Senior Thesis in Mathematics

Random Vector Norms

Author:
Jackson Hurley

Advisor:
Stephan Ramon Garcia

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 14, 2023

Contents

Introduction

This thesis accompanies a paper [5] I co-wrote with my professors, Ángel Chávez and Stephan Ramon Garcia. In the paper, we introduce a new family of norms to measure matrices. The paper is technical and cannot be understood without first understanding some advanced linear algebra 1 and probability theory 3. Even then, its dense proofs require a fluency not common to undergraduates.

The central theorem of the paper, Theorem 100, proves that taking the moments of random variables multiplied by the eigenvalues of a Hermitian matrix produces norms on the set of Hermitian matrices. The paper then goes through several common distributions and makes norms using them before applying those norms to different kinds of matrices.

In Chapter 1, I provide the linear algebra necessary to understand the crux of the paper. I start by defining Hermitian matrices and prove that they have certain properties which will be important later. Then, I describe what it means for a matrix to be positive definite, so that I can later introduce "positive definite" random vectors. Third, I introduce the concept of a norm of a matrix, and I explain the requirements for a function to be a norm.

In Chapter 2, I prove that for any symmetric norm on $\mathbb{R}^n$ we can define a second norm on the set of Hermitian matrices by applying the first norm to the eigenvalues. As far as I or my advisor can tell, this is a novel result, and the class of norms it defines is larger than class of the random vector norms we are concerned with in the rest of the thesis.

In [1] and [5], Professor Chávez wrote a more abstract proof using algebraic objects called normal decomposition systems as described in [9]. His proof applied only to the norms described in Chapter 5. In earlier iterations of this thesis, I attempted to distill [9]'s somewhat dry proofs about general convex spaces into more concrete statements about Hermitian matrices comprehensible to a student of linear algebra. In doing so, I found that some of the steps could be proven directly using relatively accessible results from linear algebra without resorting to [9]'s formalisms, and that the proof was generally applicable to any norm on $\mathbb{R}^n$. Professors Garcia, Chávez, and I are considering including my new proof in another paper clarifying and expanding upon [5].

In Chapter 3, I cover the random part of random vector norms. I introduce probability distributions and two kinds of functions to describe them (probability distribution functions and cumulative distribution functions). Then I introduce moments and moment-generating functions, which will eventually be used to translate our norms into polynomials of a matrix's eigenvalues. While using the linear algebra concepts I described in Chapter 1, I define random vectors, and I prove that the second moment matrix of a random vector of independent, identically distributed random variables is positive definite. In defining all these concepts, I give examples of common distributions that show up in the paper, and I give special attention to the Pareto distribution, which is the subject of my own contribution to the paper.

In Chapter 4, I provide an introduction to measure theory and how it relates to the probability theory that I just described in Chapter 3. I redefine random variables and their moments in measure-theoretical terms that allow us to prove the triangle inequality of our norms using the Minkowski's Inequality on L^p norms.

Chapter 5 finally rigorously defines the class of norms induced by random vectors. It then goes on to clearly explain some of the interesting corollaries and applications that are quickly proven in the paper. These applications are intended to illustrate the main theorem. I provide examples and illustrations, and I discuss the norms for which the exponent p is odd or non-integer. Following [1], I also introduce Hunter's Theorem and show how we can use our norms to prove it.

In Chapter 6, I introduce the Pareto distribution, which was the subject of much of my contribution to [5]. Unfortunately, a tangential section on fractals and Pareto distributions, a chapter examining moment problems (given a list of real numbers, can we make a distribution with those moments?), and a potential application to economics, which I mentioned in my thesis presentation, all had to be cut for the sake of turning in a complete and bounded thesis in polynomial time.

0.1 Sources

Since this thesis follows the paper [5], my sources for the later chapters are also cited in the paper. While writing the linear algebra background chapter, I revisited the Advanced Linear Algebra textbook [7] from Professor Garcia's class, which I took here at Pomona. To define eigenvalue norms, I used [4], [8], and [14], but I knew what to look for from [9]. I then found a concise summary of all these results in [2]. For the probability chapter, I used my notes as well as the textbook [6] from Professor Radunskaya's Math 151 class. A more thorough grounding of probability theory in measure theory can be found in [3].

Chapter 1

Linear Algebra

This thesis defines a class of norms on the set of Hermitian matrices. Before we can do that, we need to understand what makes a function a norm and what makes a matrix Hermitian.

1.1 Norms

Definition 1. A *vector space* $\mathcal{V}$ is a set endowed with the operations addition and scalar multiplication.

Example 2. The set of $n \times n$ Hermitian matrices H_n is a vector space. For $A, B \in \mathrm{H}_n$, and $c \in \mathbb{R}$, the matrices $A + B$ (addition) and cA (scalar multiplication) are both Hermitian matrices.

Definition 3. An inner product $\langle \cdot, \cdot \rangle$ on a vector space $\mathcal{V}$ is a function from a couple X, Y in $\mathcal{V}$ to a field of scalars (in this thesis, either the real numbers or the complex numbers) that is symmetric, homogeneous, and distributive, and for which $\langle A, A \rangle$ is positive definite. That is, for $A, B, C \in \mathcal{V}$

1. $\langle A, A \rangle = \langle A, A \rangle$,
2. $\langle A, B + C \rangle = \langle A, B \rangle + \langle A, C \rangle$,
3. $\langle cA, B \rangle = c\langle A, B \rangle$ for $c \in \mathbb{R}$, and
4. $\langle A, A \rangle \geq 0$ with equality only at $A = 0$.

Example 4. In $\mathbb{C}^n$, we define the inner product such that for complex vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$,

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1 \bar{y_1} + x_2 \bar{y_2} + \cdots + x_n \bar{y_n} = \sum_{i=1}^{n} x_i \bar{y_i}.$$

Proposition 5. *The Cauchy-Schwarz inequality states that for any inner product* $\langle \cdot, \cdot \rangle$ *on an inner product space* $\mathcal{V}$*,*

$$|\langle X, Y \rangle|^2 \leq \langle X, X \rangle \langle Y, Y \rangle$$

For a proof, see 5.4 in [7].

Definition 6. A *norm* is a function from a vector space to the non-negative real numbers that is positive definite, homogeneous, and satisfies the triangle inequality.

In other words, to prove that a function $\|\cdot\|$ from a vector space $\mathcal{V}$ to the real numbers is a norm, one must prove that for all for all A and B in $\mathcal{V}$,

1. $\|A\|$ equals zero when $A = 0$ and is otherwise positive.
2. $\|cA\| = |c|\|A\|$ for any scalar $c \in \mathbb{R}$, and
3. $\|A + B\| \leq \|A\| + \|B\|$.

1.1.1 The Euclidean Norm in $\mathbb{R}^2$

The Euclidean norm on 2×1 vectors well illustrates these basic properties of norms. The Euclidean norm of a vector is the familiar notion of magnitude or length.

Definition 7. The *Euclidean norm* $\|\cdot\|_2$ in $\mathbb{R}^n$ is the square root of sum of the squares of the components. If $\mathbf{x} \in \mathbb{R}^2$,

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2}. \tag{1.1}$$

Proposition 8. *The function* $\|\cdot\|_2$ *is a norm.*

Proof. Let $\mathbf{x} \in \mathbb{R}^2$.

1. $\|\mathbf{x}\|_2$ equals zero when $\mathbf{x} = 0$ and is otherwise positive.

   Proof. $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2}$ is positive unless $x_1^2 = x_2^2 = 0$.

2. For any scalar c, $\|c\mathbf{x}\|_2 = |c| \, \|\mathbf{x}\|_2$.

   Proof. Let c be a real number.

   $$\|c\mathbf{x}\|_2 = \sqrt{(cx_1)^2 + (cx_2)^2} = \sqrt{c^2}\sqrt{x_1^2 + x_2^2} = |c| \, \|\mathbf{x}\|.$$

   Thus, $\|\cdot\|$ is homogeneous.

3. Let $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^2$. Then $\|\mathbf{x}+\mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$.

   Proof. In $\mathbb{R}^2$, the triangle inequality states that in the triangle formed by the sides $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{x}+\mathbf{y}$, the length of the side $\mathbf{x}+\mathbf{y}$ is less than or equal to the sum of the lengths of the other two sides. Just drawing a triangle shows that this has to be true, but we can prove it using the Law of Cosines. Let θ be the angle between $\mathbf{x}$ and $\mathbf{y}$. The Law of Cosines gives that

$$\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta.$$

   Since for any $\mathbf{x}$ and $\mathbf{y}$, the cosine of θ is always between -1 and 1,

$$\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta \leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\|$$

   Taking the square root of both sides leaves the triangle inequality.

$$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

Since it satisfies the three conditions, $\|\cdot\|_2$ is a norm. □

Example 9. Let $\mathbf{x} = (4,2)$ and let $\mathbf{y} = (-1,2)$. We calculate the norms of $\mathbf{x}$ and $\mathbf{y}$ using the formula in Equation 1.1.

$$\|\mathbf{x}\|_2 = \sqrt{4^2+2^2} = \sqrt{20}$$

$$\|\mathbf{y}\|_2 = \sqrt{(-1)^2+2^2} = \sqrt{5}.$$

Now let $c = \frac{1}{2}$. Then $c\mathbf{x} = (2,1)$.

$$\|c\mathbf{x}\|_2 = \sqrt{2^2+1^2} = \sqrt{5} = \frac{\sqrt{20}}{2} = |c|\|\mathbf{x}\|_2$$

This demonstrates homogeneity.

As we can see from the graph, $\mathbf{x}$ and $\mathbf{y}$ satisfy the triangle inequality.

$$\|\mathbf{x}+\mathbf{y}\| = \sqrt{3^2+4^2} = 5 < \|\mathbf{x}\| + \|\mathbf{y}\| = \sqrt{20} + \sqrt{5} \approx 6.708.$$

1.1.2 Unit Balls

A norm $\|\cdot\|$ on a vector space $\mathcal{V}$ is uniquely defined by its unit ball, the set of $u \in \mathcal{V}$ such that $\|u\| = 1$. This is true because every non-zero element x in $\mathcal{V}$ has a corresponding unit vector u such that $x = cu$ for a positive scalar c, and $\|x\| = c\|u\| = c$.

Proposition 10. *Let $x \in \mathcal{V}$. Then $x = \|x\|u$ for some $u \in \mathcal{V}$ such that $\|u\| = 1$.*

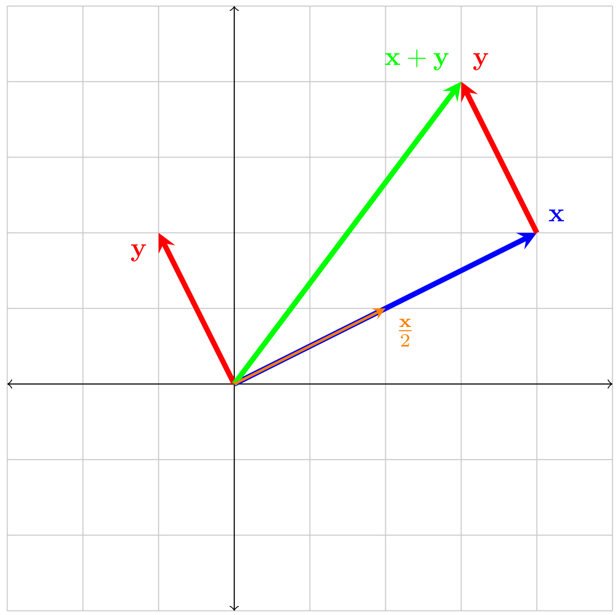


Figure 1.1: Illustration of the triangle inequality for $\mathbf{x} = (4, 2)$ and $\mathbf{y} = (-1, 2)$

Proof. If $x = 0$, $x = \|x\|u = 0u$ for any unit vector u. Let x be a nonzero element of $\mathcal{V}$. Since $\|\cdot\|$ is a norm, $\|x\| \neq 0$ by positive-definiteness. Let $u = \frac{x}{\|x\|}$. Then $\|u\| = \frac{1}{\|x\|}\|x\| = 1$, and $x = \|x\|u$. □

Example 11. Let $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ The p-norm on $\mathbb{R}^n$ is defined

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}.$$

In $\mathbb{R}^2$, this is

$$\sqrt[p]{|x_1|^p + |x_2|^p}.$$

The unit circles of the norms $\|\mathbf{x}\|_p$ for p = 1, 2, 3, and ∞ are shown in Figure 1.2.

1.1.3 Convexity

Definition 12. A function $f : \mathcal{V} \to \mathbb{R}$ is *convex* if and only if for all t such that $0 \leq t \leq 1$ and $A, B \in \mathcal{V}$,

$$f(tA + (1-t)B) \leq tf(A) + (1-t)f(B).$$

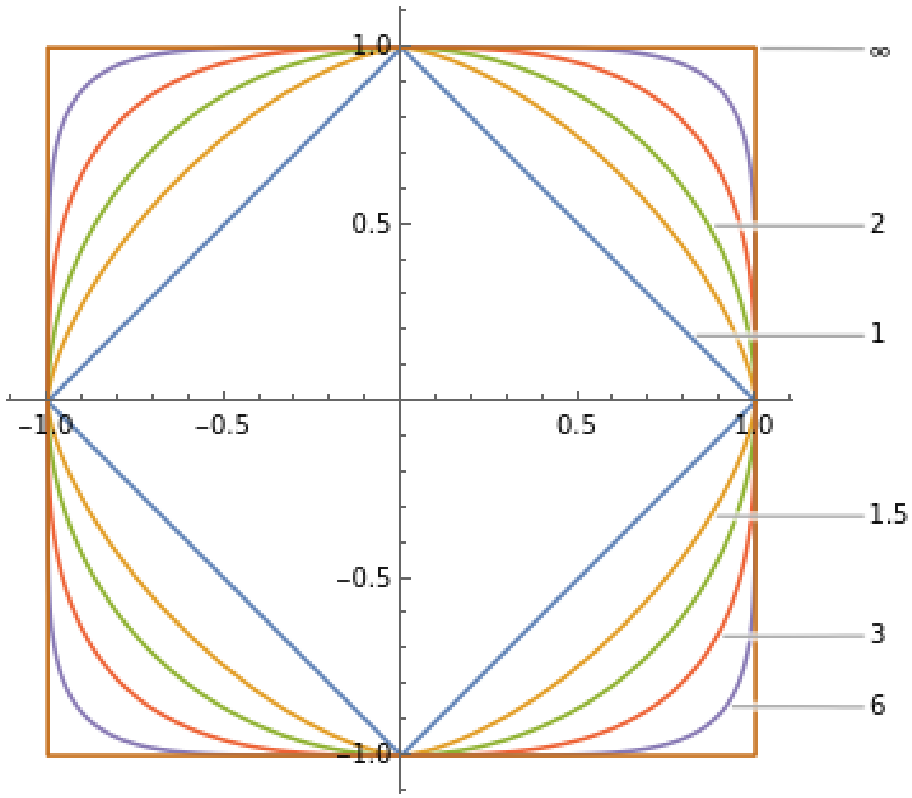


Figure 1.2: Unit circles of p-norms in $\mathbb{R}^2$. From the inside diamond to the outside square, $\|\cdot\|_1$, $\|\cdot\|_{1.5}$, $\|\cdot\|_2$, $\|\cdot\|_3$, $\|\cdot\|_6$, $\|\cdot\|_\infty$.

Proposition 13. *Let $f : \mathcal{V} \to \mathbb{R}$ be homogeneous. Then f is convex if and only if f satisfies the triangle inequality.*

Proof. ($\Longrightarrow$) Let f be homogeneous and convex. The case where $t = 1/2$ guarantees the triangle inequality, since

$$f(A+B) = f(\frac{2A}{2} + \frac{2B}{2}) \leq \frac{f(2A)}{2} + \frac{f(2B)}{2} = f(A) + f(B),$$

($\Longleftarrow$) Let f be homogeneous and satisfy the triangle inequality. Then

$$f(tA + (1-t)B) \leq f(tA) + f((1-t)B) = tf(A) + (1-t)f(B).$$

Hence f is convex. □

1.2 Hermitian Matrices

1.2.1 Conjugate Transpose

Definition 14. Let $A \in \mathrm{M}_{m\times n}$ be a complex matrix with entries $a_{ij} = \mathrm{Re}\,(a_{ij}) + i\,\mathrm{Im}\,(a_{ij})$. The *conjugate transpose* of A, denoted A^*, is the $n \times m$ matrix with entries $\overline{a_{ij}} = \mathrm{Re}\,(a_{ji}) - i\,\mathrm{Im}\,(a_{ji})$.

1.2.2 Defining Hermitian Matrices

Definition 15. A square matrix A is *Hermitian* if $A = A^*$.

Example 16. The following matrix is Hermitian.

$$\begin{bmatrix} 2 & 5+i & -3i \\ 5-i & -8 & 0 \\ 3i & 0 & 4 \end{bmatrix}$$

The reason we focus on Hermitian matrices in this thesis is that though Hermitian matrices can be complex, the eigenvalues of Hermitian matrices are always real. This result will allow us to create norms on the set of Hermitian matrices (which are functions from the set of Hermitian matrices to the real numbers) using polynomials of the eigenvalues of Hermitian matrices.

