)$.

Theorem 4.19 (Generic Chaining)**.** *Let $\{X_t\}$ be a mean zero random process. Then*

$$\mathbf{E}\left(\sup_{t\in T} X_t\right) \lesssim \|X\|_{\psi_2} \gamma_2(T).$$

Proof. TODO □

We say a process $\{X_t\}$ is *Gaussian* if all finite dimension distributions are Gaussian. Talagrand's majorizing measure theorem says that $\gamma_2(T)$ gives a tight bound for the expected supremum of the process in this case.

Theorem 4.20 (Talagrand's Majorizing Measures Theorem)**.** *Let $\{X_t : t \in T\}$ be a mean zero Gaussian process, inducing a canonical metric $d(t,s) = \|X_t - X_s\|_2$ on T. Then $\mathbf{E}(\sup X_t)$ is comparable to $\gamma_2(T)$.*

Since $\gamma_2(T)$ *upper bounds* any subgaussian process, the majorizing measures theorem often enables us to replace an arbitrary subgaussian process with a Gaussian process.

Corollary 4.21 (Talagrand's Comparison Inequality)**.** *Let $\{X_t : t \in T\}$ be a mean zero process and let $\{Y_t : t \in T\}$ be a Gaussian process. If $\|X_t - X_s\|_{\psi_2} \leqslant K\|Y_t - Y_s\|_2$, then*

$$\mathbf{E}(\sup X_t) \lesssim \mathbf{E}(\sup Y_t).$$

As a special case of this result, if $\{X_x : x \in T\}$ is a mean zero process with $T \subset \mathbf{R}^n$, then

$$\mathbf{E}(\sup X_x) \lesssim \|X\|_{\psi_2} \cdot w(T).$$

4.7 Chevet's Inequality

Talagrand's comparison inequality is often useful for studying subgaussian random variables by switching to studying Gaussian processes. Here, we give an application which gives a bound on

$$\sup\{Ax \cdot y : x \in T, y \in S\},$$

where A is random, and T and S are arbitrary sets. We now give a bound with respect to the Gaussian width of T and S, and in terms of the *radius* $\text{rad}(T) = \sup_{x \in T} |x|$ and $\text{rad}(S) = \sup_{y \in S} |y|$.

Theorem 4.22. *Let A be an $m \times n$ random matrix whose entries are independant, mean-zero, subgaussian random variables with $\|A_{ij}\|_{\psi_2} \leqslant K$. Let $T \subset \mathbf{R}^n$ and $S \subset \mathbf{R}^m$ be bounded sets. Then*

$$\mathbf{E} \sup\{Ax \cdot y : x \in T, y \in S\} \lesssim K(w(T) rad(S) + w(S) rad(T)).$$

Proof. We employ Talagrand's comparison inequality. Without loss of generality, let $K = 1$. We would like to bound $X_{uv} = Au \cdot v$, for $u \in T$ and $v \in S$. A simple calculation shows

$$\|X_{uv} - X_{wz}\|_{\psi_2} \leqslant |u - w| \text{rad}(S) + |v - z| \text{rad}(T).$$

Define $Y_{uv} = (g \cdot u)\text{rad}(S) + (h \cdot v)\text{rad}(T)$, where $g \sim N(0, I_n)$ and $h \sim N(0, I_m)$ are independant, Gaussian vectors. We calculate that

$$\|Y_{uv} - Y_{wz}\|_2^2 = |u - w|^2 \text{rad}(T)^2 + |v - z|^2 \text{rad}(S)^2.$$

Comparing these increments, we apply Talagrand's comparison inequality to conclude that

$$\begin{aligned} \mathbf{E} \sup X_{uv} \lesssim \mathbf{E} \sup Y_{uv} &= \mathbf{E} \sup (g \cdot u)\text{rad}(S) + \mathbf{E} \sup (h \cdot v)\text{rad}(T) \\ &= w(T)\text{rad}(S) + w(S)\text{rad}(T). \end{aligned}$$ □

Chapter 5

Gaussian Processes

We now try and derive lower bounds for **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 increments naturally give T the structure of a metric space, with $d(t,s) = 0$ if and only if $X_t = X_s$.

