MATH 547, STATISTICAL LEARNING THEORY, FALL 2019

STEVEN HEILMAN

Contents

Date: August 28, 2021

1. Introduction

> *With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.*
>
> John von Neumann

In Statistical Learning Theory, i.e. the statistical theory of machine learning, we will focus on the following questions:

- How do we choose a model that fits the data? (This is the one basic question in statistics.)
- How do we find the best model parameter? (This is the another basic question in statistics.)
- What algorithm for solving a given problem is most efficient? (Here "most efficient" can have several meanings, such as run time or memory usage, in the worst case or average case.)
- Even if we can efficiently find the best model parameter in a statistical problem, is that parameter meaningful? (As the quote of von Neumann points out, it may not be meaningful to fit a model with too many parameters to data. Moreover, finding correlations in data in not often meaningful. For example, the number of Nobel prizes awarded to a country is highly correlated with that country's chocolate consumption, but this correlation is not at all meaningful. See also Exercise 1.1.)
- Can we design algorithms that function when data arrives in a stream? (That is, suppose we only have access to a small amount of a larger group of data at any time, e.g. due to memory constraints on a computer. Can we still find good algorithms in this case?)

When answering these questions, we should also consider the following dichotomies for algorithms:

- Deterministic algorithms vs. randomized algorithms (i.e. those that use randomness)
- Exact vs. approximation approximation algorithms. (Sometimes an exact efficient algorithm does not exist, while an efficient algorithm that is approximately correct does exist.)
- Theoretical vs. practical algorithms. (Some algorithms work well in theory, but implied constants in their run times can be so large that such algorithms are impractical. Also, some algorithms work well according to practitioners, but there is no known theoretical guarantee that they work well for an arbitrary data set.)
- Supervised vs. unsupervised learning. (Supervised learning uses labelled data, unsupervised learning does not.)

In practice it is also important to consider when an algorithm can run in parallel on different computers, but we will not focus on this topic in this course.

Landau's Asymptotic Notation. Let $f,g\colon\mathbb{R}\to\mathbb{C}$. We use the notation $f(t)=o(g(t))$, $\forall\, t\in\mathbb{R}$ to denote $\lim_{t\to\infty}\left|\frac{f(t)}{g(t)}\right|=0$. We use the notation $f(t)=O(g(t))$ to denote that $\exists$ $c>0$ and $t_0\in\mathbb{R}$ such that $|f(t)|\leq c\,|g(t)|$ for all $t\geq t_0$. We write $f(t)=\Omega(g(t))$ when $\exists$ $c>0$ and $t_0\in\mathbb{R}$ such that $|f(t)|\geq c\,|g(t)|$ for all $t\geq t_0$. We write $f(t)=\Theta(g(t))$ when $f(t)=O(g(t))$ and $g(t)=O(f(t))$.

Standard Norm/Inner Product Notation. For any $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, define the standard inner product $\langle x, y\rangle := \sum_{i=1}^n x_i y_i$. We also denote $\|x\| := (\sum_{i=1}^n x_i^2)^{1/2}$ as the standard norm on $\mathbb{R}^n$.

For more notation see Section 12.

Exercise 1.1. Let $x^{(1)}, \ldots, x^{(m)}$ be m vectors in $\mathbb{R}^n$ with $\|x^{(i)}\| = 1$ for all $1 \leq i \leq m$. Let $\varepsilon > 0$. Assume that $m > (1 + 2/\varepsilon)^n$. Show that there exists $i, j \in \{1, \ldots, m\}$ such that $\|x^{(i)} - x^{(j)}\| < \varepsilon$.

Consequently, the vectors $x^{(i)}$ and $x^{(j)}$ are highly correlated, so that $\langle x^{(i)}, x^{(j)}\rangle > 1 - \varepsilon^2/2$. That is, if you have enough vectors on a unit sphere, at least two of them will be correlated with each other.

(If you want a big hint, look ahead to Proposition 4.3.)

To better understand our basic questions and dichotomies, we consider them for several specific examples.