Theorem 17. *The eigenvalues of a Hermitian matrix are real.*

Proof. Let λ be the eigenvalue of a Hermitian matrix A associated with an eigenvector $\mathbf{x}$. Then

$$A\mathbf{x} = \lambda\mathbf{x}.$$

Multiplying both sides from the right by the conjugate transpose of $\mathbf{x}$, we obtain

$$\mathbf{x}^* A\mathbf{x} = \mathbf{x}^*\lambda\mathbf{x} = \lambda\|\mathbf{x}\|^2.$$

Taking the conjugate transpose of both sides gives us

$$\mathbf{x}^* A^* \mathbf{x} = \overline{\lambda} \|\mathbf{x}\|^2.$$

Since $A = A^*$, we can replace the A^* with A. We conclude that $\lambda = \bar{\lambda}$, so λ is real. □

1.2.3 Spectral Theorem

See [7] Theorem 14.2.2.

Definition 18. A matrix U is *unitary* if $UU^* = I$. In other words, the columns $\mathbf{x}_i$ of U form an orthonormal basis in $\mathbb{C}_n$, since

$$(UU^*)_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Lemma 19. *Let A be Hermitian and let $A\mathbf{x} = \lambda_1 \mathbf{x}$ and $A\mathbf{y} = \lambda_2 \mathbf{y}$ with $\lambda_1 \neq \lambda_2$. Then $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.*

Proof. Since $A = A^*$, the conjugate transpose of $A\mathbf{y} = \lambda_2 \mathbf{y}$ is $\mathbf{y}^* A = \mathbf{y}^* \lambda_2$. Multiplying on the right by $\mathbf{x}$,

$$\mathbf{y}^* A \mathbf{x} = \mathbf{y}^* \lambda_2 \mathbf{x} = \lambda_2 \langle \mathbf{x}, \mathbf{y} \rangle. \tag{1.2}$$

Since $A\mathbf{x} = \lambda_1 \mathbf{x}$,

$$\mathbf{y}^* A \mathbf{x} = \mathbf{y}^* \lambda_1 \mathbf{x} = \lambda_1 \langle \mathbf{x}, \mathbf{y} \rangle. \tag{1.3}$$

Subtracting 1.3 from 1.2,

$$(\lambda_2 - \lambda_1) \langle \mathbf{x}, \mathbf{y} \rangle = 0.$$

Since $\lambda_1 \neq \lambda_2$, $\langle \mathbf{x}, \mathbf{y} \rangle$ must be 0. □

Definition 20. Define the *eigenvalue vector* $\boldsymbol{\lambda}(\cdot) : \mathrm{H}_n \to \mathbb{R}^n$ such that for $A \in \mathrm{H}_n$, $\boldsymbol{\lambda}(A)$ is the vector of the eigenvalues of A in non-increasing order. In other words, $\boldsymbol{\lambda}(A) = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

Proposition 21. *Let A be a Hermitian matrix, and let*

$$\Lambda = \operatorname{diag}(\boldsymbol{\lambda}) = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

Then A is unitarily diagonalizable. Specifically, there is a unitary $U \in \mathrm{M}_n$ such that

$$A = U \Lambda U^*.$$

Proof. Let the *eigenspace* E_{λ_k} refer to the null space of $A - \lambda_k I$. A vector $\mathbf{x}$ is in E_{λ_k} if and only if $\mathbf{x}$ is either the zero vector or an eigenvector associated with λ_k, since if $(A - \lambda_k I)\mathbf{x}$, then $A\mathbf{x} = \lambda_k \mathbf{x}$.

By Lemma 19, any two nonzero vectors in the eigenspaces corresponding to distinct eigenvalues are orthogonal, so the eigenspaces are orthogonal.

Since the dimension of an eigenspace E_{λ_k} is the multiplicity of λ_k in $\boldsymbol{\lambda}$, we can choose $\dim E_{\lambda_k}$ orthogonal vectors $\mathbf{u}_i$ in E_{λ_k}. If we thus choose $\dim E_{\lambda_k}$ orthonormal eigenvectors for each λ_k, we end up with n orthonormal $\mathbf{u}_i$. Like any set of n independent vectors, the $\mathbf{u}_i$ form an orthonormal basis of $\mathbb{C}^n$.

We can now define U as the matrix with columns $\mathbf{u}_i$. Since $\mathbf{u}_i$ are eigenvectors associated with λ_i,

$$A\mathbf{u}_i = \lambda_i \mathbf{u}_i = \mathbf{u}_i \lambda_i.$$

So $AU = U\Lambda$, since $U\Lambda$ multiplies the ith column of U by λ_i. Multiplying on the right by U^*,

$$AUU^* = U\Lambda U^*.$$

Since U is unitary, $UU^* = I$, so $AUU^* = A = U\Lambda U^*$. $\square$

1.3 Positive Definiteness

To make our norms, we will need functions that take vectors of the eigenvalues of a matrix to non-negative real numbers. We do this by using positive definite random matrices. Before we get into random matrices, it is worth recalling what is meant by a positive definite matrix in the first place.

Definition 22. A Hermitian matrix $A \in \mathrm{H}_n(\mathbb{C})$ is *positive definite* if $\langle A\mathbf{x}, \mathbf{x}\rangle > 0$ for all nonzero vectors $\mathbf{x}$ in $\mathbb{C}^n$.

An equivalent definition is that a positive definite matrix is a Hermitian matrix with positive eigenvalues [7] (15.1.3).

Proposition 23. *A Hermitian matrix A is positive definite if and only if all its eigenvalues are positive.*

Proof. ($\Longrightarrow$) Let A be positive definite and let λ be an eigenvalue of A. For some non-zero $\mathbf{x}$, $A\mathbf{x} = \lambda\mathbf{x}$. Multiplying each side by $\mathbf{x}^*$,

$$\langle A\mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^* A\mathbf{x} = \mathbf{x}^*\lambda\mathbf{x} = \lambda\langle \mathbf{x}, \mathbf{x}\rangle.$$

Since $\langle \mathbf{x}, \mathbf{x}\rangle$ and $\langle A\mathbf{x}, \mathbf{x}\rangle$ are positive, $\lambda = \frac{\langle A\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}$ is positive.
($\Longleftarrow$) Let A be a Hermitian matrix with positive eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. By Proposition 21 A has the decomposition $A = U\Lambda U^*$, where

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n).$$

Then

$$\mathbf{x}^* A\mathbf{x} = \mathbf{x}^* U \Lambda U^* \mathbf{x} = (U^*\mathbf{x})^* \Lambda (U^*\mathbf{x}) = \sum_{i=1}^{n} \overline{(U^*\mathbf{x})_i} (U^*\mathbf{x})_i \lambda_i.$$

Since the λ_i are positive and $\overline{(U^*\mathbf{x})_i}(U^*\mathbf{x})_i$ is positive, $\mathbf{x}^* A\mathbf{x}$ is positive. □

Definition 24. A Hermitian matrix $A \in \mathrm{H}_n(\mathbb{C})$ is *positive semi-definite* if $\mathbf{x}^* A\mathbf{x} \geq 0$ for all nonzero vectors $\mathbf{x}$ in $\mathbb{C}^n$.

Remark 25. By a similar proof to 1.3, we can prove that all the eigenvalues of a positive semi-definite matrix are real and non-negative.

Proposition 26. *The sum of a positive definite and a positive semi-definite matrix is positive definite.*

Proof. Let A be positive definite and B positive semi-definite. Let $\mathbf{x} \neq 0$. Then

$$\mathbf{x}^*(A + B)\mathbf{x} = \mathbf{x}^* A\mathbf{x} + \mathbf{x}^* B\mathbf{x} \geq \mathbf{x}^* A\mathbf{x} + 0 > 0.$$

So $A + B$ is positive definite. □

1.4 Matrix Norms

Since the set of $n \times n$ complex-valued matrices is closed under addition and scalar multiplication, it is a vector space, so we can also define norms on it.

1.4.1 Norms on Hermitian Matrices

This thesis concerns a class of norms defined on the set of Hermitian matrices, which is a subspace of the vector space of complex-valued matrices. Certain properties of Hermitian matrices make it easier to define norms on them than on the larger set of square matrices. The first useful property will be helpful with establishing the positive-definiteness of our norms.

Proposition 27. *If the eigenvalues of a Hermitian matrix are zero, that matrix equals zero.*

Proof. Let $A \in \mathrm{H}_n$, and let all the eigenvalues of A be zero. Since A is Hermitian, for some unitary $U \in \mathrm{M}_n$, $A = U\Lambda U^*$, where $\Lambda = \mathrm{diag}(0, 0, \ldots, 0) = 0$. Thus, $A = U0U^* = 0$. □

1.4.2 Singular Values

Definition 28. Let $A \in \mathrm{M}_n$. The *singular values* of A are the square roots of the eigenvalues of the matrix A^*A.

Proposition 29. *Singular values exist and are real and non-negative.*

Proof. Let $A \in \mathrm{M}_n$. Since the eigenvalues of positive semi-definite matrices are non-negative, it suffices to show that A^*A is positive semi-definite. Let $\mathbf{x}$ be a vector in $\mathbb{C}^n$.

$$\langle A^*A\mathbf{x}, \mathbf{x}\rangle = \langle A^*(A\mathbf{x}), \mathbf{x}\rangle = \langle A\mathbf{x}, A\mathbf{x}\rangle.$$

The inner product of a vector $\mathbf{x}$ with itself is always real and non-negative, and equal to zero only if $\mathbf{x} = 0$. So A^*A is positive semi-definite. Thus, the square roots of its eigenvalues, which are the singular values of A, exist and are real and non-negative. □

Because they are so well-behaved, many of the most well-known matrix norms can be defined using symmetric functions of singular values.

Proposition 30. *The singular values of a matrix A are zero if and only if A is the zero matrix.*

Proof. Let $A \in \mathrm{M}_n$.
($\Longrightarrow$) If $A = 0$, $A^*A = 0$, so the singular values of A, which are the eigenvalues of A^*A, are also zero.
($\Longleftarrow$) Let the singular values of A be zero. Then the eigenvalues of A^*A are zero. Since A^*A is positive definite, it is Hermitian. By Proposition 27, since its eigenvalues are zero, $A^*A = 0$.

Let $\mathbf{x}$ be an arbitrary vector in $\mathbb{C}^n$. Since $A^*A = 0$,

$$\langle \mathbf{x}, A^*A\mathbf{x}\rangle = \langle A\mathbf{x}, A\mathbf{x}\rangle = 0.$$

So $A\mathbf{x} = 0$. Since $\mathbf{x}$ is arbitrary, $A = 0_{n\times n}$. □

Proposition 31. *Singular values are homogeneous.*

Proof. Let $A \in \mathrm{M}_{m\times n}$, and let c be a non-negative real number. If the eigenvalues of A^*A are $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, where σ_i are singular values of A, $(cA)^*(cA) = c^2A^*A$ has eigenvalues $c^2\sigma_1^2, c^2\sigma_2^2, \ldots, c^2\sigma_n^2$.

So cA has singular values $c\sigma_1^2, c\sigma_2^2, \ldots, c\sigma_n^2$. □

Since the singular values are already positive definite and homogeneous, it is relatively straightforward to define norm functions using them.

The nuclear norm $\|A\|_1$ simply sums the singular values.

$$\|A\|_1 = \sigma_1 + \sigma_2 + \cdots + \sigma_n$$

The spectral norm $\|\cdot\|_\infty$ returns the largest singular value.

Perhaps most familiar to newer students of linear algebra is the Frobenius norm $\|\cdot\|_2$, which is equivalent to the Euclidean norm in $\mathbb{R}^n$ on the vector of singular values. .

Chapter 2

Eigenvalue Norms

Our norms use neither the singular values nor the entries of Hermitian matrices. Instead, they are functions of the eigenvalues. Since the eigenvalues of Hermitian matrices are not necessarily already positive, our norms will have to be positive definite functions.

Norms on the set of Hermitian matrices that are, like ours, defined using eigenvalues, correspond to norms on $\mathbb{R}^n$.

Definition 32. A function f on $\mathbb{R}^n$ is *symmetric* if $f(\mathbf{x}) = Px$ for all $\mathbf{x} \in \mathbb{R}^n$ and $n \times n$ permutation matrices P.

Example 33. If $n = 3$, f is symmetric if

$$\begin{aligned} f(x_1, x_2, x_3) &= f(x_1, x_3, x_2) = f(x_2, x_1, x_3) \\ &= f(x_2, x_3, x_1) = f(x_3, x_1, x_2) = f(x_3, x_2, x_1). \end{aligned}$$

Theorem 34. *Let $\|\cdot\|_{\mathrm{H}}$ be a norm on the Hermitian matrices such that $\|\cdot\|_{\mathrm{H}} = \|\boldsymbol{\lambda}(\cdot)\|$ for some symmetric function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$. Then $\|\cdot\|$ is a norm on $\mathbb{R}^n$.*

Remark 35. See 20 for the definition of the function $\boldsymbol{\lambda}(\cdot) : \mathrm{H}_n \to \mathbb{R}^n$.

Proof. Let $\mathbf{a} \in \mathbb{R}^n$ with $\mathbf{a} = a_1, a_2, \ldots, a_n$. Define the matrix $\operatorname{diag} \mathbf{a} \in \mathrm{H}_n$

$$\operatorname{diag} \mathbf{a} = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{bmatrix}.$$

The eigenvalues of a diagonal matrix are its diagonal entries, so $\boldsymbol{\lambda}(\operatorname{diag} \mathbf{a}) = P\mathbf{a}$ for some permutation matrix P. Since $\|\cdot\|$ is symmetric,

$$\|\mathbf{a}\| = \|P\mathbf{a}\| = \|\boldsymbol{\lambda}(\operatorname{diag} \mathbf{a})\| = \|\operatorname{diag} \mathbf{a}\|_{\mathrm{H}}.$$

Since $\|\cdot\|_{\mathrm{H}}$ is a norm on the Hermitian matrices, and since addition and scalar multiplication work the same for vectors and diagonal matrices, $\|\cdot\|$ inherits all the properties of a norm in $\mathbb{R}^n$:

1. $\mathbf{a} = 0 \iff \operatorname{diag}\mathbf{a} = 0_{n\times n}$. Since $\|\cdot\|_{\mathrm{H}}$ is positive definite, $\|\mathbf{a}\| = \|\operatorname{diag}\mathbf{a}\|_{\mathrm{H}}$ is positive definite.

2. $\|c\mathbf{a}\| = \|\operatorname{diag} c\mathbf{a}\|_H = \|c\operatorname{diag}\mathbf{a}\|_H = |c|\|\operatorname{diag}\mathbf{a}\|_{\mathrm{H}} = |c|\|\mathbf{a}\|$.

3. $\|\mathbf{a}+\mathbf{b}\| = \|\operatorname{diag}(\mathbf{a}+\mathbf{b})\|_{\mathrm{H}} \leq \|\operatorname{diag}(\mathbf{a})\|_{\mathrm{H}} + \|\operatorname{diag}(\mathbf{b})\|_{\mathrm{H}} = \|\mathbf{a}\| + \|\mathbf{b}\|$. □

We now proceed to prove the reverse: not only does every norm on the set H_n defined using eigenvalues induce a norm on $\mathbb{R}^n$, but also, every norm on $\mathbb{R}^n$ induces a norm on H_n.

Theorem 36. *Let $\|\cdot\|$ be a symmetric norm on $\mathbb{R}^n$, and define $\boldsymbol{\lambda}(\cdot) : \mathrm{H}_n \to \mathbb{R}^n$ such that $\boldsymbol{\lambda}(A) = (\lambda_1(A), \cdots, \lambda_n(A))$ is the vector of the eigenvalues of A in non-increasing order. Then $\|\cdot\|_{\mathrm{H}} = \|\boldsymbol{\lambda}(\cdot)\|$ is a norm on the set of Hermitian matrices H_n.*

Positive Definiteness

Proof. By 27, $A = 0 \iff \boldsymbol{\lambda}(A) = 0$. Thus, if $\|\cdot\|$ is positive definite, so is $\|\cdot\|_{\mathrm{H}} = \|\boldsymbol{\lambda}(\cdot)\|$. □

Homogeneity

Proof. All matrices satisfy $\boldsymbol{\lambda}(cA) = c\boldsymbol{\lambda}(A)$, so if $\|\cdot\|$ is homogeneous,

$$\|cA\|_{\mathrm{H}} = \|\boldsymbol{\lambda}(cA)\| = \|c\boldsymbol{\lambda}(A)\| = |c|\|\boldsymbol{\lambda}(A)\| = |c|\|A\|_{\mathrm{H}}.$$

Thus, $\|\cdot\|_{\mathrm{H}}$ is homogeneous. □

Triangle Inequality

As is often the case, the triangle inequality is hardest to prove. I wrote this whole section before finding it, but Chapter II of [2] includes everything until Theorem 47.

2.1 Majorization

Definition 37. Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\mathbb{R}^n$ with non-increasing elements. Then $\mathbf{x}$ *majorizes* $\mathbf{y}$ if

$$\sum_{i=1}^{k} x_i \geq \sum_{i=1}^{k} y_i$$

for all $1 \leq k \leq n$, and

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i.$$

When $\mathbf{x}$ majorizes $\mathbf{y}$, we write $\mathbf{x} \succ \mathbf{y}$.