Example. *Let $\{X_t : t \geqslant 0\}$ be a Brownian motion. The metric induced on $[0,\infty)$ is given by $d(t,s) = |t-s|^{1/2}$. Similarily, if we consider independant normal random variables $Z_1, Z_2, \dots$ 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. In particular, this means that the finite dimensional distributions of the process are uniquely determined by the metric, if the variances of the random variables are known, and thus the expectation $\mathbf{E}\sup X_t$. This means we should be able to obtain bounds on the expectation purely from the geometry of the underlying metric space.

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 $s \in \mathbf{R}$,

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

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,*

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

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 $f(x) = h(x_1) \dots h(x_n)$ is an approximation of $\mathbf{I}(\max(x_1, \dots, x_n) < s)$. Then for $i \neq j$

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

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

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

Taking complements, we find

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

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 Inequality

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 $\alpha \to \infty$, $f_\alpha(x) \to \max(x_1,\dots,x_n)$. 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

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

so

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

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

Example. *If A is an $m \times n$ matrix with independant, standard normal entries, then the process $X_{xy} = y \cdot Ax$ is a Gaussian process on $S^{n-1} \times S^{m-1}$. If we define $Y_{xy} = g \cdot x + g' \cdot y$, where $g \sim N(0, I_n)$ and $g' \sim N(0, I_m)$ are independant from one another, then we find that*

$$\mathbf{E}((X_{xy} - X_{zw})^2) \leqslant |x - z|^2 + |y - w|^2 = \mathbf{E}((g \cdot x + g' \cdot y))^2.$$

Thus the Sudakov-Fernique inequality implies that

$$\begin{aligned} \mathbf{E} \sup X_{xy} \leqslant \mathbf{E} \sup Y_{xy} &= (\mathbf{E} \sup g \cdot x) + (\mathbf{E} \sup g' \cdot y) \\ &= \mathbf{E}\,|g| + \mathbf{E}\,|g'| = \sqrt{n} + \sqrt{m}. \end{aligned}$$

Thus $\mathbf{E}\,\|A\| \leqslant \sqrt{n} + \sqrt{m}$. We note that the map $A \mapsto \|A\|$ is 1-Lipschitz, so we can apply Gaussian concentration to conclude that there is an absolute constant c such that

$$\mathbf{P}\left(|\|A\| - \sqrt{n} - \sqrt{m}| \geqslant t\right) \leqslant \exp(-c \cdot t^2).$$

Thus the expectation calculation gives us a concentration bound automatically.

5.3 Gaussian Width

Before we move on, it is important to discuss the basic properties of Gaussian width, which will become very useful later on.

Theorem 5.6. *Let $T \subset \mathbf{R}^n$. Then*

- *$w(T)$ is finite if and only if T is bounded.*
- *Gaussian width is invariant under isometries of $\mathbf{R}^n$, i.e. $w(T) = w(U \cdot T + s)$ for any $U \in O(n)$ and $s \in \mathbf{R}^n$.*
- *The Gaussian width of a set is equal to the Gaussian width of it's convex hull, i.e. $w(T) = w(Conv(T))$.*
- *For any $T, S \subset \mathbf{R}^n$,*
$$w(T+S) = w(T) + w(S) \quad \text{and} \quad w(aT) = |a| w(T).$$
- *For any set $T \subset \mathbf{R}^n$,*
$$w(T) = \frac{w(T-T)}{2} = \frac{\mathbf{E} \sup_{x,y \in T} g \cdot (x-y)}{2}.$$
- *Forany set $T \subset \mathbf{R}^n$,*
$$\frac{diam(T)}{\sqrt{2\pi}} \leqslant w(T) \leqslant \frac{\sqrt{n}}{2} \cdot diam(T).$$
- *If A is an $m \times n$ matrix, then $w(AT) \leqslant \|A\| w(T)$.*