Example 1.2 (**Computing Determinants**)**.** Let $n > 0$ be an integer. Suppose we want to compute the determinant of a real $n \times n$ matrix A with entries A_{ij}, $i, j \in \{1, \ldots, n\}$. An inefficient but straightforward way to do this is to directly use a definition of the determinant. Let S_n denote the set of all permutations on n elements. For any $\sigma \in S_n$, let $\text{sign}(\sigma) := (-1)^j$, where σ can be written as a composition of j transpositions (Exercise: this quantity is well-defined). (A transposition $\sigma \in S_n$ satisfies $\sigma(i) = i$ for at least $n-2$ elements of $\{1, \ldots, n\}$.) Then

$$\det(A) = \sum_{\sigma \in S_n} \text{sign}(\sigma) \prod_{i=1}^n A_{i\sigma(i)}.$$

This sum has $|S_n| = n!$ terms. So, if we use this formula to directly compute the determinant of A, in the worst case we will need to perform at least $(n+1) \cdot n!$ arithmetic operations. This is quite inefficient. We know a better algorithm from linear algebra class. We first perform row operations on A to make it upper triangular. Suppose B is an $n \times n$ real matrix such that BA represents one single row operation on A (i.e. adding a multiple of one row to another row, or swapping the positions of two rows). Then there are real $n \times n$ matrices $B_1, \ldots, B_m$ such that

$$B_1 \cdots B_m A \qquad (*)$$

is an upper triangular matrix. The matrices $B_1, \ldots, B_m$ can be chosen to first eliminate the left-most column of A under the diagonal, then the second left-most column entries under the diagonal, and so on. That is, we can choose $m \leq n(n-1)/2$, and each row operation involves at most $3n$ arithmetic operations. So, the multiplication of $(*)$ uses at most

$$3mn \leq 2n^3$$

arithmetic operations. The determinant of the upper diagonal matrix $(*)$ is then the product of its diagonal elements, and

$$\det(B_1 \cdots B_m A) = \det(B_1) \cdots \det(B_m) \det(A).$$

That is,

$$\det(A) = \frac{\det(B_1 \cdots B_m A)}{\det(B_1) \cdots \det(B_m)}.$$

So, $\det(A)$ can be computed with at most $2n^3+m+n \leq 4n^3 = O(n^3)$ arithmetic operations. Can we do any better?

It turns out that this is possible. Indeed, if it is possible to multiply two $n \times n$ real matrices with $O(n^a)$ arithmetic operations for some $a > 0$, then it is possible to compute the determinant of an $n \times n$ matrix with $O(n^a)$ arithmetic operations. The naïve way to multiply two real $n \times n$ matrices requires $O(n^3)$ arithmetic operations, so that $a = 3$ is achievable. However $a < 2.3728639$ is the best known upper bound [Gal14] (building upon Coppersmith-Winograd, Stothers, and Williams.) I do not think the algorithm with such a value of a has been implemented in practice, since the implied constants in its analysis are quite large, and apparently the algorithm does not parallelize. On the other hand, Strassen's algorithm has been implemented, and it has $a = \log 7/\log 2 \approx 2.807$.

What if we only have access to the matrix in the streaming setting? That is, suppose n is so large or the memory of the computer is so limited that we can only store a few of the rows of the matrix at one time. And once we view an entry of the matrix, it cannot be viewed again. In this case, it is impossible to know the determinant of the whole matrix. For example, suppose the first $n-1$ rows of the matrix A are known linearly independent vectors, and the last row of the matrix is known, except its last entry. If the last row is identical to the first row (except for their last entries), then changing the last entry could make the determinant of A zero or nonzero. So, it seems like we need to know essentially the whole matrix to even approximately know the determinant of the matrix, or even more fundamentally, the rank of the matrix. Indeed this is the case [CW09].

All of the above algorithms are deterministic, and they compute the determinant exactly (up to machine precision).