Example 38. The vector $\mathbf{x} = (5, 2, -1)$ majorizes $\mathbf{y} = (3, 3, 0)$.

Definition 39. A matrix D is *doubly stochastic* if the sum of each row equals 1, the sum of each column equals 1, and all entries are non-negative.

The following two lemmas are due to Hardy, Littlewood, and Pólya (1929) [8] and Birkhoff (1946) [4], respectively. An elegant constructive proof of Lemma 40 can be found in [15].

Lemma 40. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and let $\mathbf{x}$ majorize $\mathbf{y}$. Then there exists a doubly stochastic matrix D such that $\mathbf{y} = D\mathbf{x}$.*

Lemma 41. *Let D be an $n \times n$ doubly stochastic matrix. Then there exist permutation matrices P_i and non-negative coefficients c_i such that $\sum_{i=1}^{n^2} c_i = 1$ and*

$$D = \sum_{i=1}^{n^2} c_i P_i.$$

Proof. We can construct our c_i and P_i by repeatedly subtracting off permutation matrices scaled to the smallest non-zero entries. Since these non-zero entries are less than or equal to the sum of each row and column, and permutation matrices have rows and columns that sum to 1, subtracting off each $c_i P_i$ leaves a matrix that still has non-zero entries whose rows and columns all have the same sum. We then repeat this process at most n^2 times, since each time we subtract a $c_i P_i$, we eliminate at least one non-zero element, and there are at most n^2 non-zero elements in an $n \times n$ matrix to begin with. The following example is helpful. ☐

Example 42. Consider the doubly stochastic matrix

$$D = \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.3 & 0.2 & 0.5 \\ 0.1 & 0.4 & 0.5 \end{bmatrix}.$$

We start Birkhoff's algorithm by subtracting off the lowest non-zero entry $c_1 = 0.1$ times a permutation matrix whose 1s correspond to non-zero elements.

$$D - c_1 P_1 = D - 0.1 \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.6 & 0.3 & 0 \\ 0.3 & 0.2 & 0.4 \\ 0 & 0.4 & 0.5 \end{bmatrix}.$$

Here, all entries are still non-negative and all the rows and columns sum to $1 - 0.1 = 0.9$. The new smallest nonzero element is 0.2. Subtracting off $c_2 = 0.2$ times another permutation matrix that includes the 0.2, we get

$$D - c_1 P_1 - 0.2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.4 & 0.3 & 0 \\ 0.3 & 0 & 0.4 \\ 0 & 0.4 & 0.3 \end{bmatrix}.$$

All rows and columns now sum to $1 - 0.1 - 0.2 = 0.7$. We repeat the process twice more, with $c_3 = 0.3$ and $c_4 = 0.4$, until we conclude that $D - (c_1 P_1 + c_2 P_2 + c_3 P_3 + c_4 P_4) = 0$.

Definition 43. Let $A \in \mathrm{H}_n$. Let V be an m-dimensional subspace of $\mathbb{C}^n$. The *partial trace*, denoted $\operatorname{tr} A\,|_V$, is the sum

$$\operatorname{tr} A\,|_V = \sum_{i=1}^{m} v_i^* A v_i, \tag{2.1}$$

where $v_1, v_2, \ldots, v_m$ is any orthonormal basis of V.

The following Lemma follows Terence Tao's blog post [14].

Lemma 44. *Let $A \in \mathrm{H}_n$ and let $\boldsymbol{\lambda}(A)$ be the vector of eigenvalues of A in non-increasing order. Then for $1 \leq k \leq n$, and subspaces V of $\mathbb{C}^n$,*

$$\sum_{i=1}^{k} \lambda_i(A) = \lambda_1(A) + \cdots + \lambda_k(A) = \sup_{\dim(V)=k} \operatorname{tr} A\,|_V.$$

Corollary 45. *For all $A, B \in \mathrm{H}_n$ and $1 \leq k \leq n$, the extreme partial trace satisfies the triangle inequality.*

$$\lambda_1(A+B) + \cdots + \lambda_k(A+B) \leq \lambda_1(A) + \cdots + \lambda_k(A) + \lambda_1(B) + \cdots + \lambda_k(B).$$

Proof. By Lemma 44,

$$\sum_{i=1}^{k} \lambda_i(A+B) = \sup_{\dim(V)=k} \operatorname{tr}(A+B)\,|_V$$

Using 2.1, we can distribute.

$$\operatorname{tr}(A+B)\,|_V = \sum_{i=1}^{m} v_i^*(A+B)v_i = \sum_{i=1}^{m} (v_i^* A v_i + v_i^* B v_i) = \operatorname{tr} A\,|_V + \operatorname{tr} B\,|_V.$$

The supremum function is convex, so

$$\sup_{\dim(V)=k} (\operatorname{tr} A\,|_V + \operatorname{tr} B\,|_V) \leq \sup_{\dim(V)=k} \operatorname{tr} A\,|_V + \sup_{\dim(V)=k} \operatorname{tr} B\,|_V$$

Using 2.1 again, we get

$$\sup_{\dim(V)=k} \operatorname{tr} A\,|_V + \sup_{\dim(V)=k} \operatorname{tr} B\,|_V = \sum_{i=1}^{k} \lambda_i(A) + \sum_{i=1}^{k} \lambda_i(B)$$

Thus, the function $\sum_{i=1}^{k} \lambda_i(\cdot)$ satisfies the triangle inequality on H_n. □

Theorem 46. *Let $\boldsymbol{\lambda}(A)$ denote the vector eigenvalues of a Hermitian matrix A ordered from largest to smallest. Then $\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B)$ majorizes $\boldsymbol{\lambda}(A+B)$ for all $A, B \in \mathrm{H}_n$.*

Proof. The inequality when $k \leq n$ follows directly from Corollary 45. For all $1 \leq k \leq n$,

$$\sum_{i=1}^{k} \lambda_i(A+B) \leq \sum_{i=1}^{k} \lambda_i(A) + \lambda_i(B).$$

At n, we have equality, since the sum of the eigenvalues of a matrix is its trace:

$$\sum_{i=1}^{n} \lambda_i(A+B) = \operatorname{tr}(A+B) = \operatorname{tr} A + \operatorname{tr} B = \sum_{i=1}^{n} \lambda_i(A) + \lambda_i(B).$$

Therefore, $\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B)$ majorizes $\boldsymbol{\lambda}(A+B)$. $\square$

Theorem 47. *We can write $\boldsymbol{\lambda}(A+B)$ as the sum of non-negative constants c_i times permutation matrices P_i times $\boldsymbol{\lambda}(A)+\boldsymbol{\lambda}(B)$, where the sum of the c_i is 1:*

$$\boldsymbol{\lambda}(A+B) = \sum_{i=1}^{n!} c_i P_i(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B)).$$

Proof. Since $\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B)$ majorizes $\boldsymbol{\lambda}(A+B)$, Lemma 40 ensures that we can produce a stochastic matrix D such that

$$\boldsymbol{\lambda}(A+B) = D(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B)).$$

From Lemma 41, we can express D as a sum of permutations times non-negative constants adding to 1. $\square$

Triangle Inequality

Proof. Let $A, B \in \mathrm{H}_n$. By Lemma 47,

$$\|A+B\|_{\mathrm{H}} = \|\boldsymbol{\lambda}(A+B)\| = \|\sum_{i=1}^{n!} c_i P_i(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\|.$$

Since $\|\cdot\|$ satisfies the triangle inequality,

$$\|\sum_{i} c_i P_i(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\| \leq \sum_{i} \|c_i P_i(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\|.$$

Since $\|\cdot\|$ is homogeneous,

$$\sum_{i} \|c_i P_i(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\| = \sum_{i} c_i \|P_i(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\|.$$

Since $\|\cdot\|$ is symmetric, it is invariant under permutations, so

$$\sum_{i} c_i \|P_i(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\| = \sum_{i} c_i \|(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\|.$$

Since the c_i sum to 1,

$$\sum_i c_i \|(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\| = \|(\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\|.$$

Finally, since $\|\cdot\|$ satisfies the triangle inequality,

$$\|\boldsymbol{\lambda}(A) + \boldsymbol{\lambda}(B))\| \leq \|\boldsymbol{\lambda}(A)\| + \|\boldsymbol{\lambda}(B))\| = \|A\|_{\mathrm{H}} + \|B\|_{\mathrm{H}}$$

Thus, $\|\cdot\|_{\mathrm{H}}$ satisfies the triangle inequality. □

Chapter 3

Probability Theory

The matrix norms in [5], which we introduce in Chapter 5 are constructed using properties of random vectors. In Chapter 6, we will introduce the Pareto distribution and examine the matrix norms it can induce. This chapter aims to provide enough background in probability theory to define our norms and to understand the Pareto distribution.

3.1 Random Variables

Definition 48. A real *random variable* X is a variable that takes on a range of values on the real line according to a probability distribution. For a more rigorous definition of a random variable using measure theory, see Section 4.3.

Some well-known probability distributions are the Bernoulli distribution, the normal distribution, and the exponential distribution.

Definition 49. A *discrete* real random variable is one in which the number of values that X can take on is finite or at most countably infinite.

Example 50. A Bernoulli random variable X is one that takes on the value 1 with probability p and 0 with probability $1-p$. Since X can take on only two possible values, X is discrete.

Definition 51. A *continuous* real random variable X is one which can take on any value in one or more intervals of the real line, but for which the probability that X has any specific value in $\mathbb{R}$ is zero.

Definition 52. The *probability distribution function* or *PDF* of a continuous real random variable X represents X's density of likelihood for every point in $\mathbb{R}$. The probability that X falls within an interval $[a, b]$ is the integral from a to b of the PDF. No point can have negative probability, and since total probability must add up to 1,

$$\int_{-\infty}^{\infty} \text{PDF}(x)\, dx = 1.$$

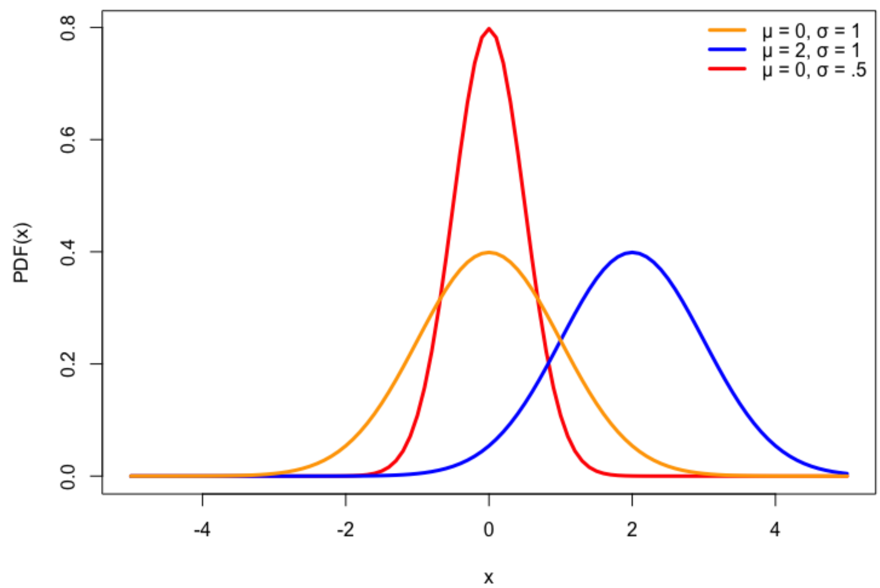


Figure 3.1: Normal PDFs

Example 53. The normal distribution is known as the 'bell curve' because its Probability Distribution Function resembles a bell. In a normal curve with mean μ and standard deviation σ, the PDF is

$$\text{PDF}(x) = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

The PDFs of normal distributions with varying parameters μ and σ are shown in Figure 3.1.

Definition 54. The *support* of a continuous random variable is the closure of the set of points $x \in \mathbb{R}$ for which $\text{PDF}_X(x) > 0$. A random variable X cannot take on any value that is not in its support.

Example 55. The exponential distribution is used to estimate the time between spontaneous events whose occurrence is independent of other such events. Its positive real parameter λ determines the steepness of the curve; that is, as λ grows, so grows the probability that X is near zero. If X is an exponential random variable with parameter λ, its PDF is

$$\text{PDF}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

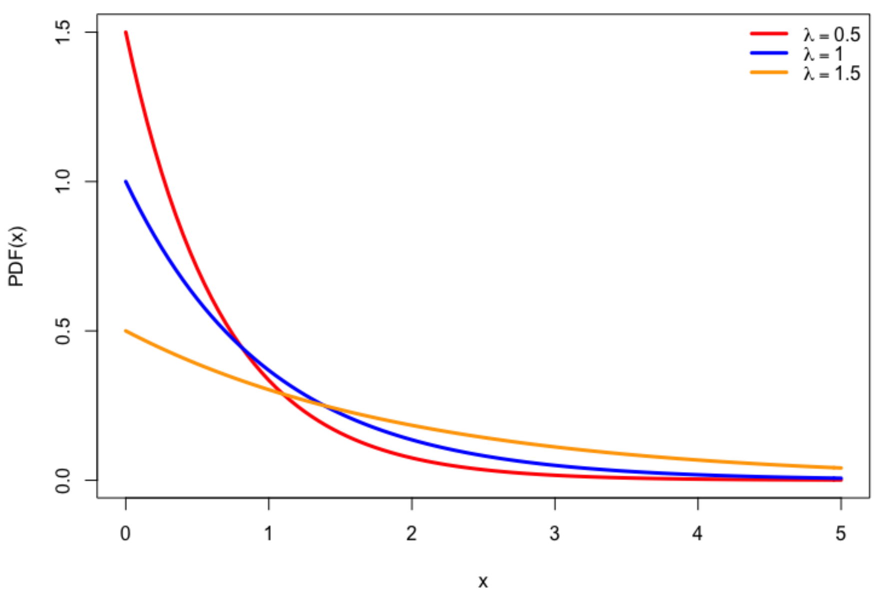


Figure 3.2: Exponential PDFs

Since $\lambda e^{-\lambda x}$ is always greater than zero for positive λ, if X is exponential, the support of X in $\mathbb{R}$ is $[0, \infty)$.

Definition 56. The *cumulative distribution function* or CDF of a continuous random variable gives for each x in $\mathbb{R}$ the probability that X takes on a value less than or equal to x. If X is discrete, the CDF of x is the sum of the probabilities of all values less than or equal to x. If X is continuous, it is the integral from negative infinity to x of the PDF of X. Since the probabilities and probability densities are always positive, and the integral over all $\mathbb{R}$ is 1,

$$\lim_{x \to -\infty} \mathrm{CDF}(x) = 0,$$

and

$$\lim_{x \to \infty} \mathrm{CDF}(x) = 1.$$

Example 57. The CDFs of normal and exponential distributions are shown in Figure 3.3.

3.1.1 Moments

Definition 58. The *expected value*, or *mean*, of a random variable X, denoted $\mathbf{E}[X]$, is the average of the values of X weighted by probability or probability density.

For a discrete random variable, let $\mathrm{p}(x)$ be the probability of x in X. The expected value of X is calculated by taking the sum over all the values x that X can take on of $x\mathrm{p}(x)$.

$$\mathbf{E}[X] = \sum_{x \colon \mathrm{p}(x) > 0} x\mathrm{p}(x).$$

For continuous random variables, the integral takes the place of the sum, and the probability density takes the place of discrete probabilities.

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x\mathrm{PDF}(x)\, dx.$$

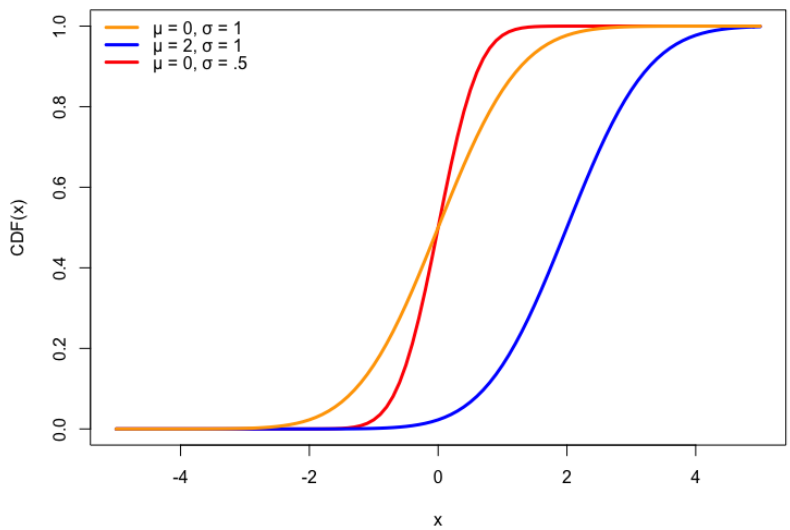
Remark 59. The existence of an expected value is not guaranteed. For example, in Chapter 6, we examine the Pareto Distribution, which does not have a mean when its parameter α is less than 1.

Definition 60. The nth *moment* of a random variable X, denoted $\mu_n(X)$, is the expected value of X^n.