Proof. There exists a constant c such that for any particular $x \in \mathbf{R}^n$,

$$\mathbf{P}(g \cdot x \leqslant t|x|) \lesssim 1 - \exp(-c \cdot t^2)$$

consider a sequence a_n such that

$$\sum_{k=1}^{\infty} 1 - \exp(c \cdot a_n^2) < \infty.$$

If we consider a sequence $\{x_k\} \subset T$ with $|x_k| \geqslant n/a_n$, then a union bound combined with the Borel-Cantelli lemma implies that $g \cdot x_k \geqslant a_n|x_k| \geqslant n$ almost surely for sufficiently large k, which means

$$\sup g \cdot x \geqslant \sup g \cdot x_k = \infty \quad \text{a.s.}$$

So obviously $\mathbf{E} \sup g \cdot x = \infty$.

If $U \in O(n)$, and $y \in \mathbf{R}^n$, then

$$w(T)\, \mathbf{E} \sup g \cdot (Ux + y) = \mathbf{E} \sup (Ug) \cdot x + \mathbf{E}\, g \cdot y = \mathbf{E} \sup g \cdot x = w(T).$$

This proves the second point. Any element x of $\text{Conv}(X)$ can be written as $\sum \alpha_i x_i$ for $x_i \in X$, and $\sum \alpha_i = 1$. But then

$$g \cdot x = \sum \alpha_i (g \cdot x_i) \leqslant \max_i g \cdot x_i.$$

Thus $\sup_{x \in T} g \cdot x = \sup_{x \in \text{Conv}(T)} g \cdot x$, which proves the third point. Next, we calculate that

$$w(T + S) = \sup_{(x+y) \in T+S} g \cdot (x + y) = \sup_{x \in T} g \cdot x + \sup_{y \in T} g \cdot y = w(T) + w(S).$$

The fact that $w(\alpha T) = |\alpha| w(T)$ is obvious. Thus the fourth property is established. And

$$w(T - T) = w(T) + w(-T) = 2w(T),$$

so the fifth point is established.

To establish the lower bound $w(T) \geqslant \text{diam}(T)/\sqrt{2\pi}$, we may assume without loss of generality that T is compact. We then just take two points $x, y \in T$ with $|x - y| = \text{diam}(T)$, and calculate

$$\begin{aligned} w(T) &\geqslant 0.5 \cdot \mathbf{E} \max(g \cdot (x - y), g \cdot (y - x)) \\ &\geqslant 0.5 \cdot \mathbf{E}\, |g \cdot (x - y)| = 0.5 \cdot (2/\pi)^{1/2} \cdot |x - y| = (2\pi)^{-1/2} \text{diam}(T). \end{aligned}$$

Conversely, we have

$$\begin{aligned} w(T) &= 0.5 \cdot \mathbf{E} \left(\sup g \cdot (x - y) \right) \leqslant 0.5 \cdot \sup |g||x - y| \\ &\leqslant 0.5 \cdot \text{diam}(T) \sup |g| = 0.5 \cdot \sqrt{n} \cdot \text{diam}(T). \end{aligned}$$

Thus the fifth point has been established.

To establish the final point, we must show

$$\mathbf{E}\left(\sup_{x\in T} g \cdot Ax\right) \leqslant \|A\| \sup_{x\in T}(g \cdot x).$$

If we let $X_x = g \cdot x$, and $Y_x = g \cdot Ax$, then

$$\|Y_x - Y_y\|_2 = \|A^T g \cdot (x-y)\|_2 \leqslant \sqrt{m} \cdot \|A\||x-y| = \|A\|\|X_x - X_y\|_2.$$

Applying the Sudakov-Fernique inequality, we conclude

$$\mathbf{E} \sup Y_x \leqslant \|A\| \, \mathbf{E} \sup X_x = \|A\| w(T). \qquad \square$$

The Gaussian width of a set should be compared to the spherical width of a set, i.e.

$$w_s(T) = \mathbf{E}\left(\sup_{x\in T} \theta \cdot x\right),$$

where θ is a uniformly randomly chosen vector on the unit sphere. Because the distribution is translation invariant, the spherical width of a set also satisfies the translating invariance and homgeneity bounds that the Gaussian width also satisfies. The spherical width and Gaussian width are essentially comparable, once the spherical width is scaled by $\sqrt{n}$.