Remark 1.3. Interestingly, computing the permanent of a matrix

$$\mathrm{per}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} A_{i\sigma(i)}$$

is #**P**-complete. However, for any $\varepsilon > 0$, there is a $(1+\varepsilon)$ polynomial time randomized approximation algorithm for computing the permanent of a matrix with nonnegative entries [JSV04]. That is, for any $\varepsilon > 0$, and $0 < \delta < 1$ there is a randomized algorithm such that the following holds. For any $n \times n$ matrix A of nonnegative real numbers, the algorithm runs in time that is polynomial in $1/\varepsilon, n$, and $\log(1/\delta)$, and with probability at least $1-\delta$ the algorithm outputs a real number p such that

$$p \leq \mathrm{per}(A) \leq (1+\varepsilon)p.$$

On the other hand, for any constant c, the problem of approximating the permanent of an arbitrary matrix A is #**P**-hard [Aar11].

Example 1.4 (**Least Squares**)**.** Suppose we want to solve a least squares minimization problem. Suppose $w \in \mathbb{R}^m$ is an unknown vector, and A is a known $m \times n$ real matrix. Let $Z \in \mathbb{R}^n$ be a vector of i.i.d. standard Gaussian random variables. Our observation is $y := Aw + Z$, and the goal is to recover the unknown vector w. In linear least squares regression, we try to determine the best linear relationship w between the rows of A and the observation y. Assume that $n \leq m$ and the matrix A has full rank (so that $A^T A$ is invertible). The vector $x \in \mathbb{R}^m$ that minimizes the quantity

$$\|y - Ax\|^2 := \sum_{i=1}^{n} (y_i - (Ax)_i)^2$$

is then

$$x := (A^T A)^{-1} A^T y. \qquad (*)$$

Equivalently, x minimizes

$$\mathbb{E} \, \|y - x\|^2$$

over all choices of vectors $x \in \mathbb{R}^m$ such that $x = By$ for some $n \times m$ real matrix B, and such that $\mathbb{E}x = w$. (Since Z is the only random variable here, $\mathbb{E}$ denotes expected value with respect to Z.)

So, solving the linear least squares minimization problem could "just" use equation $(*)$. However, inverting a matrix directly is inefficient and could introduce numerical error. Below are some alternative ways of computing x. From $(*)$, we first write

$$A^T A x = A^T y,$$

and we then do any one of the following.

- Compute the Cholesky decomposition of $A^T A$. That is, we write $A^T A = R^T R$ where R is an upper triangular $n \times n$ matrix with positive diagonal elements. Then, $R^T R x = A^T y$, and solve the following simpler problems: (1) solve $R^T z = A^T y$ for the unknown $z \in \mathbb{R}^n$, then (2) solve $Rx = z$ for $x \in \mathbb{R}^n$. This is our desired x, since
$$R^T R x = R^T z = A^T y.$$
- Compute the QR decomposition of the matrix A. That is, we write $A = QR$ where Q is an $m \times n$ rectangular matrix with $Q^T Q = I_n$, where I_n denotes the $n \times n$ identity matrix, and R is an upper triangular $n \times n$ matrix with positive diagonal elements. Then $(A^T A)^{-1} = (R^T Q^T Q R)^{-1} = (R^T R)^{-1}$, so $(A^T A)^{-1} A^T = (R^T R)^{-1} R^T Q^T$, and $(R^T R)^{-1} R^T = R^{-1}$, so we have
$$x \overset{(*)}{=} R^{-1} Q^T y.$$
- Compute the singular value decomposition of the matrix A. (That is, we write $A = USV$, where U is an $m \times m$ orthogonal matrix, V is an $n \times n$ orthogonal matrix, and S is an $m \times n$ diagonal matrix with nonnegative entries.) Then $A^T A x = V^T S^2 V x = A^T y$, and solve the following simpler problems: (1) solve $V^T z = A^T y$ for the unknown $z \in \mathbb{R}^n$, then (2) solve $S^2 V x = z$ for $x \in \mathbb{R}^n$. This is our desired x, since
$$A^T A x = V^T S^2 V x = V^T z = A^T y.$$

These algorithms all use $O(n^3)$ arithmetic operations, and they assume we have access to the whole matrix A.