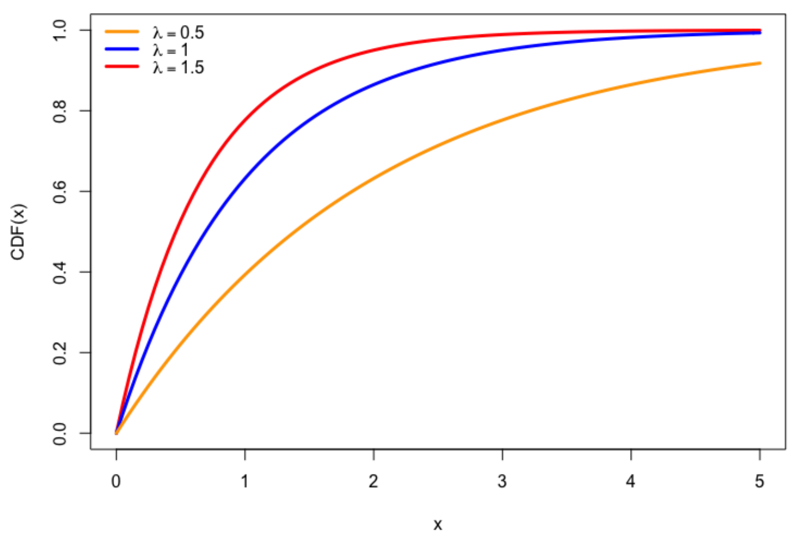
Definition 61. The *variance* σ^2 of a random variable X is $\mu_2 - \mu_1^2$, where μ_i is the ith moment of X:

$$\sigma^2 = \mu_2 - \mu_1^2 = \mathbf{E}[X^2] - 2\mathbf{E}[X\mu_1] + \mu_1)^2 = \mathbf{E}\left[(X - \mu_1)^2\right].$$

A random variable X is *constant* almost everywhere if and only if its variance $\sigma^2 = \mathbf{E}\left[(X - \mu_1)^2\right]$ is zero.



(a) Normal Distribution CDFs



(b) Exponential Distribution CDFs

Figure 3.3

Example 62. For the Bernoulli distribution with probability $\mathrm{p}(1) = p$, the nth moments are

$$\mu_n = \mathbf{E}\left[X^n\right] = \sum_{i=0}^{1} i^n \mathrm{p}(i) = 0^n\left(1-p\right) + 1^n p = p.$$

Example 63. If X has exponential distribution with parameter λ, the moments are

$$\mu_n = \mathbf{E}\left[X^n\right] = \int_0^\infty x^n \lambda e^{-\lambda x}\, dx. \tag{3.1}$$

We can integrate by parts until we get rid of the exponent. Using $u(x) = e^{-\lambda x}$ and $v'(x) = x^n$,

$$\lambda \int_0^\infty x^n e^{-\lambda x}\, dx = \lambda \frac{x^n e^{-\lambda x}}{-\lambda}\Big|_0^\infty - \lambda \int_0^\infty \frac{n x^{n-1} e^{-\lambda x}}{-\lambda}\, dx.$$

Since we are evaluating from 0 to infinity, every term uv on the left is zero, since for each natural number $k > 0$, $0^k e^{-\lambda 0} = 0$ and $\lim_{x\to\infty} x^k e^{-\lambda x} = 0$. The negative from integration by parts and from the $-\lambda$ in the denominator cancel, so this becomes

$$\lambda \int_0^\infty \frac{n! x^0 e^{-\lambda x}}{\lambda^n}\, dx = \frac{\lambda n! e^{-\lambda x}}{-\lambda^{n+1}}\Big|_0^\infty = 0 - \left(\frac{-n!}{\lambda^n}\right) = \frac{n!}{\lambda^n} = \mu_n.$$

If moments exist for all n, it is sometimes useful to construct functions whose Taylor series coefficients are related to the moments of a distribution.

3.1.2 Moment Generating Functions

Definition 64. Let X be a real random variable with that whose moments μ_n exist for all $n \in \mathbb{N}$. The *moment generating function*, or *MGF*, of X is the function

$$\mathrm{MGF}(t) = 1 + \mu_1 t + \frac{\mu_2}{2!} t^2 + \frac{\mu_3}{3!} t^3 + \cdots.$$

By the definition of μ_n, this is

$$1 + \mathbf{E}\left[X\right] t + \frac{\mathbf{E}\left[X^2\right]}{2!} t^2 + \frac{\mathbf{E}\left[X^3\right]}{3!} t^3 + \cdots = \mathbf{E}\left[1 + Xt + \frac{X^2}{2!} t^2 + \frac{X^3}{3!} t^3 + \cdots\right],$$

which is the Taylor expansion of

$$\mathbf{E}\left[e^{tX}\right]$$

in a neighborhood around $t = 0$.

Example 65. The use of the Taylor series identity makes finding the moments of our same exponential variable X from 63 much easier than the somewhat tedious process of integration by parts of 3.1. Let X be exponential with parameter λ. The MGF of X is

$$\mathbf{E}\left[e^{tX}\right] = \int_0^\infty e^{tx}\left(\lambda e^{-\lambda x}\right)\, dx = \frac{\lambda e^{(t-\lambda)x}}{t-\lambda}\bigg|_0^\infty = \frac{\lambda}{t-\lambda} = \frac{1}{1-\frac{t}{\lambda}}$$

The Taylor series of $\frac{1}{1-\frac{t}{\lambda}}$ is

$$\frac{1}{1-\frac{t}{\lambda}} = 1 + \frac{t}{\lambda} + \frac{t^2}{\lambda^2} + \frac{t^3}{\lambda^3} + \cdots = 1 + \mu_1 t + \frac{\mu_2}{2!}t^2 + \frac{\mu_3}{3!}t^3 + \ldots.$$

So $\mu_n = \frac{n!}{\lambda^n}$.

Proposition 66. *Let independent random variables X and Y have moment generating function M_X and M_Y respectively. Then the moment generating function of $Z = X + Y$ is the product $\mathrm{M}_X\mathrm{M}_Y$.*

Proof. Let $Z = X + Y$. Then

$$\mathrm{M}_\mathrm{Z}(t) = \mathbf{E}\left[e^{tZ}\right] = \mathbf{E}\left[e^{t(X+Y)}\right] = \mathbf{E}\left[e^{tX}e^{tY}\right].$$

Since X and Y are independent, so are e^{tX} and e^{tY}, so the expected value of their product is the product of their expected values. Hence

$$\mathrm{M}_\mathrm{Z}(t) = \mathbf{E}\left[e^{tX}e^{tY}\right] = \mathbf{E}\left[e^{tX}\right]\mathbf{E}\left[e^{tY}\right] = \mathrm{M}_X(t)\mathrm{M}_Y(t).$$

Thus, the MGF of a sum of independent random variables is the product of their MGFs. □

3.2 Random Vectors

Definition 67. A real *random vector* $\mathbf{X}$ is a column vector $[X_1, X_2, \ldots, X_n]^\mathsf{T}$ where each X_i is a real random variable.

Definition 68. When all X_i in a random vector $\mathbf{X}$ have the same distribution and are independent of each other, the random variables are called *i.i.d.* (independent and identically distributed).

3.2.1 Second Moment Matrix

Definition 69. For a random vector $\mathbf{X}$ of dimension n, the *second moment matrix* $\Sigma(\mathbf{X})$ is the $n \times n$ matrix of the expected values of X_iX_j.

$$\Sigma(\mathbf{X}) = \mathbf{E}[\mathbf{X}\mathbf{X}^\mathsf{T}] = \begin{bmatrix} \mathbf{E}[X_1^2] & \mathbf{E}[X_1X_2] & \cdots & \mathbf{E}[X_1X_n] \\ \mathbf{E}[X_2X_1] & \mathbf{E}[X_2^2] & \cdots & \mathbf{E}[X_2X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}[X_nX_1] & \mathbf{E}[X_nX_2] & \cdots & \mathbf{E}[X_n^2] \end{bmatrix}.$$

Definition 70. We call a random vector *positive definite* when its second moment matrix is positive definite.

Theorem 71. *Let* $\mathbf{X}$ *be a random vector of dimension* n *composed of* n *i.i.d. random variables* X_i *with at least two moments* μ_1 *and* μ_2*. If the* X_i *are non-constant, the second moment matrix* $\Sigma(\mathbf{X})$ *is positive definite.*

Proof. Since the X_i are independent, the non-diagonal entries $\Sigma(\mathbf{X})_{ij}$ where $i \neq j$ are

$$\Sigma(\mathbf{X})_{ij} = \mathbf{E}[X_i X_j] = \mathbf{E}[X_i]\mathbf{E}[X_j] = \mu_1^2.$$

The diagonal entries are

$$\Sigma(\mathbf{X})_{ii} = \mathbf{E}[X_i^2] = \mu_2.$$

So the second moment matrix $\Sigma(\mathbf{X})$ equals

$$\begin{bmatrix} \mu_2 & \mu_1^2 & \cdots & \mu_1^2 \\ \mu_1^2 & \mu_2 & \cdots & \mu_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1^2 & \mu_1^2 & \cdots & \mu_2 \end{bmatrix} = \mu_1^2 \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} + (\mu_2 - \mu_1^2) \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

This is $\mu_1^2 J + \sigma^2 I$, where J is the all-ones matrix, σ^2 is the variance (see 61), and I is the identity matrix $\mathrm{diag}(1, 1, \ldots, 1)$. Since all its columns are the same, their span has dimension of 1, and the null space has dimension $n-1$. So J has just one nonzero eigenvalue.

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} n \\ n \\ \vdots \\ n \end{bmatrix}.$$

Since J is Hermitian and its only non-zero eigenvalue, n, is positive, the all-ones matrix is positive semi-definite. The identity matrix is positive definite. If the X_i are non-constant, the variance $\sigma^2 > 0$.

Since the sum of a positive definite and a positive semi-definite matrix is positive definite (See Proposition 26), the second moment matrix $\Sigma(X)$ is positive definite. □

Chapter 4

Measure Theory

The treatment in Section 3.1 of random variables and their moments is intuitive and adequate for understanding how our norms are constructed and what they look like. However, for a more rigorous definition of a random variable, and in order to prove the triangle inequality for our norms (see Subsection 4.3.2), we need measure theory.

4.1 Measure Spaces

Definition 72. Let Ω be a non-empty set, and let $\mathcal{P}(\Omega)$ denote the power set of Ω, the set of all subsets of Ω. The set $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is a *σ-algebra* on Ω if it is non-empty and closed under complement, countable intersection, and countable intersection. That is, if for all sets $A, B \in \mathcal{F}$,

1. $A^{\mathsf{c}} \in \mathcal{F}$,
2. $A \cup B \in \mathcal{F}$, and
3. $A \cap B \in \mathcal{F}$.

Definition 73. Let Ω be a non-empty set, and let $\mathcal{F}$ be a σ-algebra on Ω. The double $(\Omega, \mathcal{F})$ is called a *measurable space*.

Definition 74. Let Ω be a set, and let $\mathcal{F}$ be a σ-algebra on S. A *measure* μ is a set function from $\mathcal{F}$ to $[0, \infty]$ that is positive semi-definite, takes $\emptyset$ to 0, and is finitely additive:

1. $\mu(A) \geq 0$ for all $A \in \mathcal{F}$,
2. $\mu(\emptyset) = 0$, and
3. $\mu(A) + \mu(B) = \mu(A \cup B)$ if A and B are disjoint sets in $\mathcal{F}$.

Definition 75. A *measure space* $(\Omega, \mathcal{F}, \mu)$ is a triplet composed of a set S, a σ-algebra $\mathcal{F}$ on Ω, and a measure μ on $\mathcal{F}$.

Remark 76. Unfortunately, it is not possible to define a measure on the power set of $\mathbb{R}$ with all three characteristics. Instead, by using a smaller σ-algebra called the Borel σ-algebra, we will be able to define probability measures.

Definition 77. The *Borel σ-algebra* of $\mathbb{R}$, denoted $\mathcal{B}$, is the σ-algebra generated by the open sets of $\mathbb{R}$. That is, a set $B \subseteq \mathbb{R}$ is an element of $\mathcal{B}$ if it is the result of a countably many intersection, union, and complement operations on open sets of $\mathbb{R}$.

Definition 78. Let $(\Omega_1, \mathcal{F}_1, \mu)$ be a measure space and let $(\Omega_2, \mathcal{F}_2)$ be a measurable space. A function $f : \Omega_1 \to \Omega_2$ is a *measurable function* if for every subset $A_2 \in \mathcal{F}_2$, the preimage of A_2 under f, denoted $f^{-1}(A_2)$, is a subset in $\mathcal{F}_1$ and is thus measurable by μ.

Definition 79. Let $(\Omega_1, \mathcal{F}_1, \mu)$ be a measure space and let $(\Omega_2, \mathcal{F}_2)$ be a measurable space. The *pushforward measure* μ_f of a measurable function $f : \Omega_1 \to \Omega_2$ is the measure which satisfies $\mu_f(A_2) = \mu(f^{-1}(A_2))$ for all $A_2 \in \mathcal{F}_2$.

4.2 Lebesgue Integration

Definition 80. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and, considering the measurable space $(\mathbb{R}, \mathcal{B})$, define a measurable function $f : \Omega \to \mathbb{R}$.

Define the function $f^* : [1, \infty) \to \mathbb{R}$ such that $f^*(t)$ is the μ-measure of subset $E_t \in \mathcal{F}$ such that $\omega \in E_t$ whenever $f(\omega) > t$.

$$f^*(t) = \mu\left(\{\omega \in \Omega \mid f(\omega) > t\}\right).$$

In other words, $f^*(t)$ returns the measure of the preimage $f^{-1}((t, \infty))$. Since any set (t, ∞) is an element of $\mathcal{B}$, and f is a measurable function, every $E_t = f^{-1}(t, \infty)$ is a μ measurable set, so f^* is well defined.

Definition 81. The *Lebesgue integral* of f over Ω, denoted

$$\int_E f \, d\mu,$$

is the Riemann integral

$$\int_0^\infty f^*(t) \, dt.$$

Remark 82. Our choice of $*$ in f^* is intended to inspire notions of the transpose. Instead of calculating the area under f by using f-tall and progressively narrower rectangles, as in normal Reimann integration, we use f^*-wide and progressively shorter rectangles stacked on top of each other.

Intuitively, Lebesgue integrals find the area under a curve by summing up dt-thick layers, one layer at a time, much like a 3D printer. If Riemann integration is like building a picket fence whose parapets are the shape of f, with many progressively thinner vertical slats (of width dx) placed next to each other, Lebesgue integration is like building a masonry wall of the same size and with the same f-shaped parapets with progressively shorter bricks (of height dt).

The advantage of Lebesgue integration is that if the x-axis has an irregular measure, that can be accounted for every time you add a layer. For instance, many measurable functions from spaces with probability measures are not amenable to Riemann integration for this reason.

Proposition 83. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, equip $\mathbb{R}$ with the Borel σ-algebra in the measurable space $(\mathbb{R}, \mathcal{B})$, and let f be a measurable function that induces the pushforward measure μ_f on $(\mathbb{R}, \mathcal{B})$, and let g be a measurable function from $\mathbb{R}$ to $\mathbb{R}$. If they exist, the Lebesgue integrals*

$$\int_\Omega g \circ f \, d\mu \text{ and } \int_\mathbb{R} g \, d(\mu_f)$$

are equal.

Proof. Integrating the Lebesgue integral on the left, we would construct a function $(g \circ f)^*$ such that $(g \circ f)^*(t) = \{\omega \in \Omega \,|\, g(f(\omega)) > t\}$ and integrate it from $t = 0$ to infinity.

Evaluating the integral on the right, we construct g^* such that $g^*(t)$ is the μ_f measure of the set $B_t \in \mathcal{B}$ that contains those points x in $\mathbb{R}$ such that $g(x) > t$. Formally,

$$g^*(t) = \mu_f(B_t) = \mu_f(\{x \in \mathbb{R} \,|\, g(x) > t\}).$$

The preimage of B_t under f is the set

$$f^{-1}(B_t) = \{\omega \in \Omega \,|\, f(\omega) \in B_t\}.$$

The real number $f(\omega)$ is in B_t if and only if it satisfy the condition $g(f(\omega)) > t$. Thus,

$$f^{-1}(B_t) = \{\omega \in \Omega \,|\, g(f(\omega)) > t\}.$$

Hence the g^* we construct when integrating the integral on the left equals the $(g \circ f)^*$ we construct when integrating the integral on the right, and

$$\int_\Omega g \circ f \, d\mu = \int_0^\infty (g \circ f)^*(t) \, dt = \int_0^\infty g^*(t) \, dt = \int_\mathbb{R} g \, d\mu_f.$$

So the two integrals are equal. □

Remark 84. The use of pushforward measures allows us to construct new random variables as measurable functions of existing random variables, knowing that the integrals of functions on these new random variables using their pushforward measures will be just as well-behaved as the originals.

4.3 Probability Spaces

Definition 85. A *probability space* is a measure space $(\Omega, \mathcal{F}, P)$ such that $P(\Omega) = 1$. The measure P on a probability space is called a *probability measure.*

Definition 86. Let $(\Omega, \mathcal{F}, P)$ be a probability space (i.e., $P(\Omega) = 1$), and consider the measurable space $(\mathbb{R}, \mathcal{B})$. A real random variable X is a measurable function from Ω to $\mathbb{R}$. It induces the pushforward measure P_X, which by Definition 79 satisfies $P_X(\mathbb{R}) = P(\Omega) = 1$.