Theorem 5.7. *There exists an absolute constant C such that for any $T \subset \mathbf{R}^n$,*

$$w(T) = w_s(T)\left(\sqrt{n} + O(1)\right).$$

Proof. We note that g is uniformly distributed on the sphere once normalized, so $\sup g \cdot x$ is identically distributed to $|g| \sup \theta \cdot x$, where θ is uniformly distributed on the unit sphere and independant of g. Thus

$$w(T) = \mathbf{E} \sup g \cdot x = \mathbf{E}\,|g|\,\mathbf{E} \sup \theta \cdot x = \left(\sqrt{n} + O(1)\right) w_s(T). \qquad \square$$

Example. *If B is the unit ball, then*

$$\sup_{x\in B} g \cdot x = |g|$$

Thus

$$w(B) = \mathbf{E} \sup_{x\in B} g \cdot x = \mathbf{E}\,|g| = \sqrt{n} + O(1).$$

The unit ball has the same width as the unit sphere S^{n-1}.

Example. *Consider the unit cube in the l^∞ norm, i.e. $B^\infty = [-1,1]^n$. Then*

$$\sup_{x \in B^\infty} g \cdot x = \sup_{x \in \{-1,1\}^n} g \cdot x = |g_1| + \cdots + |g_n|.$$

Thus

$$w(B^\infty) = \mathbf{E}\,|g_1| + \cdots + \mathbf{E}\,|g_n| = \sqrt{2/\pi} \cdot n.$$

Thus B^∞ has almost the same width as the ball of radius $\sqrt{n}$ around the origin, which contains B^∞.

Example. *Consider the unit ball in the l^1 norm, i.e.*

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

Then

$$\sup_{x \in B^1} g \cdot x = \max(g_1, \dots, g_n),$$

and so

$$w(B^1) = \mathbf{E} \max(g_1, \dots, g_n) = \Theta\left((\log n)^{1/2}\right).$$

Thus B^1 has a proportionally tiny *width as compared to the unit balls in the l^2 and l^∞ norm. It has essentially the same width as the radius $1/\sqrt{n}$ ball in the l^2 norm that it contains. Thus we can conclude that the majority of the mass of the l^1 ball is concentrated near the origin.*

Example. *If X is a collection of n vertices with $diam(X) \leqslant 1$, then $w(X) \lesssim (\log n)^{1/2}$. We might as well assume $|x| = 1$ for all $x \in X$, because this would only increase $\sup g \cdot x$, and then we are just considering the supremum of n unit variance Gaussians, which gives the result.*

It is often useful to work with a squared version of Gaussian width, i.e. $h(T) = (\mathbf{E} \sup (g \cdot x)^2)^{1/2}$. This is equivalent to the Gaussian width, up to a constant factor.

Theorem 5.8. *For any $T \subset \mathbf{R}^n$,*

$$w(T - T) \leqslant h(T - T) \leqslant w(T - T) + O(diam(T)) \lesssim w(T - T).$$

In particular, $h(T - T)$ is comparable with $w(T)$ for any set T.