What if we only have streaming access to A? Put another way, is there an algorithm that only needs to access a small amount of A at any single time? The answer is yes. Recall that A is an $m \times n$ matrix with $m \geq n$. In some cases, m will be much larger than n. The recursive algorithm described below only needs to store an $n \times n$ matrix of A at any given time, and it is given access to one row of A at a time.

For a more modern view of this problem, see e.g. [KKP17].

Exercise 1.5. Let A be an $m \times n$ real matrix with $m \geq n$. Show that A has rank n if and only if $A^T A$ is positive definite.

(Hint: $A^T A$ is always positive semidefinite.)

Algorithm 1.6 (**Recursive Least Squares/ Online Learning**)**.** Let $m \geq n$, let A be an $m \times n$ real matrix. Let $a^{(1)}, \ldots, a^{(m)} \in \mathbb{R}^n$ be row vectors which are the rows of A (data), and let $b \in \mathbb{R}^m$. For any $j \geq n$, let

$$A_j := \begin{pmatrix} a^{(1)} \\ \vdots \\ a^{(j)} \end{pmatrix}, \qquad b^{(j)} := \begin{pmatrix} b_1 \\ \vdots \\ b_j \end{pmatrix}.$$

Assume that A_n has rank n (so that $A_n^T A_n$ is invertible). Define

$$x^{(n)} := (A_n^T A_n)^{-1} A_n^T b^{(n)} \in \mathbb{R}^n, \qquad P_n := (A_n^T A_n)^{-1}.$$

For any $j \geq n$, define

$$P_{j+1} = P_j - \frac{P_j (a^{(j+1)})^T a^{(j+1)} P_j^T}{1 + a^{(j+1)} P_j (a^{(j+1)})^T}.$$
$$x^{(j+1)} = x^{(j)} + P_{j+1} (a^{(j+1)})^T (b_{j+1} - a^{(j+1)} x^{(j)}).$$

The vectors $x^{(n)}, \ldots, x^{(m)}$ recursively minimize the quantity $\|Ax - b\|^2$ in the following sense.

Proposition 1.7. *Let $\lambda > 0$. Let $x^{(n)}, \ldots, x^{(m)}$ be the output of Algorithm 1.6. Let $n \leq j \leq m$. Define $f_j \colon \mathbb{R}^n \to \mathbb{R}$ by*

$$f_j(x) := \frac{1}{2} \sum_{i=1}^{j} (\langle x, a^{(i)} \rangle - b_i)^2, \qquad x \in \mathbb{R}^n.$$

Then $x^{(j)}$ minimizes f_j on $\mathbb{R}^n$. In particular, when $j = m$, $x^{(m)}$ minimizes $\|Ax - b\|^2$.

Proof. We induct on j. The case $j = n$ follows by definition of $x^{(n)}$ and by $(*)$. We now complete the inductive step. Assume the Proposition holds for j, and consider the case $j+1$. Define $G_j := (A_j^T A_j)$.

First, note that

$$G_{j+1} = \begin{pmatrix} A_j^T & (a^{(j+1)})^T \end{pmatrix} \begin{pmatrix} A_j \\ a^{(j+1)} \end{pmatrix} = A_j^T A_j + (a^{(j+1)})^T a^{(j+1)} = G_j + (a^{(j+1)})^T a^{(j+1)}. \qquad (*)$$

By the inductive hypothesis and $(*)$, we have $x^{(j)} = G_j^{-1} A_j^T b^{(j)}$. So,

$$A_j^T b^{(j)} = G_j G_j^{-1} A_j^T b^{(j)} = G_j x^{(j)} \overset{(*)}{=} (G_{j+1} - (a^{(j+1)})^T a^{(j+1)}) x^{(j)}. \qquad (**)$$