Lemma 87. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then we can take the integral of $f^*(t)\,dt$ from 0 to 1, rather than from 0 to infinity, since for any event in $\mathcal{F}$, $0 \le P(F) \le 1$.*

Theorem 88. *For a probability space $(\Omega, \mathcal{F}, P)$, Lebesgue integration satisfies*

$$\int_\Omega \mathbf{1}\, dP = \int_0^1 \mathbf{1}^*(t)\, dt = \mu(\Omega) = 1.$$

Let $f : \Omega \to \mathbb{R}$ such that $f(x) = 1$ for all $x \in \Omega$. Then we construct $f^* : \mathbb{R} \to \mathbb{R}$ such that $f^*(t)$ is the probability that $f(x) > t$, or in measure theoretic terms, the P-measure of the set such that $f(x) = 1 > t$. This holds for all $x \in \Omega$ when $t < 1$. So

$$f^*(t) = \begin{cases} P(\Omega) & \text{if } 0 \le t < 1, \\ P(\emptyset) & \text{if } 1 \le t. \end{cases}$$

The area under f^* is a unit square.

We can also calculate the Lebesgue integral on other important functions.

Example 89. Let f be the *indicator function* for some measurable set $B \in \mathcal{B}$. That is, let

$$f(x) = \begin{cases} 1 & \text{if } x \in B, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Then $f^*(t)$ is the measure of the set such that $f(x) > t$. This holds for all $x \in B$ when $t < 1$, but holds for no x when $t \ge 1$. So

$$f^*(t) = \begin{cases} P(B) & \text{if } 0 \le t < 1 \\ P(\emptyset) & \text{if } 1 \le t \end{cases}$$

When we take the Lebesgue integral

$$\int_\Omega f\, dP = \int_0^1 f^*(t)\, dt,$$

the area under f^* is a rectangle of width 1 and height $P(B)$.

We use Lebesgue integrals to calculate the moments of a random variable.

Definition 90. Let $(\Omega, \mathcal{F}, P)$ be a probability space $X : \Omega \to \mathbb{R}$ be a real-valued measurable function that induces the pushforward measure P_X on the measure space $(\mathbb{R}, \mathcal{B}, P_X)$. Then we can define the pth moment of X as

$$\int_{\mathbb{R}} X^p \, dP_X.$$

Example 91. If X is a continuous random variable, then the Lebesgue integral $\int_{\mathbb{R}} X^n \, dP_X$ can be computed using Riemann integration. Since the measurable function X is continuous, we can write $dP_X = PDF_X(x) \, dx$ so the nth moment of X satisfies

$$\int_{\Omega} X^n \, dP = \int_{\mathbb{R}} X^n \, dP_X = \int_{-\infty}^{\infty} x^n PDF(x) \, dx$$

4.3.1 L^p Spaces

We can use Lebesgue integration to define L^p spaces, which are spaces of functions f such that the Lebesgue integral over Ω of $|f|^p$ is less than infinity.

Definition 92. Recalling Chapter 4, let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let f be a measurable function from S to $\mathbb{R}$. The L^p *norm* is the function $\| \cdot \|_p$ defined by

$$\|f\|_p = \left(\int_X |f|^p \, d\mu \right)^{1/p}.$$

Remark 93. Note that the L^p norm of a random variable Y is pth root of the pth moment of Y.

Definition 94. Let $L^p(\Omega, \mu)$ be the set of measurable functions f from Ω to $\mathbb{R}$ equipped with the Borel σ-algebra such that

$$\|f\|_p = \left(\int_X |f|^p \, d\mu \right)^{1/p} \leq \infty.$$

Such sets L^p are known as L^p spaces, and they are important for functional analysis. The property of greatest interest to us is that when $p \geq 1$, the L^p norms used to define them satisfy the triangle inequality.

4.3.2 Hölder's Inequality and Minkowski's Inequality

Proving Hölder's and Minkowski's Inequalities from scratch requires proving a handful of non-trivial integral inequalities, all named after early-20th-century mathematicians. One would start with Jensen's inequality, then use that to prove Young's inequality, which one can use to prove Hölder's inequality, after which Minkowski's Inequality is relatively tame.

To prove that our norms satisfy positive definiteness, we will use Hölder's inequality, and in proving that our norms satisfy the triangle inequality, we will merely seek to satisfy the requirements of the Minkowski's Inequality as stated.

Proposition 95. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let f and g be real-valued measurable functions on Ω. Then*

$$\|fg\|_1 \leq \|f\|_q \|g\|_r.$$

for all $q, r \in [0, \infty]$ such that

$$\frac{1}{q} + \frac{1}{r} = 1.$$

Corollary 96. *For $p \geq 2$, let $q = p/2$ and let $r = \frac{p}{p-2}$. Let $\mu = P$ be a probability measure, and let $f = |Y^2|$ for some random variable Y and $g = 1$.*

$$\mathbf{E}\left[|Y^2|\right] = \int_\Omega |Y^2| 1 \, dP \leq \left(\int_\Omega |Y^2|^{p/2} \, dP \right)^{2/p} \left(\int_\Omega 1^r \, dP \right)^{1/r}$$

$$\left(\int_\Omega |Y^2|^{p/2} \, dP \right)^{2/p} \left(\int_\Omega 1^r \, dP \right)^{1/r} = \mathbf{E}\left[Y^p\right]^{2/p}.$$

Minkowski's Inequality is the triangle inequality for L^p norms.

Proposition 97. *If $(\Omega, \mathcal{F}, \mu)$ is a measure space, $\|\cdot\|_p$ is the L^p norm for some $1 < p < \infty$, and $f, g \in S$,*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

We now proceed to define our norms without using the formalisms of measure theory, but when it comes to proving the triangle inequality, we will again resort to the rigorous measure-theoretic definitions of this section.

Chapter 5

Random Vector Norms

In this chapter, we use all the linear algebra and probability theory from Chapters 1 and 3 to construct our norms using random vectors and prove that they are indeed norms. We then use a special case of our norms to prove Hunter's Theorem, a result about complete homogeneous symmetric polynomials. We provide plenty of examples and plots of our norms in $\mathbb{R}^2$ and $\mathbb{R}^3$. I then prove another result that goes beyond [5], namely, that the norms are continuous with respect to the exponent p.

5.1 Random Vector Norms

Definition 98. Let $\langle \cdot, \cdot \rangle$ be the inner product (or dot product) of two $n \times 1$ vectors familiar from basic linear algebra such that for $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^* \mathbf{x} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

Note that for a random vector $\mathbf{X}$, the dot product $\langle \mathbf{X}, \boldsymbol{\lambda} \rangle$ is the sum

$$\langle \mathbf{X}, \boldsymbol{\lambda} \rangle = X_1 \lambda_1 + X_2 \lambda_2 + \cdots + X_n \lambda_n,$$

where each λ_i is a real number and each X_i is a real random variable. This sum $\langle \mathbf{X}, \boldsymbol{\lambda} \rangle$ is another real random variable, and therefore, it has an expected value. Since the product of random variables is also a random variable, $|\langle \mathbf{X}, \boldsymbol{\lambda} \rangle|^p$ is also a random variable, and its absolute value means it is defined on $[0, \infty)$.

Definition 99. The *Gamma Function* Γ is a complex extension of the factorial function ! on the nonnegative integers. It has many interesting properties, but in this thesis we are concerned only with its restriction to the interval $[1, \infty)$, on which it is smooth, positive, and exactly what out might expect out of a function satisfying the equation

$$\Gamma(n) = (n-1)!$$

for $n = 1, 2, 3, \cdots$. On half of the complex plane with positive real component, Γ is defined by

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt.$$

5.2 Main Theorem

Theorem 100. *Define the function $\|\cdot\|_{\mathbf{X},p} : \mathbb{R}_n \to \mathbb{R}$ for an i.i.d. random vector $\mathbf{X}$ with X_i and $p \geq 1$ by*

$$\|\boldsymbol{\lambda}\|_{\mathbf{X},p} = \left(\frac{\mathbf{E}|\langle \mathbf{X}, \boldsymbol{\lambda} \rangle|^p}{\Gamma(p+1)} \right)^{1/p}.$$

Then $\|\cdot\|_{\mathbf{X},p}$ is a norm.

Remark 101. In this thesis, we are primarily concerned with cases in which p is an even integer, since norms defined using even p can be defined using moment generating functions. Accordingly, we will replace $\Gamma(p+1)$ with $p!$ unless we need to address cases involving non-integer p (see Section 5.5.1):

$$\|\boldsymbol{\lambda}\|_{\mathbf{X},p} = \left(\frac{\mathbf{E}|\langle \mathbf{X}, \boldsymbol{\lambda} \rangle|^p}{p!} \right)^{1/p}.$$

Corollary 102. *Define the function ${}_{\mathrm{H}}\|\cdot\|_{\mathbf{X},p} : \mathrm{H}_n \to \mathbb{R}$ for a positive definite random vector $\mathbf{X}$ with X_i i.i.d., and $p \geq 1$ by*

$${}_{\mathrm{H}}\|A\|_{\mathbf{X},p} = \left(\frac{\mathbf{E}|\langle \mathbf{X}, \boldsymbol{\lambda}(A) \rangle|^p}{\Gamma(p+1)} \right)^{1/p}$$

Since $\mathbf{X}$ is i.i.d., $\|\cdot\|_{\mathbf{X},p}$ is symmetric, so by Theorem 36, ${}_{\mathrm{H}}\|\cdot\|_{\mathbf{X},p}$ is a norm.

Remark 103. This remark concerns errors in our paper [5]. In Remark 3.4 of [5], we claimed that our proof did not require the X_i to be independent, merely that they be identically distributed (which we mistakenly wrote as "i.i.d"). Even ignoring the typo, the claim is erroneous in its strong form. Suppose X_1 is a random variable, and $X_2 = X_1$ has the identical distribution, but is dependent. Let $\lambda_1 = -\lambda_2$. Then $\mathbf{E}|\lambda_1 X_1 + \lambda_2 X_2| = \mathbf{E}|\lambda_1 X_1 - \lambda_1 X_1| = 0$. This violates the property of positive definiteness.

There may be some room for the X_i to be neither totally independent nor i.i.d., however. The function $\cdot^{\downarrow} : \mathbb{R}^n \to \mathbb{R}^n$, which sorts any vector into a non-increasing vector with the same entries, such that $\boldsymbol{\lambda}^{\downarrow} = (\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_n})$ with $\lambda_{i_1} \geq \lambda_{i_2} \geq \cdots \geq \lambda_{i_n}$, is symmetric. If we define our norm as

$$\|\cdot\|_{\mathbf{X},p} = \left(\frac{\mathbf{E}|\langle \mathbf{X}, \boldsymbol{\lambda}^{\downarrow} \rangle|^p}{\Gamma(p+1)} \right)^{1/p},$$

then the X_i can be any random vector such that the second moment matrix is positive definite and the triangle inequality holds, because the composition with $\cdot^{\downarrow}$ guarantees that the norm is symmetric. A characterization of the random vectors $\mathbf{X}$ that satisfy the triangle inequality remains an open question.

In [5]'s proof of Theorem 1b, we also misleadingly claimed that for diagonal $A, B \in \mathrm{H}_n$, $\boldsymbol{\lambda}(A+B) = \boldsymbol{\lambda}(A)+\boldsymbol{\lambda}(B)$. In fact, this is true only if the diagonals of A and B have the same ordering. One can work around this by restricting to the set of diagonal matrices with non-increasing diagonal entries, or by assuming that the norm is symmetric and using Theorem 47.

Proving Theorem 100

To prove that $\|\cdot\|_{\mathbf{X},p}$ is a norm, we must show three things:

1. $\|\boldsymbol{\lambda}\|_{\mathbf{X},p}$ equals zero when $\boldsymbol{\lambda} = 0$ and is otherwise positive.
2. $\|\cdot\|_{\mathbf{X},p}$ is homogeneous, so $\|c\boldsymbol{\lambda}\|_{\mathbf{X},p} = c\|\boldsymbol{\lambda}\|_{\mathbf{X},p}$ for any positive scalar c.
3. $\|\cdot\|_{\mathbf{X},p}$ satisfies the triangle inequality.

5.2.1 Positive Definiteness

Lemma 104. *Let $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ be positive definite. The function $\|\cdot\|_{\mathbf{X},p}$ is positive definite.*

Proof. We rely on the positive definiteness of $\mathbf{X}$ as we have defined positive definiteness for random vectors in 1.3. All we need to do is fit the expression into a form that includes the positive definite matrix $\mathbf{X}\mathbf{X}^{\mathsf{T}}$. We manage this by using Hölder's inequality as shown in 96. Let $Y = \langle \mathbf{X}, \boldsymbol{\lambda}\rangle$. Then

$$\mathbf{E}\left[|Y^p|\right]^{2/p} \geq \mathbf{E}\left[|Y^2|\right].$$

We first examine the left side.

$$\mathbf{E}\left[|Y^p|\right]^{2/p} = \left(\mathbf{E}\left|\langle \mathbf{X}, \boldsymbol{\lambda}\rangle^p\right|^{1/p}\right)^2 = \Gamma(p+1)^{2/p}\|\boldsymbol{\lambda}\|^2_{\mathbf{X},p}.$$

Since square roots and multiplication by positive scalars preserve sign,

$$\|\boldsymbol{\lambda}\|_{\mathbf{X},p} \geq \frac{\left(\mathbf{E}\left[|Y^2|\right]\right)^{1/2}}{\Gamma(p+1)^{1/p}}.$$

We now turn to Y^2. Since the dot product is commutative,

$$Y^2 = \langle \mathbf{X}, \boldsymbol{\lambda}\rangle^2 = \langle \boldsymbol{\lambda}, \mathbf{X}\rangle\langle \mathbf{X}, \boldsymbol{\lambda}\rangle = \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{X}\mathbf{X}^{\mathsf{T}}\boldsymbol{\lambda}.$$

Since $\mathbf{E}\left[\mathbf{X}\mathbf{X}^{\mathsf{T}}\right]$ is positive definite, the expected value

$$\mathbf{E}\left[|Y^2|\right] = \mathbf{E}|\boldsymbol{\lambda}^{\mathsf{T}}\left(\mathbf{X}\mathbf{X}^{\mathsf{T}}\right)\boldsymbol{\lambda}| > 0$$

if and only if $\boldsymbol{\lambda}$ is the zero vector. Thus, $\|\boldsymbol{\lambda}\|_{\mathbf{X},p} = 0$ if and only if $\boldsymbol{\lambda} = 0$. □

Corollary 105. *If the X_i are nonzero and i.i.d., then by 71, $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ is positive definite, so $\|\cdot\|_{\mathbf{X},p}$ is positive definite.*

5.2.2 Homogeneity

The second property, homogeneity, is relatively straightforward.

Lemma 106. *The function* $\|\cdot\|_{\mathbf{X},p}$ *is homogeneous. In other words, for any scalar* c*,*

$$\|cA\|_{\mathbf{X},p} = |c|\|\boldsymbol{\lambda}\|_{\mathbf{X},p}.$$

Proof. Inner products and expected values preserve constants, and the absolute value brackets preserve magnitude, so

$$\mathbf{E}|\langle \mathbf{X}, c\boldsymbol{\lambda}\rangle| = \mathbf{E}|c\langle \mathbf{X}, \boldsymbol{\lambda}\rangle| = |c|\mathbf{E}|\langle \mathbf{X}, \boldsymbol{\lambda}\rangle|.$$

All that remains for the poor constant is to pass through the exponents p and $1/p$ before arriving safely at the far left of the expression. Since $|c|$ is positive, the fractional exponent poses no challenge. Thus,

$$\|cA\|_{\mathbf{X},p} = |c|\left(\frac{\mathbf{E}|\langle \mathbf{X}, \boldsymbol{\lambda}\rangle|^p}{p!}\right)^{1/p} = |c|\|\boldsymbol{\lambda}\|_{\mathbf{X},p}. \quad \square$$

5.2.3 Triangle Inequality

Our Norms As L^p Norms

Let $\mathbf{X}$ be an i.i.d. real random vector on the probability space $(\Omega, \mathcal{F}, P)$, and let $Y = \langle \mathbf{X}, \boldsymbol{\lambda}\rangle$. Since Y is the sum of constants times random variables X_i, Y is a random variable. Our norm $\|\cdot\|_{\mathbf{X},p}$ is defined by

$$\sqrt[p]{\Gamma(p+1)}\|\boldsymbol{\lambda}\|_{\mathbf{X},p} = \left(\mathbf{E}|\langle \mathbf{X}, \boldsymbol{\lambda}\rangle|^p\right)^{1/p} = \left(\mathbf{E}|Y|^p\right)^{1/p}.$$

Using the Lebesgue definition of a moment,

$$\left(\mathbf{E}|Y|^p\right)^{1/p} = \left(\int_\Omega |Y|^n \, dP\right)^{1/p}.$$

This is the L^p norm of Y, so it satisfies Minkowski's inequality. Replacing Y with $|\langle \mathbf{X}, \boldsymbol{\lambda}\rangle|$, we obtain

$$\left(\int |\langle \mathbf{X}, \boldsymbol{\lambda}\rangle|^p \, dP\right)^{1/p}.$$

Lemma 107. *Let* $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathbb{R}^n$*. Then*

$$\|\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2\|_{\mathbf{X},p} \leq \|\boldsymbol{\lambda}_1\|_{\mathbf{X},p} + \|\boldsymbol{\lambda}_2\|_{\mathbf{X},p}.$$