Proof. Because $T - T$ is symmetric, since the L^2 norm of a random variable is always greater than the L^1 norm,

$$\begin{aligned} h(T-T) &= \left(\mathbf{E} \sup_{x,y \in T} (g \cdot (x-y))^2 \right)^{1/2} \\ &= \left(\mathbf{E} \left(\sup_{x,y \in T} |g \cdot (x-y)| \right)^2 \right)^{1/2} \\ &\geqslant \mathbf{E} \sup_{x,y \in T} |g \cdot (x-y)| \\ &= \mathbf{E} \sup_{x,y \in T} g \cdot (x-y) = w(T-T). \end{aligned}$$

TODO: UPPER BOUND. □

It's also often useful to discuss the *Gaussian complexity* of a set T, which is defined as $\gamma(T) = \mathbf{E} \sup |g \cdot x|$. We obviously have $w(T) \leqslant \gamma(T)$, but we need T to be symmetric about the origin to conclude that $w(T) = \gamma(T)$. But they are closely related in general.

Lemma 5.9. *For any set T, and $y \in T$,*

$$(1/3) \cdot (w(T) + |y|) \leqslant \gamma(T) \leqslant 2(w(T) + |y|).$$

Proof. TODO □

5.4 Stable Dimension

If $T \subset \mathbf{R}^n$, it's algebraic dimension $\dim(T)$ is the dimensional of the smallest vector space containing T. The algebraic dimension is very unstable, since small changes in T can vastly change the dimension. The Gaussian width helps us come up with a geometric quantity which acts like the algebraic dimension, but is more stable under small pertubations. We define the *stable dimension* of a bounded set T as

$$d(T) = \frac{h(T-T)^2}{\text{diam}(T)^2} = \frac{\mathbf{E} \sup_{x,y \in T} (g \cdot (x-y))^2}{\text{diam}(T)^2} \sim \frac{w(T)^2}{\text{diam}(T)^2}.$$

The stable dimension is always bounded by the algebraic dimension, but can be significantly smaller.

Lemma 5.10. *For any T, $d(T) \leqslant \dim(T)$.*

Proof. Suppose T lies in a k dimensional subspace. Without loss of generality, we can assume it has diameter one. If Q denotes the orthogonal projection onto this k dimensional subspace, then

$$\mathbf{E} \sup_{x,y \in T} (g \cdot (x-y))^2 = \mathbf{E} \sup_{x,y \in T} (Qg \cdot (x-y))^2 \leqslant \mathbf{E}|Qg|^2 = k.$$

□

The intersection of a unit ball with a k dimensional plane has $d(T) = \dim(T)$. On the other hand, $d(T) \lesssim \log |T|$.

Example. *Let B be the unit ball in n dimensions, and let A be an $m \times n$ matrix. If A has singular value decomposition $\sum s_i u_i v_i^T$, then*

$$\sup_{x \in B} (g \cdot Ax)^2 = \sup_{x \in B} \sum s_i^2 (g \cdot u_i)^2 (x \cdot v_i)^2$$

This random variable is identically distributed to

$$\sup_{x \in B} \sum s_i^2 g_i^2 x_i^2 = \max(s_1^2 g_1^2, \dots, s_n^2 g_n^2)$$

TODO: FINISH. and so $h(AB) = \|A\|_F^2$.

The *stable rank* of an $m \times n$ matrix A if $r(A) = \|A\|_F^2 / \|A\|^2$. The normal rank of A is the algebraic dimension of $A(B)$, whereas the stable rank is the stable dimension of $A(B)$. Thus $r(A) \leqslant \operatorname{rank}(A)$ for any set A.

5.5 Sudakov's Minorization Inequality

We now establish a lower bound on the expectation of the supremum of a Gaussian process using covering numbers. Since Dudley's inequality upper bounds the supremum using covering numbers, we can fairly easily see how sharp the two inequalities are with respect to one another.