From $(*)$, the minimum of f_{j+1} on $\mathbb{R}^n$ occurs when

$$\begin{aligned}
x &= G_{j+1}^{-1}A_{j+1}^T b^{(j+1)} = G_{j+1}^{-1}\begin{pmatrix} A_j \\ a^{(j+1)} \end{pmatrix}^T \begin{pmatrix} b^{(j)} \\ b_{j+1} \end{pmatrix} = G_{j+1}^{-1}(A_j^T b^{(j)} + b_{j+1}(a^{(j+1)})^T) \\
&\overset{(**)}{=} G_{j+1}^{-1}\Big(G_{j+1}x^{(j)} - (a^{(j+1)})^T a^{(j+1)} x^{(j)} + b_{j+1}(a^{(j+1)})^T\Big) \\
&= x^{(j)} + G_{j+1}^{-1}(a^{(j+1)})^T(b_{j+1} - a^{(j+1)}x^{(j)}).
\end{aligned}$$

Comparing this formula to the definition of $x^{(j+1)}$ in Algorithm 1.6, it remains to manipulate the G_{j+1}^{-1} term. Applying Exercise 1.8 to $(*)$,

$$G_{j+1}^{-1} = (G_j + (a^{(j+1)})^T a^{(j+1)})^{-1} = G_j^{-1} - \frac{G_j^{-1}(a^{(j+1)})^T a^{(j+1)} G_j^{-1}}{1 + a^{(j+1)} G_j^{-1} (a^{(j+1)})^T}.$$

Finally, note that $P_n = G_n^{-1}$, and since the matrices P_j and G_j^{-1} satisfy the same recursion, we get $P_j = G_j^{-1}$, completing the proof.

□

Exercise 1.8. Show the following identity. Let A be an $r \times r$ real matrix, let U be an $r \times s$ real matrix, and let V be an $s \times r$ real matrix. Assume that A is invertible and that $I + VA^{-1}U$ is invertible, where I is the $s \times s$ identity matrix. Then $A + UV$ is invertible and

$$(A + UV)^{-1} = A^{-1} - (A^{-1}U)(I + VA^{-1}U)^{-1}(VA^{-1}).$$

In particular, if $s = 1$, we get the Sherman-Morrison formula:

$$(A + UV)^{-1} = A^{-1} - \frac{A^{-1}UVA^{-1}}{1 + VA^{-1}U}.$$

Example 1.9 (Minimum Vertex Cover). Suppose we have a set of vertices $V := \{1, \ldots, n\}$ and a set of undirected edges $E \subseteq \{\{i, j\} \colon i, j \in V\}$. The goal of the minimum vertex cover problem is to find the smallest vertex cover of the graph $G = (V, E)$. A vertex cover is a subset $S \subseteq V$ such that every $\{i, j\} \in E$ satisfies $i \in S$ or $j \in S$. More generally, for any $i \in V$, let $c_i \in \mathbb{R}$, $c_i \geq 0$. We are asked to minimize the weighted sum

$$\sum_{i \in S} c_i$$

over all $S \subseteq V$ such that every $\{i, j\} \in E$ satisfies $i \in S$ or $j \in S$. (To recover the unweighted minimum vertex cover problem, let $c_i := 1$ for all $1 \leq i \leq n$.) For a somewhat contrived example, we could think of the vertices as cities, and the set S as a subset of cities where cell phone towers are placed. And each cell phone tower is designed to cover the city in which it resides, and any adjacent cities.

There is a simple polynomial time algorithm that, given $G = (V, E)$ whose minimum vertex cover has size $a > 0$, finds a vertex cover $S \subseteq V$ such that

$$a \leq |S| \leq 2a.$$

This algorithm is therefore called a 2-approximation algorithm for the minimum vertex cover problem.