Proof. We take the norm of $\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2$

$$\sqrt[p]{\Gamma(p+1)}\|\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2\|_{\mathbf{X},p} = \left(\int |\langle \mathbf{X}, \boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2\rangle|^p \, dP\right)^{1/p}.$$

We distribute the inner product,

$$\left(\int |\langle \mathbf{X}, \boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)\rangle|^p \, dP\right)^{1/p} = \left(\int |\langle \mathbf{X}, \boldsymbol{\lambda}_1\rangle + \langle \mathbf{X}, \boldsymbol{\lambda}_2\rangle|^p \, dP\right)^{1/p}.$$

Applying Minkowski's Inequality, the previous expression is

$$\leq \left(\int |\langle \mathbf{X}, \boldsymbol{\lambda}_1\rangle|^p \, dP\right)^{1/p} + \left(\int |\langle \mathbf{X}, \boldsymbol{\lambda}_2\rangle|^p \, dP\right)^{1/p}.$$

Thus,

$$\sqrt[p]{\Gamma(p+1)}\|\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2\|_{\mathbf{X},p} \leq \sqrt[p]{\Gamma(p+1)}\|\boldsymbol{\lambda}_1\|_{\mathbf{X},p} + \sqrt[p]{\Gamma(p+1)}\|\boldsymbol{\lambda}_2\|_{\mathbf{X},p}.$$

Cancelling the constant $\sqrt[p]{\Gamma(p+1)}$, we conclude that

$$\|\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2\|_{\mathbf{X},p} \leq \|\boldsymbol{\lambda}_1\|_{\mathbf{X},p} + \|\boldsymbol{\lambda}_2\|_{\mathbf{X},p}$$

for all $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathbb{R}^n$. □

Proof of Theorem 100

Since the function $\|\cdot\|_{\mathbf{X},p} : \mathbb{R}^n \to \mathbb{R}$ satisfies the conditions of positive definiteness (Lemma 104), homogeneity (Lemma 106), and the triangle inequality (Theorem 107), $\|\cdot\|_{\mathbf{X},p}$ is a norm on $\mathbb{R}^n$. ■

5.3 Moments

Theorem 108. *Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be the random vector in $\mathbb{R}^n$ such that the X_i are i.i.d. random variables with moments μ_i for $i \in \mathbb{N}$. Let* M *be the moment generating function of all the X_i. Then*

$$\|\boldsymbol{\lambda}\|_{\mathbf{X},p}^p = \frac{\mathbf{E}|\langle \mathbf{X}, \boldsymbol{\lambda}\rangle|^p}{n!} = [t^p]\,\mathrm{M}_{\boldsymbol{\lambda}}(t),$$

where

$$\mathrm{M}_{\boldsymbol{\lambda}}(t) = \prod_{i=1}^{n} \mathrm{M}(\lambda_i t),$$

and $[t^p]\,\mathrm{M}_{\boldsymbol{\lambda}}(t)$ refers to the coefficient of the pth term in the Taylor expansion of $\mathrm{M}_{\boldsymbol{\lambda}}$.

Proof. The moment generating function of $X_i\lambda_i$ is

$$\mathrm{M}_{X_i\lambda_i}(t) = \mathbf{E}\left[e^{tX_i\lambda_i}\right] = \mathrm{M}_{X_i}(\lambda_i t)$$

By Proposition 66, the moment generating function of the sum of independent random random variables X_i is the product of each moment generating function. Let $Y = X_1\lambda_1 + X_2\lambda_2 + \cdots + X_n\lambda_n$. Then

$$\mathrm{M}_{\mathrm{Y}}(t) = \prod_{i=1}^{n} \mathrm{M}_{X_i\lambda_i}(t) = \prod_{i=1}^{n} \mathrm{M}_{X_i}(t\lambda_i)$$

□

Theorem 109. *Let* $\mathbf{X}$ *be i.i.d. and let* μ_k *be the* k*th moment of the* X_i*. Then*

$$\|\boldsymbol{\lambda}\|_{\mathbf{X},p} = \sum_{|\pi|=p} \binom{p}{\pi} \prod_{i=1}^{n} \mu_{\pi_i}\lambda_i^{\pi_i}. \tag{5.1}$$

Proof. We first examine the expression whose expected value is the pth moment.

$$\langle \mathbf{X}, \boldsymbol{\lambda}\rangle^p = (X_1\lambda_1 + X_2\lambda_2 + \cdots + X_n\lambda_n)^p$$

We can represent this pth power of a polynomial using multiindices.

$$\sum_{|\pi|=p} \binom{p}{\pi} (X\lambda)^{\pi}$$

where $\pi = \pi_1, \pi_2, \ldots, \pi_n$ with non-negative integer π_i, $\binom{\cdot}{\cdot}$ gives multinomial coefficients, defined as

$$\binom{p}{\pi} = \binom{p}{\pi_1, \pi_2, \cdots, \pi_n} = \frac{p!}{\pi_1!\pi_2!\ldots\pi_n!},$$

and

$$(X\lambda)^{\pi} = (X_1\lambda_1)^{\pi_1}(X_2\lambda_2)^{\pi_2}\cdots(X_n\lambda_n)^{\pi_n}.$$

Because the X_i are independent, the expected value operator can be distributed to each power of X_i.

$$\mathbf{E}\left|\sum_{|\pi|=p}\binom{p}{\pi}(X\lambda)^{\pi}\right| = \sum_{|\pi|=p}\binom{p}{\pi}\mathbf{E}\left|(X_1\lambda_1)^{\pi_1}\right|\mathbf{E}\left|(X_2\lambda_2)^{\pi_2}\right|\ldots\mathbf{E}\left|(X_n\lambda_n)^{\pi_n}\right|$$

Since λ_i are constant and $\mathbf{E}|X_i^k| = \mu_k$, the above expression equals

$$\sum_{|\pi|=p}\binom{p}{\pi}\prod_{i=1}^{n}\mu_{\pi_i}\lambda_i^{\pi_i},$$

which is our desired expression. □

5.4 Hunter's Theorem

Definition 110. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. The *complete homogeneous symmetric* (CHS) polynomial of degree k for n variables is the sum of all possible monomials of degree k with coefficient 1. In other words,

$$h_k(\mathbf{x}) = \sum_{1 \le i_1 \le i_2 \le \ldots i_k \le n} \prod_{j=1}^{k} x_{i_j}.$$

Example 111. Let $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$. The polynomial h_1 in three variables is defined by

$$h_1(\mathbf{x}) = x_1 + x_2 + x_3.$$

The polynomial h_2 in three variables gives

$$h_2(\mathbf{x}) = x_1^2 + x_1x_2 + x_1x_3 + x_2^2 + x_2x_3 + x_3^2.$$

The polynomial h_3 in three variables gives

$$h_3(\mathbf{x}) = x_1^3 + x_1^2x_2 + x_1^2x_3 + x_1x_2^2 + x_1x_2x_3 + x_1x_3^2 + x_2^3 + x_2^2x_3 + x_2x_3^3 + x_3^3.$$

Theorem 112. *Complete homogeneous symmetric polynomials are positive definite.*

Proof. Let $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $A = \operatorname{diag}(\boldsymbol{\lambda})$. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ such that the X_i are i.i.d. exponential random variables with parameter 1 (see Example 55), and let M be the moment generating function for all the X_i. From Example 65, M generates the moments

$$\mu_n = \frac{n!}{1^n} = n!.$$

By Theorem 108, $\|\boldsymbol{\lambda}\|_{\mathbf{X},p}^p$ is the pth term in the MGF $\mathrm{M}_{\boldsymbol{\lambda}}(t) = \prod_{i=1}^n \mathrm{M}(\lambda_i t)$. From 5.1, the pth term in $\mathrm{M}_{\boldsymbol{\lambda}}$ is the sum

$$\sum_{|\pi|=p} \binom{p}{\pi} \prod_{i=1}^{n} \mu_{\pi_i} \lambda_i^{\pi_i} = \sum_{|\pi|=p} \frac{p!}{\pi_i!} \prod_{i=1}^{n} \pi_i! \lambda_i^{\pi_i} = p! \sum_{|\pi|=p} \lambda_i^{\pi_i} = p! h_p(\boldsymbol{\lambda}). \tag{5.2}$$

Our norm $\|\boldsymbol{\lambda}\|_{\mathbf{X},p}$ is the pth root of the pth moment of $|\langle \mathbf{X}, \boldsymbol{\lambda} \rangle|$ divided by $p!$. By 5.2,

$$\|\boldsymbol{\lambda}\|_{\mathbf{X},p} = (h_p(\boldsymbol{\lambda}))^{1/p}.$$

Since $\|\boldsymbol{\lambda}\|_{\mathbf{X},p}$ is a norm, it is positive definite. Thus, $\|\boldsymbol{\lambda}\|_{\mathbf{X},p}^p = h_p(\boldsymbol{\lambda})$ is positive definite. $\square$

5.5 Examples

Since a norm is uniquely defined by its unit ball (Proposition 10), we can get a good picture of our norms by plotting the unit balls in two and three dimensions.

Example 113. For any even p, norms on $\mathbb{R}^n$ induced by normal distributions with mean 0 and variance σ^2 (Figures 5.1) reproduce multiples of the Euclidean norm. When the mean is non-zero, the result is elliptical, but changing p does not affect the shape. This remarkable good behavior can be explained by the fact that the sum of normally distributed random variables centered at zero $Z = \langle \mathbf{X}, \boldsymbol{\lambda} \rangle$ is normally distributed and centered at zero with variance $\sigma_Z^2 = \sum_{i=1}^n \lambda_i \sigma^2$. For even p, the pth moment of Z, which is $p! \|\boldsymbol{\lambda}\|_{\mathbf{X},p}^p$, is $\sigma_Z^p (p-1)!!$, where $k!!$ is the product of all natural numbers less than or equal to k with the same parity as k (See Section 5-4 in [11]). Our norm is thus

$$\left(\frac{p!! \left(\sum_{i=1}^n \lambda_i \sigma^2 \right)^{p/2}}{p!} \right)^{1/p} = \left(\frac{p-1!!}{p!} \right)^{1/p} \left(\sigma^2 \sum_{i=1}^n \lambda_i \right)^{1/2} = \frac{\sigma \|\boldsymbol{\lambda}\|}{(p!!)^{1/p}}.$$

Example 114. The exponential distribution with parameter $\lambda = 1$ (Figure 5.3) produces norms whose unit circles have interesting shapes. The 1-norm on $\mathbb{R}^2$ resembles a racetrack. When λ_1 and λ_2 have the same sign, $\|\boldsymbol{\lambda}\|_{\mathbf{X},1} = |\lambda_1| + |\lambda_2| = \lambda_1 + \lambda_2$, which is linear, but when they have different signs, the unit circle is semicircular. The unit circles of $\|\boldsymbol{\lambda}\|_{\mathbf{X},p}$ for $p = 2, 3, 4$ are football-shaped, and as p increases, they come to resemble the square that is the unit circle of $\|\cdot\|_\infty$ (See Figure 1.2).

5.5.1 Odd and non-integer p

The proof that our norms satisfy the triangle inequality relies on Minkowski's inequality, which does not require p to be an even integer. Because we take the expected value of the absolute value of $\langle \mathbf{X}, \boldsymbol{\lambda} \rangle^p$, our norms are well-defined for all $p \geq 1$, not just even integers. In fact, we can do better.

Theorem 115. *Let $\mathbf{X}$ have at least m moments. The function $f : [1, m] \to \mathbb{R}$ defined by $f(p) = \|\boldsymbol{\lambda}\|_{\mathbf{X},p}$ is continuous for all $\|\boldsymbol{\lambda}\| \in \mathbb{R}^n$.*

Proof. Let P_Y be the pushforward measure of the random variable $Y = \langle \mathbf{X}, \boldsymbol{\lambda} \rangle$. Then

$$\Gamma(p+1)(f(p))^p = \mathbf{E}|\langle \mathbf{X}, \boldsymbol{\lambda} \rangle|^p = \int |\langle \mathbf{X}, \boldsymbol{\lambda} \rangle|^p \, dP = \int |x|^p \, dP_Y.$$

For all $x \in \mathbb{R}$ and $1 \leq p \leq m$, $|x|^p \leq |x| + |x|^m$, so

$$\int |x|^p \, dP_Y \leq \int |x| + |x|^m \, dP_Y = \int |x| \, dP_Y + \int |x|^m \, dP_Y.$$

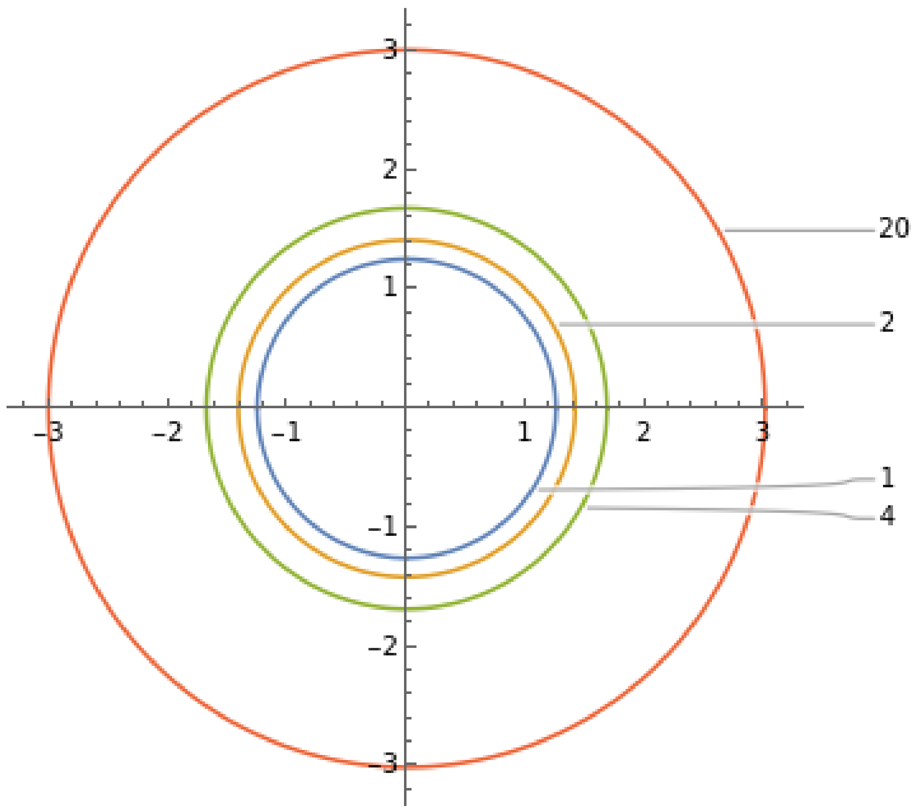


Figure 5.1: Unit circles for $\|\cdot\|_{\mathbf{X},1}$, $\|\cdot\|_{\mathbf{X},2}$, $\|\cdot\|_{\mathbf{X},4}$, and $\|\cdot\|_{\mathbf{X},20}$, where the X_i are normal with $\mu = 0$ and $\sigma = 1$

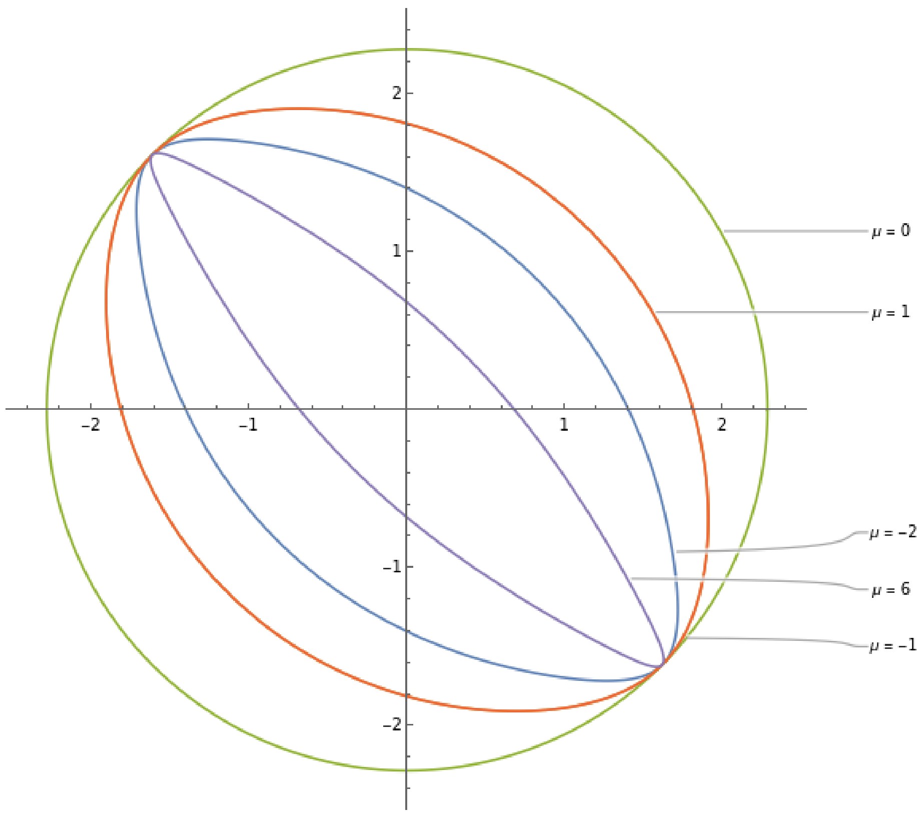


Figure 5.2: Normal unit circles with $p = 10$ and mean $-2, -1, 0, 1$, and 6. The curve is the same for $\mu = -1$ and $\mu = 1$.