Theorem 5.11. *If $\{X_t\}$ is a centered, Gaussian process, then for any $\varepsilon > 0$,*

$$\mathbf{E} \sup X_t \gtrsim \varepsilon \sqrt{\log N(T, \varepsilon)}.$$

Proof. We deduce the result from the Sudakov-Fernique inequality. Assume that $N(T,\varepsilon) = N$, and consider a ε net S of cardinality N. We certainly have $\sup_{t\in T} X_t \geqslant \sup_{s\in S} X_s$. Define $Y_s = (\varepsilon 2^{-1/2})g_s$, where $\{g_s\}$ is a family of independant standard normal distributions. Then

$$\mathbf{E}(Y_s - Y_{s'})^2 = \varepsilon^2 \leqslant \mathbf{E}(X_s - X_{s'})^2$$

Thus the Sudakov Fernique inequality implies that

$$c\varepsilon\sqrt{\log N} \leqslant \mathbf{E}\sup Y_s \leqslant \mathbf{E}\sup X_s \leqslant \mathbf{E}\sup X_t.$$

This completes the proof. □

Example. *If P is a polytope with n vertices, then $w(P) \lesssim (\log n)^{1/2}$. But Sudakov's minorization inequality implies that $w(P) \gtrsim \varepsilon(\log(P,\varepsilon))^{1/2}$, so we conclude that $N(P,\varepsilon) \lesssim n^{1/\varepsilon^2}$.*

5.6 Two Sided Sudakov Inequality

There is a gap between Sudakov's minorization and Dudley's inequality. But when talking about coverings on $\mathbf{R}^n$, the gap is logarithmically large. The two-sided Sudakov inequality makes this precise.

Theorem 5.12. *Let $T \subset \mathbf{R}^n$ and define*

$$s(T) = \sup_{\varepsilon\geqslant 0} \varepsilon \cdot (\log N(T,\varepsilon))^{1/2}.$$

Then $s(T) \lesssim w(T) \lesssim \log(n) \cdot s(T)$.

Proof. TODO. □

5.7 Diameters of Random Projections

Our goal in this section is, given a random projection $Q : \mathbf{R}^n \to \mathbf{R}^m$, and some set T, does P shrink the diameter of T with high probability. For instance, if T is a finite set, then the Johnson-Lindenstrauss lemma shows that with high probability, provided $m \gtrsim \log|T|$,

$$\operatorname{diam}(Q(T)) \approx (m/n)^{1/2} \cdot \operatorname{diam}(T).$$

If T is infinite, then this need not be the case. For instance, if T is the unit ball in $\mathbf{R}^n$, then $Q(T)$ is the unit ball in $\mathbf{R}^m$, so $\text{diam}(Q(T)) = \text{diam}(T)$ for any projection Q. We now show that random projections *do* shrink the diameter of a set, but only up to the spherical width of the set.

Theorem 5.13. *Let T be a bounded set in $\mathbf{R}^n$, and $Q : \mathbf{R}^n \to \mathbf{R}^m$ a random projection onto an m dimensional subplace of $\mathbf{R}^n$. Then with probability at least $1 - 2e^{-m}$,*

$$diam(Q(T)) \lesssim w_s(T) + (m/n)^{1/2} diam(T).$$

Proof. Without loss of generality, assume $\text{diam}(T) \leqslant 1$. We note that if $z \in S^{m-1}$, then $Q^T z$ is uniformly distributed on S^{n-1}. Now

$$\text{diam}(Q(T)) = \sup_{x \in T-T} |Qx| = \sup_{x \in T-T} \max_{z \in S^{m-1}} (Qx \cdot z).$$

We now use a covering argument to bound this quantity. Let N be a $1/2$ net of S^{m-1} with $|N| \leqslant 5^m$. Then

$$\text{diam}(Q(T)) \leqslant 2 \max_{z \in N} \sup_{x \in T-T} (Q^T z \cdot x).$$

If we fix $z \in N$, then

$$\mathbf{E} \sup_{x \in T-T} (Q^T z \cdot x) = w_s(T - T) = 2 w_s(T).$$

Using the concentration inequality for Lipschitz functions on th esphere, the map $f(\theta) = \sup_{x \in T-T} x \cdot \theta$ has $\|f\|_{\text{Lip}} \leqslant 1$, because

$$\sup_{x \in T-T} x \cdot \eta = \sup_{x \in T-T} x \cdot \theta - x \cdot (\eta - \theta) \geqslant \sup_{x \in T-T} x \cdot \theta - |\eta - \theta|,$$

and so

$$|f(\theta) - f(\eta)| = \sup_{x \in T-T} x \cdot \theta - \sup_{x \in T-T} x \cdot \eta \leqslant |\eta - \theta|.$$