It is known that there exists an $\alpha > 1$ such that finding an α-approximation to the minimum vertex cover problem is NP-complete. So, it seems impossible to efficiently solve the minimum vertex cover problem. More specifically, if we were given $G = (V, E)$ whose minimum vertex cover has size $c > 0$, and we could finds a vertex cover $S \subseteq V$ in polynomial time such that

$$c \leq |S| \leq \alpha c,$$

then $P = NP$, and we would solve one of the Millennium Prize Problems. Since it is widely believed that $P \neq NP$, it is doubtful that the Minimum Vertex Cover problem can be solved in time polynomial in n. In fact, the constant $\alpha = 2$ is believed to be the best possible approximation. That is, it is conjectured that it is NP-hard to find a 1.9999-approximation to the minimum vertex cover problem [KR08].

Let us describe the deterministic algorithm.

Algorithm 1.10 (**Greedy Algorithm for Unweighted Min-Vertex Cover**)**.** Given a graph $G = (V, E)$, begin with $S := \emptyset$. Iterate the following procedure until E is empty.

- Choose some edge $\{i, j\} \in E$.
- Redefine $S := S \cup \{i\} \cup \{j\}$, and remove from E any edge whose endpoint is either i or j.

When E is empty, output S.

Proposition 1.11. *Algorithm 1.10 is a 2-approximation algorithm for the unweighted minimum vertex cover problem.*

Proof. Let a be the size of the minimum vertex cover of a graph $G = (V, E)$. Let E' be the set of edges that is found in the first step of all iterations of Algorithm 1.10. By step two of the iteration, each edge in E' is encountered exactly once in step one, and the vertices of each edge in E' are all disjoint from each other. Moreover, steps one and two of the iteration imply that S is a vertex cover. Also, step two implies that the minimum vertex cover must contain at least one endpoint of each edge in E', otherwise some edge would be not covered by the minimum vertex cover. So, $|E'| \leq a$. In summary,

$$|S| = 2\,|E'| \leq 2a.$$

Lastly, $|S| \geq a$ by definition of c. □

There is also a randomized 2-approximation algorithm for the general (weighted) minimum vertex problem using linear programming. The Minimum Vertex Cover problem is:

$$\text{minimize} \quad \langle c, x\rangle \quad \text{subject to the constraints}$$

$$x_i \in \{0, 1\}, \quad \forall\, i \in V$$

$$x_i + x_j \geq 1, \quad \forall\, \{i, j\} \in E$$

We then solve the following linear program and try to relate its solution to the above problem.

$$\text{minimize} \quad \langle c, x\rangle \quad \text{subject to the constraints}$$

$$x_i \in [0, 1], \quad \forall\, i \in V$$

$$x_i + x_j \geq 1, \quad \forall\, \{i, j\} \in E$$

Recall that an $n \times n$ matrix A is **symmetric** if $A = A^T$. We denote the $n \times n$ identity matrix as I or I_n. Recall that an $n \times n$ matrix A is **orthogonal** if A is a real matrix such that $A^T A = AA^T = I$.

Theorem 1.12 (**Spectral Theorem for Real Symmetric Matrices**). *Let A be a real $n \times n$ symmetric matrix. Then there exists an $n \times n$ orthogonal matrix Q and a real diagonal matrix D such that*

$$A = Q^T D Q.$$

Equivalently, there exists an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A with real eigenvalues.

Example 1.13 (**Singular Value Decomposition (SVD) and Principal Component Analysis (PCA)**)**.** Suppose A is an $m \times n$ real matrix. We think of each row of A as a vector of data. The matrix $A^T A$ is then an $n \times n$ real symmetric matrix. By the Spectral Theorem 1.12, there exists an $n \times n$ orthogonal matrix Q and a real diagonal matrix D such that

$$A^T A = Q^T D Q.$$

Similarly, the matrix AA^T is an $m \times m$ real symmetric matrix. The Spectral Theorem 1.12 implies that there exists an $m \times m$ orthogonal matrix R and a real diagonal matrix G such that

$$AA^T = R^T G R.$$

Note that all entries of D are nonnegative, since if $v \in \mathbb{R}^n$ is an eigenvector of $A^T A$ with eigenvalue $\lambda \in \mathbb{R}$, then

$$\lambda \left\|v\right\|^2 = \lambda\langle v, v\rangle = \langle v, A^T A v\rangle = \langle Av, Av\rangle = \left\|Av\right\|^2.$$