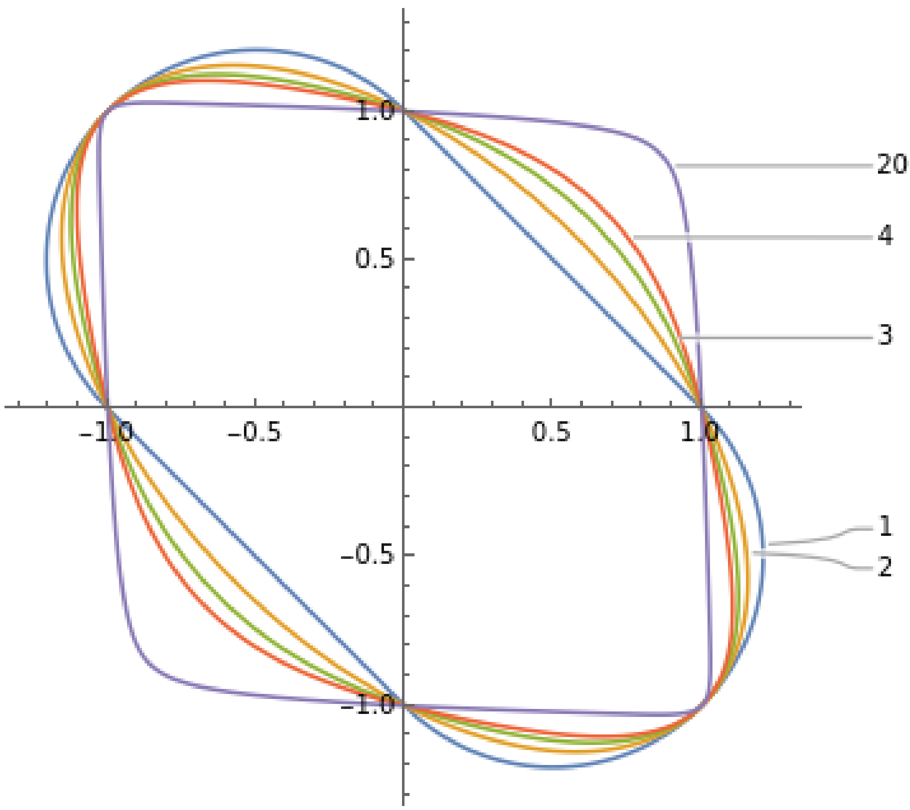


Figure 5.3: Unit circles for $\|\cdot\|_{\mathbf{X},1}$, $\|\cdot\|_{\mathbf{X},2}$, $\|\cdot\|_{\mathbf{X},3}$, $\|\cdot\|_{\mathbf{X},4}$, and $\|\cdot\|_{\mathbf{X},20}$, where the X_i are exponential with parameter $\lambda = 1$

These are the 1 and m norms, so

$$\int |x|^p \, dP_Y \leq \int |x| \, dP_Y + \int |x|^m \, dP_Y = \|\boldsymbol{\lambda}\|_{\mathbf{X},1} + \Gamma(m+1)\|\boldsymbol{\lambda}\|_{\mathbf{X},m}^m.$$

Thus, we can bound $f(p)$. Because the interval of possible integral values

$$\left[\|\boldsymbol{\lambda}\|_{\mathbf{X},1}, \|\boldsymbol{\lambda}\|_{\mathbf{X},1} + \Gamma(m+1)\|\boldsymbol{\lambda}\|_{\mathbf{X},m}^m\right]$$

is compact, for any $p \in [1, m]$ and sequence $p_1, p_2, \ldots$ that converges to p, the integrals $\int |x|_i^p \, dP_Y$ converge to $\int |x|^p \, dP_Y$. So $\Gamma(p+1)f^p$ is continuous. Since $\Gamma(p+1)$ is a positive continuous function for positive p and p is bounded between 1 and m, f is continuous. □

This continuity is clear in Figure 5.3, where each p norm appears squeezed between the $p-1$ and $p+1$ norms. Coming up with polynomials to describe norms with odd p is difficult since the moment-generating function we use for even p does not work. When p is not an integer, it quickly becomes necessary to use a computer.

5.5.2 Three Dimensions

By using 3D modeling, we can plot the unit balls of our norms in three variables.

Example 116. Bernoulli random variables produce interesting norms (see Figures 5.4, 5.5, 5.6, and 5.7).

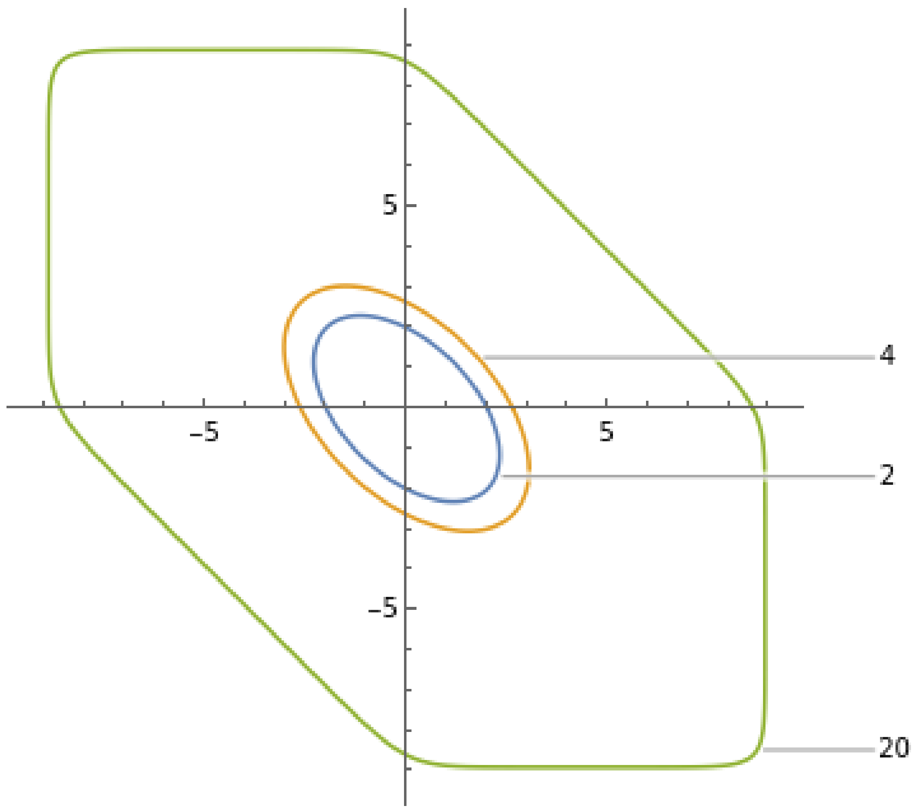


Figure 5.4: Unit circles for $\| \cdot \|_{\mathbf{X},2}$, $\| \cdot \|_{\mathbf{X},4}$, and $\| \cdot \|_{\mathbf{X},20}$, where the X_i are Bernoulli with parameter 1/2. For parameter 1/2, Bernoulli random variables produce norms in two variables whose unit circles, as the norm's exponent p increases, approach a hexagon resembling a multiple of $\| \cdot \|_1$ when λ_1 and λ_2 have the same sign, and $\| \cdot \|_\infty$ when they do not.

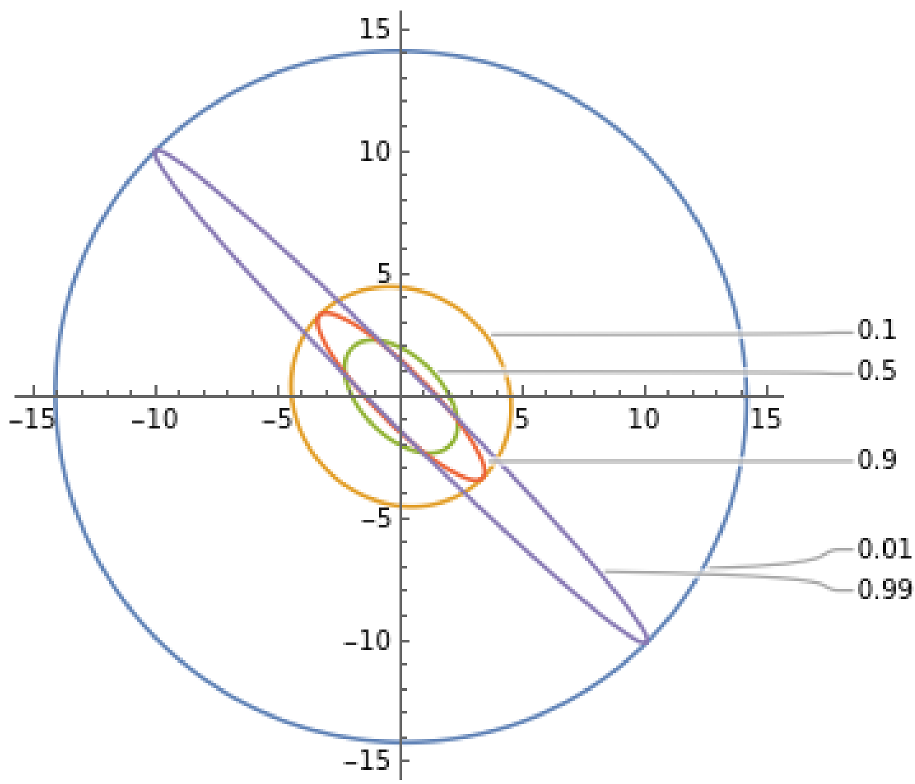


Figure 5.5: Unit circles for $\|\cdot\|_{\mathbf{X},2}$, where the X_i are Bernoulli with varying parameter. As the parameter increases, the unit circle grows narrower, since the most likely outcome is that X_1 and X_2 are both 1, in which case $|\langle \mathbf{X}, \boldsymbol{\lambda} \rangle| = |\lambda_1 + \lambda_2|$, which is not a norm: the equation $|\lambda_1 + \lambda_2| = 1$ produces two parallel lines, not a closed curve.

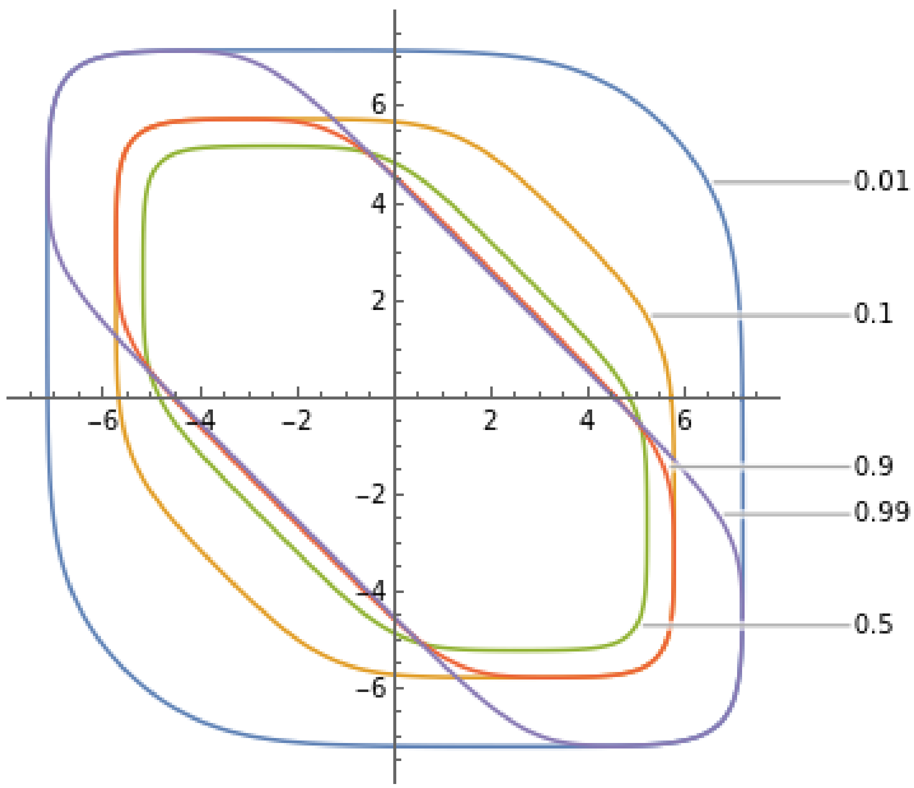


Figure 5.6: Unit circles for $\| \cdot \|_{\mathbf{X},10}$, where the X_i are Bernoulli with varying parameter. Since we are taking the norms with $p = 10$, we see roughly hexagonal curves. Again, the top left and top right corners of Bernoulli unit circles with parameter P come very close to those of parameter $1 - P$.

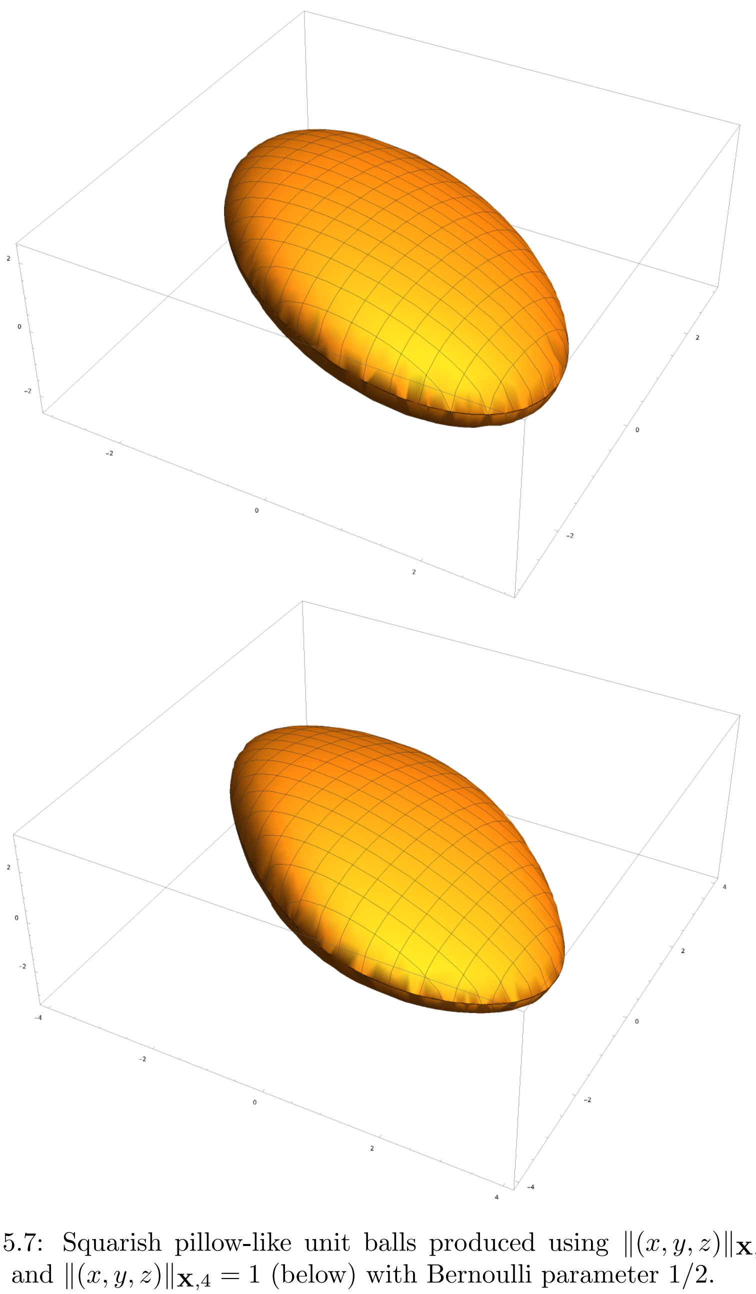


Figure 5.7: Squarish pillow-like unit balls produced using $\|(x, y, z)\|_{\mathbf{X},2} = 1$ (above) and $\|(x, y, z)\|_{\mathbf{X},4} = 1$ (below) with Bernoulli parameter 1/2.

Chapter 6

Pareto Distribution

The Pareto distribution, named after the Italian economist Vilfredo Pareto, is a power law distribution used in a variety of sciences to describe continuously the inverse relationship between rank and frequency. Pareto initially used it to describe the distribution of wealth in a society: as the amount of wealth increases, the probability that a randomly selected citizen has that amount of wealth decreases according to a roughly Pareto power law.

Pareto distributions, and the related, discrete Zipf distributions, which converge to the inverses of Pareto distributions, have been used to describe the distribution of sizes among grains of sand [13], population of settlements [12], and the frequency of words in linguistic corpora [16], as well as a host of other phenomena.

Pareto distributions have two parameters, x_m and α. Their probability density functions (PDFs) are given by

$$\mathrm{PDF}\,(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}},$$

with support on $[x_m, \infty)$.