Therefore, we conclude

$$\mathbf{P}\left(\sup_{x \in T-T} (Q^T z \cdot x) \geqslant 2 w_s(T) + t \right) \leqslant 2 \exp(-cnt^2).$$

If we now take a union bound, we find

$$\begin{aligned}\mathbf{P}\left(\max_{z\in N}\sup_{x\in T-T}(Q^T z\cdot x)\geqslant 2w_s(T)+t\right)&\leqslant 2|N|\exp(-cnt^2)\\ &\leqslant 2\cdot 5^m\exp(-cnt^2).\end{aligned}$$

Choosing $t=C(m/n)^{1/2}$, where C is large enough, we can bound the probability above by $2e^{-m}$. And so

$$\mathbf{P}\left(\operatorname{diam}(Q(T))\geqslant 4w_s(T)+2C(m/n)^{1/2}\right)\leqslant e^{-m}. \qquad \square$$

We can equivalent write the result as saying

$$\operatorname{diam}(Q(T))\lesssim\max(w_s(T),(m/n)^{1/2}\cdot\operatorname{diam}(T)).$$

The threshold bound at which the spherical mean becomes important occurs when $w_s(T)=(m/n)^{1/2}\cdot\operatorname{diam}(T)$, so

$$m=\frac{nw_s(T)^2}{\operatorname{diam}(T)^2}\sim\frac{w(T)^2}{\operatorname{diam}(T)^2}\sim d(T),$$

where $d(T)$ is the *stable dimension* of T. Thus

$$\operatorname{diam}(Q(T))\lesssim\begin{cases}(m/n)^{1/2}\cdot\operatorname{diam}(T) & : m\geqslant d(T)\\ w_s(T) & : m\leqslant d(T)\end{cases}.$$

So the diameter shrinks under the projection up to when we project onto a set which has the same dimension as the stable dimension of the set.

Chapter 6

Deviations of Random Matrices

We now use all of the tools we have developed to show that for any $m \times n$ random matrix A, $|Ax| \approx \mathbf{E}|Ax|$ with high probability, for infinitely many values x. But what is the error rate if we take points x lying in some set T? The answer is the Gaussian complexity $\gamma(T)$. We reduce to the Gaussian case by applying Talagrand's comparison inequality.

Theorem 6.1. *Let A be an $m \times n$ matrix whose rows A_i are independant, isotropic, and subgaussian. Then the process $X_x = |Ax| - \sqrt{m} \cdot |x|$ has subgaussian increments, i.e.*

$$\|X_x - X_y\|_{\psi_2} \lesssim K^2 |x - y|,$$

where $K = \max \|A_i\|_{\psi_2}$.

Proof. Assume first that $|x| = 1$, and $y = 0$. Then we need only show

$$\left\| |Ax| - m^{1/2} \right\|_{\psi_2} \lesssim K^2.$$

But this follows because the coordinates $A_i \cdot x$ of Ax are independant with $\mathbf{E}(A_i \cdot x)^2 = 1$, and $\|A_i \cdot x\|_{\psi_2} \leqslant K$, so we can apply the concentration of norm theorem to yield the result.

Now we assume $|x| = |y| = 1$. We then have to prove that

$$\||Ax| - |Ay|\|_{\psi_2} \lesssim K^2 |x - y|.$$

We first prove a version of this for the squared process. TODO: Fill in details here. □

We then obtain our main result immediately.

Theorem 6.2. *Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic, and subgaussian random vectors. Then for any $T \subset \mathbf{R}^n$,*

$$\mathbf{E} \sup_{x \in T} \Big| |Ax| - m^{1/2} \cdot |x| \Big| \lesssim K^2 \gamma(T),$$

where $\gamma(T)$ is the Gaussian complexity, and $K = \max \|A_i\|_{\psi_2}$.