That is, we must have $\lambda \geq 0$. So, the square root of D (i.e. the diagonal matrix all of whose entries are the square roots of the entries of D) is well-defined. We denote the square root of D as $\sqrt{D}$, so that $(\sqrt{D})^2 = D$. Then (if $n \leq m$)

$$A = R^T \begin{pmatrix} \sqrt{D} \\ 0 \end{pmatrix} Q.$$

This factorization of A is called a **singular value decomposition** of A. (In general, the matrices $R^T, \sqrt{D}, Q$ are not uniquely determined by A, though the entries of $\sqrt{D}$ are uniquely determined by A.) The entries of $\sqrt{D}$ are called the **singular values** of A.

So, at least theoretically, a singular value decomposition exists. How can we find it on a computer? If the matrix A is relatively small, then we can compute $A^T A$ and find its eigenvalues and eigenvectors by the Power Method (see Exercise 1.18 below). The matrix D then consists of the eigenvalues and Q contains the eigenvectors of $A^T A$. The power method is especially efficient when A has very few nonzero entries.

If the matrix A is fairly large, we can instead randomly sample a small number of the rows and columns of A, perform a singular value decomposition on this smaller matrix, and use it to approximate the SVD of A itself [KV09, HMT11, Mah11].

SVD is used widely in practice. Note that since each data vector is a vector in $\mathbb{R}^n$, and each eigenvector of $A^T A$ is also a vector in $\mathbb{R}^n$, any vector $x \in \mathbb{R}^n$ can be written uniquely as a linear combination of eigenvectors of $A^T A$ (since these eigenvectors form an orthonormal basis of $\mathbb{R}^n$.) In many applications, it is generally expected that $A^T A$ has low rank or "approximately low rank." In such cases, the matrix $A^T A$ is well understood by examining its eigenvectors with largest eigenvalues. For example, if $A^T A$ has rank k, then only k of its eigenvectors are needed to understand $A^T A$. Similarly, if $A^T A$ has k large eigenvalues with the rest of them being close to zero, then only the first k eigenvectors are needed to approximately understand A^TA, and hence A.

PCA is designed with this observation in mind. In PCA, one uses the SVD of A (or of A with each row of A subtracted by the average of all rows of A), and we examine the eigenvectors of A^TA (and AA^T) with the largest eigenvalues. In particular, the eigenvectors of A^TA with largest eigenvalues are judged to be the "principal components" of any data vector $x \in \mathbb{R}^n$. PCA is used e.g. when Netflix tries to find recommendations for movies or shows you might enjoy. We consider each row of A to contain the preferences of one user, e.g. containing their ratings for various movies (say entry $A_{ij} \in \{1, 2, 3, 4, 5\}$ is the rating of user i for movie j). Suppose the eigenvectors of A^TA with the k largest eigenvalues are $y^{(1)}, \ldots, y^{(k)}$. We can then e.g. examine the k^{th} **spectral embedding** $f_k \colon \{1, \ldots, n\} \to \mathbb{R}^k$ that maps each movie to its vector of values on these eigenvectors:

$$f_k(i) := \left(y_i^{(1)}, \ldots, y_i^{(k)}\right), \qquad \forall\, i \in \{1, \ldots, n\}.$$

In practice, one observes that "similar" movies appear in clusters in the spectral embedding set $\{f_k(i)\}_{i=1}^n \subseteq \mathbb{R}^k$. We can then infer that someone who enjoys one movie in one cluster will enjoy other movies in the same cluster. The actual application is a bit more complicated since users do not rank all movies, so extra steps are made to fill or infer the missing entries of A. Since the initial data vectors are in $\mathbb{R}^n$ but the set $\{f_k(i)\}_{i=1}^n$ is in $\mathbb{R}^k$ with $k < n$, PCA is considered a **dimension reduction** method.