When doing statistics, the parameter x_m is used to scale the x-axis and to mark the cutoff point for the object of study. For example, in examining the frequency of settlements by population, we could put the cutoff point for what we consider a settlement at x_m = 10,000 inhabitants. If the distribution is Pareto, then there will be many small towns with population just above 10,000, fewer towns with population of about 100,000, still fewer cities with population of about one million, and only a handful of metropolises.

More interesting, perhaps, than the often arbitrary lower cutoff x_m is the sharpness of the power law by which the frequency of cities with population x decreases as x increases. For example, in historically fragmented regions such as Europe, there are many small and medium-sized cities, with relatively few large metropolises. In Pareto terms, α is high. In contrast, regions such as California,

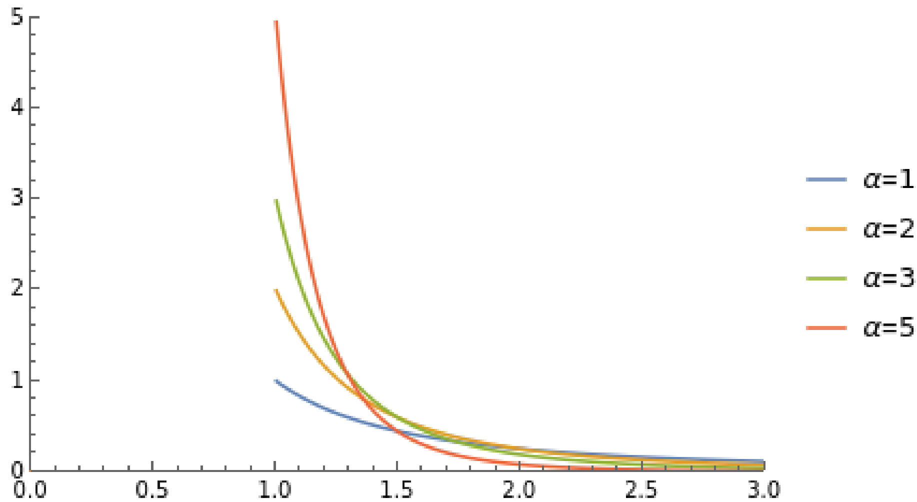


Figure 6.1: Pareto PDFs for varying α with $x_m = 1$

where the population is highly concentrated in Los Angeles, the Bay Area, and San Diego, the right tail is fatter, so α is lower.

The power laws of the Pareto distribution are also relevant to the study of fractals. In [10], the parameter $\frac{1}{\alpha}$ can be considered a measure of fractional dimension for sets whose sizes using different metrics distribute according to Pareto power laws.

6.1 Moments

The motivation for using the Pareto Distribution in relation to our random matrix norms is that not all moments exist for each α.

Pareto random variables have only as many finite moments as there are natural numbers $n < \alpha$. Accordingly, Pareto distributions never admit analytic moment generating functions, but the moments that do exist can be solved using the PDF and the usual integral for moments:

$$\mu_n = \int_1^\infty x^n \frac{\alpha}{x^{\alpha+1}}\, dx = \alpha \int_1^\infty x^{n-\alpha-1}\, dx = \frac{\alpha}{n-\alpha} x^{n-\alpha}\Big|_1^\infty = \frac{\alpha}{\alpha - n}$$

6.2 Pareto Random Vector Norms

Fortunately, our norm $\|\boldsymbol{\lambda}\|_{\mathbf{X},p}$ requires only the first p moments of the X_i, but does not require them to admit moment generating functions. As long as $\alpha > p$, then, our norm exists.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, in which the X_i are independent Pareto random variables with parameters α and $x_m = 1$. That is, let each X_i have the CDF

$$\mathrm{CDF}_X(x) = \begin{cases} 1 - \frac{1}{x^\alpha} & \text{if } x \geq x_m, \\ 0 & \text{if } x < x_m. \end{cases}$$

The pth moment of $\langle \mathbf{X}, \boldsymbol{\lambda} \rangle$ is

$$\mathbf{E}\,|\langle \mathbf{X}, \boldsymbol{\lambda} \rangle^p| = p! \|\boldsymbol{\lambda}\|_{\mathbf{X},p}^p.$$

Let $p < \alpha$ be even and let X_i be Pareto with parameters α and $x_m = 1$. Since the pth moment of each X_i exists, Theorem 100 ensures that pth moment of $\langle \mathbf{X}, \boldsymbol{\lambda} \rangle$ exists and Section 5.3 gives the formula in terms of the pth and smaller moments of X_i:

$$\mathbf{E}\,|\langle \mathbf{X}, \boldsymbol{\lambda} \rangle^p| = \sum_{|\pi|=p} \binom{p}{\pi} \mu_{\pi_i} \lambda_i^{\pi_i} = \sum_{|\pi|=p} \binom{p}{\pi} \frac{\alpha \lambda_i^{\pi_i}}{\alpha - \pi_i}. \tag{6.1}$$

Example 117. For $n = 2$,

$$\|\boldsymbol{\lambda}\|_{\mathbf{X},2}^2 = \frac{1}{2}\alpha \left(\frac{\lambda_1^2}{\alpha - 2} + \frac{2\alpha\lambda_1\lambda_2}{(\alpha-1)^2} + \frac{\lambda_2^2}{\alpha - 2} \right).$$

The norm where $p = 4$ is also an elegant polynomial.

$$\|\boldsymbol{\lambda}\|_{\mathbf{X},4}^4 = \frac{1}{24}\alpha \left(\frac{\lambda_1^4}{\alpha - 4} + \frac{4\alpha\lambda_1^3\lambda_2}{\alpha^2 - 4\alpha + 3} + \frac{6\alpha\lambda_2^2\lambda_1^2}{(\alpha-2)^2} + \frac{4\alpha\lambda_1\lambda_2^3}{\alpha^2 - 4\alpha + 3} + \frac{\lambda_2^4}{\alpha - 4} \right).$$

6.3 Limits

Sending α to its limits yields interesting results.

6.3.1 High α

As α approaches infinity, the Pareto distribution bunches up toward a single mass at 1. When the X_i resemble a constant of 1, $|\langle \mathbf{X}, \boldsymbol{\lambda} \rangle|$ resembles $|\langle \mathbf{1}, \boldsymbol{\lambda} \rangle| = |\sum_{i=1}^n \lambda_i|$, where $\mathbf{1}$ is the all-ones vector. Interestingly, though we can approach it with norms, the function $|\sum_{i=1}^n \lambda_i|$, which corresponds to $|\sum_{i=1}^n \lambda_i(A)| = |\operatorname{tr} A|$ in the Hermitian case, is not itself a norm because it is not positive definite. For example, the matrix

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

has trace of $1 + (-1) = 0$, but it is not the zero matrix.

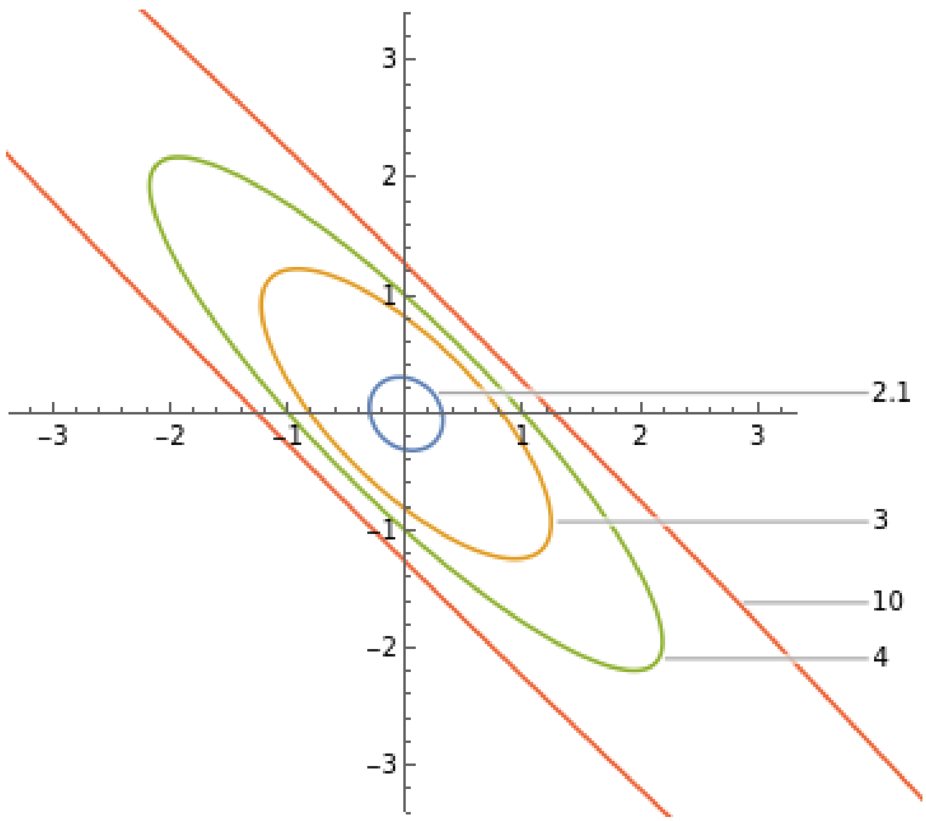


Figure 6.2: Unit circles of Pareto norms with $p = 2$ and varying α.

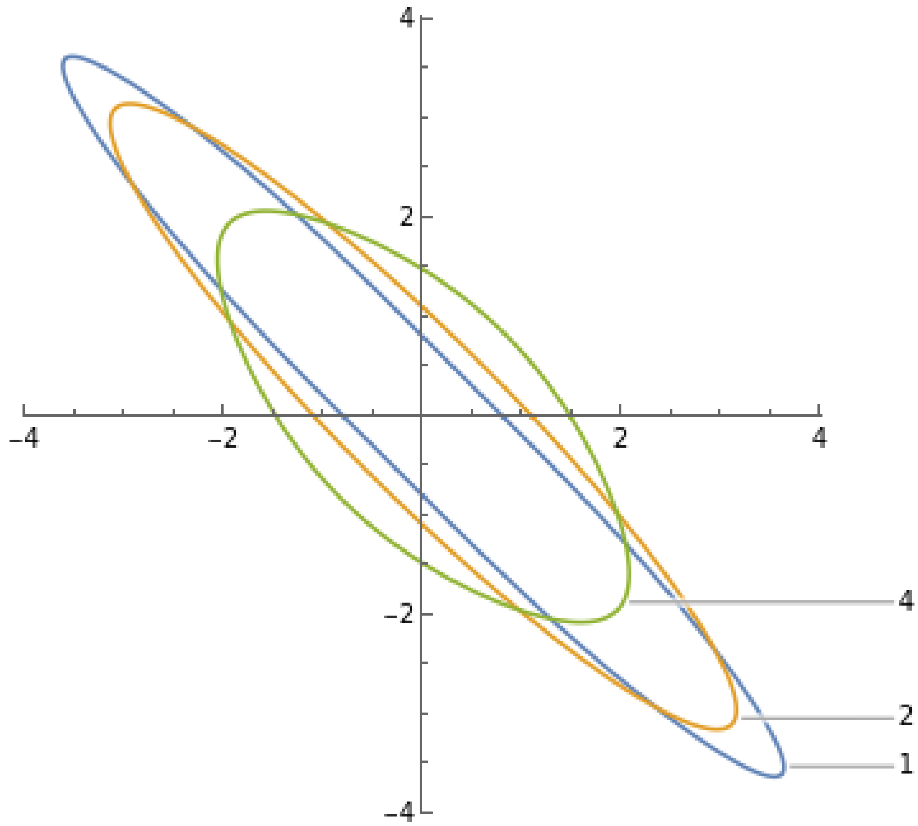


Figure 6.3: Unit circles of Pareto norms with $\alpha = 5$ and varying p.

In Figures 6.2 and 6.3, we see that the unit circles produced by the norms get longer and narrower as the difference between α and p grows. In $\mathbb{R}^2$, the limit of the unit circle as alpha approaches infinity is the curve $|\lambda_1 + \lambda_2| = 1$, which is represented by two parallel lines at $y = 1 - x$ and $y = -1 - x$.

The presence in the examples of binomial coefficients gives a hint, and indeed, the proof is straightforward.

Proposition 118. *Let $\mathbf{X}_\alpha$ be i.i.d. and let the X_i be Pareto random variables with parameter α. Then*

$$\lim_{\alpha \to p^+} \sqrt[p]{p!} \|\boldsymbol{\lambda}\|_{\mathbf{X}_\alpha, p} = |\sum_{i=1}^{n} \lambda_i|.$$

Proof. By 6.1

$$\lim_{\alpha \to p^+} \sqrt[p]{p!} \|\boldsymbol{\lambda}\|_{\mathbf{X}_\alpha, p} = \lim_{\alpha \to p^+} \left(\sum_{|\pi|=p} \binom{p}{\pi} \frac{\alpha \lambda_i^{\pi_i}}{\alpha - \pi_i} \right)^{1/p}.$$

As α approaches infinity, all moments $\mu_k = \frac{\alpha}{\alpha - k}$ approach 1, leaving

$$\lim_{\alpha \to p^+} \left(\sum_{|\pi|=p} \binom{p}{\pi} \frac{\alpha \lambda_i^{\pi_i}}{\alpha - \pi_i} \right)^{1/p} = \left(\sum_{|\pi|=p} \binom{p}{\pi} \lambda_i^{\pi_i} \right)^{1/p}.$$

By the multinomial theorem,

$$= \left(|\sum_{i=1}^{n} \lambda_i|^p \right)^{1/p} = |\sum_{i=1}^{n} \lambda_i|.$$

Thus as α approaches p from above, the norm approaches the absolute value of the sum of the λ_i, which is, again, not a norm. □

6.3.2 Low α

As α approaches p from above, the terms of

$$\sum_{|\pi|=p} \binom{p}{\pi} \frac{\alpha \lambda_i^{\pi_i}}{\alpha - \pi_i}$$

in which one of the π_i is p, and the others are zero, approach infinity, while the other terms remain small. The pth moment thus approaches

$$\sum_{i=1}^{n} \binom{p}{p} \frac{p \lambda_i^p}{\alpha - p} = \sum_{i=1}^{n} \frac{p \lambda_i^p}{\alpha - p}.$$

As α approaches p for even p, $\|\boldsymbol{\lambda}\|_{\mathbf{X}_{\alpha},p}$ approaches

$$\left|\frac{\sum_{i=1}^{n} \frac{p\lambda_i^p}{\alpha-p}}{p!}\right|^{\frac{1}{p}} = \frac{\left|\sum_{i=1}^{n} \lambda_i^p\right|^{\frac{1}{p}}}{(\alpha-p)\,(p-1)!^{\frac{1}{p}}} = \|\boldsymbol{\lambda}\|_p \left((\alpha-p)(p-1)!\right)^{-\frac{1}{p}},$$

where $\|\boldsymbol{\lambda}\|_p$ is the p-norm as shown in 1.2.

Bibliography

[1] Konrad Aguilar, Ángel Chávez, Stephan Ramon Garcia, and Jurij Volčič. Norms on complex matrices induced by complete homogeneous symmetric polynomials. *Bull. Lond. Math. Soc.*, 54(6):2078–2100, 2022.

[2] Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[3] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[4] Garrett Birkhoff. Three observations on linear algebra. *Univ. Nac. Tucumán. Revista A.*, 5:147–151, 1946.

[5] Ángel Chávez, Stephan Ramon Garcia, and Jackson Hurley. Norms on complex matrices induced by random vectors. *Canadian Mathematical Bulletin*, page 1–18, 2023. To appear.

[6] Morris H. DeGroot. *Probability and statistics*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1975.

[7] Stephan Ramon Garcia and Roger A Horn. *A Second Course in Linear Algebra*. Cambridge University Press, 2017.

[8] G. H. Hardy, J. E. Littlewood, and G. Pólya. Some simple inequalities satisfied by convex functions. *Messenger Math*, 58:145–152, 1929.

[9] A. S. Lewis. Group invariance and convex matrix analysis. *SIAM J. Matrix Anal. Appl.*, 17(4):927–949, 1996.

[10] Benoit B. Mandelbrot. *The fractal geometry of nature*. Schriftenreihe für den Referenten. [Series for the Referee]. W. H. Freeman and Co., San Francisco, Calif., 1982.

[11] Athanasios Papoulis. *Probability, random variables, and stochastic processes*. McGraw-Hill Series in Electrical Engineering. Communications and Information Theory. McGraw-Hill Book Co., New York, second edition, 1984.

[12] William J. Reed. On the rank-size distribution for human settlements. *Journal of Regional Science*, 42(1):1–17, 2002.

[13] William J. Reed and Murray Jorgensen. The double pareto-lognormal distribution—a new parametric model for size distributions. *Communications in Statistics - Theory and Methods*, 33(8):1733–1753, 2004.

[14] Terence Tao. 254a, notes 3a: Eigenvalues and sums of hermitian matrices. `https://terrytao.wordpress.com/2010/01/12/254a-notes-3a-eigenvalues-and-sums-of-hermitian-matrices/`, Jan 2010.

[15] Oscar Volij. An alternative proof of Hardy, Littlewood, and Pólya's (1929) necessary condition for majorization. *Econ. Theory Bull.*, 4(2):347–350, 2016.

[16] Shuiyuan Yu, Chunshan Xu, and Haitao Liu. Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. *Computing Research Repository (CoRR)*, abs/1807.01855, 2018.