Proof. s □

Chapter 7

Applications of Random Processes

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. $f(x) = a_0 + a_1 x + \dots + a_m x^m$, then $f(A) = a_0 + a_1 A + \dots + a_m A^m$. 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 7.1 (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 7.2. *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 7.3. *Let $A_1, \dots, A_n$ be independant, mean zero, $m \times m$ symmetric random matrices with $\|A_i\| \leqslant K$. Then*

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

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 7.4. *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 7.5 (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 SECTION 5.5.

Chapter 8

Applications to Random Graphs

Let us now apply our results to problems in random graph theory. Here the most common model to study is the Erdös-Renyi model $G(n,p)$, which is a random graph on n vertices, when any two distinct vertices are connected independently with probability p. In applications, this model presents itself as the simplest model of large networks.

It is clear that the expected degree of each vertex in $G(n,p)$ is equal to $d = (n-1)p$. In fact, for *dense graphs*, i.e. where $d \gtrsim \log n$, with high probability we can guarantee that *all* vertices have vertex approximately equal to d.

Theorem 8.1. *There exists $c > 0$ such that if $G \sim G(n,p)$ with $d = (n-1)p \geqslant C \log n/\delta^2$, then with 99% probability, all vertices of G have degree between $(1-\delta)d$ and $(1+\delta)d$.*

Proof. For a vertex v in the graph. The degree of v is the sum of $n-1$ independent $\mathrm{Ber}(p)$ random variables. Applying a stanadrd Chernoff bound for random variables gives that for $\delta \in (0,1)$,

$$\mathbf{P}(|\deg(v) - d| \geqslant \delta d) \leqslant 2e^{-cd\cdot\delta^2}.$$

Taking a union bound over all n vertices, we conclude that the probability that all vertices have degree between $(1-\delta)d$ and $(1+\delta)d$ is at most

$$2ne^{-cd\delta^2}.$$

Thus if $d \geqslant (\ln n)/100c\delta^2$, the required result is obtained. □

One can also obtain results on more sparse graphs using concentration. For instance, that if $d \lesssim \log n$, then with 99% probability, all vertices of $G \sim G(n,p)$ have degree $O(\log n)$. If the graph is *very sparse*, i.e. $d \lesssim 1$, then with 99% probability, we can obtain a slight improvement here, i.e. that all vertices have degree $O(\log n/\log\log n)$. However, we should *not* expect sparse graphs to also be regular.

Theorem 8.2. *Let $G \sim G(n,p)$ be a random graph with expected degree $d = o(\log n)$. Then with 99% probability, G has a vertex with degree at least $10d$.*

Proof. Suppose we are given a vertex v, and a set of $n/3$ other vertices $W_v \subset V - \{v\}$. Let Y be the number of vertices in W_v which correspond to an edge to v in the graph G, then $(3/d) \cdot Y$ is the sum of Bernoulli random variables with mean one. The Poisson limit theorem thus implies that as $n \to \infty$, $(1/d) \cdot Y$ converges in distribution to a Poisson distribution. Thus in particular, we conclude that for suitably large n,

$$\mathbf{P}(Y \geqslant 10d) \geqslant \frac{e^{-30}}{30!}.$$

Denote the right hand side by δ. Now suppose we can find, for each vertex v, an edge set W_v containing $n/3$ elements such that there is no pair v and v' such that $v' \in W_v$ and $v \in W_{v'}$, then the events that G contains $10d$ edges in W_v are independent as v varies. Thus applying the result above, we conclude that the probability that a vertex has degree at least $10d$ is at least

$$1 - (1-\delta)^n.$$

If n is made large enough, we can make this quantity bigger than or equal to 99%. □

Chapter 9

Applications of High Dimensional Concentration

9.1 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 9.1 (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.

9.2 Covariance Estimation

Suppose we are analyzing data in high dimensions, represented as points $X_1, \dots, X_m$ 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 9.2. *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.