PCA is also used in facial recognition with the so-called "eigenfaces" method. In this case, each row of A is a vector (say of bitmap image values) of the front view of someone's face. PCA is then performed on a standard large data set. Given a new face image $x \in \mathbb{R}^n$, we associate x to its "feature vector" $g_k(x) := (\langle x, y^{(1)} \rangle, \ldots, \langle x, y^{(k)} \rangle)$. We can then guess if two different faces $x, y \in \mathbb{R}^n$ are the same or not the same by computing the distance $\|g_k(x) - g_k(y)\|$. Alternatively, if we want to compare x to the set of images in the initial data set (on which the original PCA was performed), we can try to find $\min_{i=1,\ldots,n} \|g_k(x) - f_k(i)\|$, and then associate x with the value of $i \in \{1, \ldots, n\}$ achieving this minimum. The actual application is more complicated. Images in practice do not often show the front of a face, so extra steps can be made to infer the front view of a face from the image of a non-frontal view of a face.

Since PCA and SVD use unlabelled data, they are considered methods in **unsupervised learning**.

Example 1.14 (**k-means Clustering**)**.** Let k, m, n be positive integers. Given m vectors $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$, the k-means clustering problem asks for the partition $S_1, \ldots, S_k$ of $\{1, \ldots, m\}$ into k sets which minimizes

$$\sum_{i=1}^{k} \sum_{j \in S_i} \left\| x^{(j)} - \frac{1}{|S_i|} \sum_{p \in S_i} x^{(p)} \right\|^2. \qquad (*)$$

For each $1 \leq i \leq k$, the term $\frac{1}{|S_i|} \sum_{p \in S_i} x^{(p)}$ is the center of mass (or barycenter) of the points in S_i, so each term in the sum is the squared distance of some point in S_i from the barycenter of S_i. So, k-means clustering can be seen as a kind of geometric version of least-squares regression. We emphasize that k is fixed.

How can we solve this problem? The most basic algorithm is a "gradient-descent" procedure known as Lloyd's Algorithm.

Algorithm 1.15 (LLoyd's Algorithm). Let $x^{(1)},\ldots,x^{(m)} \in \mathbb{R}^n$. Begin by choosing $y^{(1)},\ldots,y^{(k)} \in \mathbb{R}^n$ (randomly or deterministically), and define $T_i := \emptyset$ for all $1 \leq i \leq k$. Repeat the following procedure:

- For each $1 \leq i \leq k$, re-define
$$T_i := \{j \in \{1,\ldots,m\} \colon \left\|x^{(j)} - y^{(i)}\right\| = \min_{p=1,\ldots,k} \left\|x^{(j)} - y^{(p)}\right\| \}.$$
(If more than one p achieves this minimum, assign j to an arbitrary such minimal p.) (The sets $T_1,\ldots,T_m$ are called Voronoi regions.)
- For each $1 \leq i \leq k$, re-define $y^{(i)} := \frac{1}{|T_i|}\sum_{p\in T_i} x^{(p)}$.

Once this procedure is iterated a specified number of times, output $S_i := T_i$.

Algorithm 1.15 can be considered a "gradient-descent" procedure since the first step of the iteration always decreases the quantity $(*)$ by the definition of $(*)$, and the second step of the iteration always decreases $(*)$ by Exercise 1.16.

Exercise 1.16. Let $x^{(1)},\ldots,x^{(m)} \in \mathbb{R}^n$. Let $y \in \mathbb{R}^n$. Show that
$$\sum_{j=1}^{m} \left\|x^{(j)} - \frac{1}{m}\sum_{p=1}^{m} x^{(p)}\right\|^2 \leq \sum_{j=1}^{m} \left\|x^{(j)} - y\right\|^2.$$
That is, the barycenter is the point in $\mathbb{R}^n$ that minimizes the sum of squared distances.

While Lloyd's Algorithm 1.15 decreases the value of the quantity $(*)$, iterating this algorithm many times does not guarantee that a global minimum of $(*)$ is found. To see why, recall that the local minimum of a function $f\colon \mathbb{R} \to \mathbb{R}$ may not be the same as a global minimum. So, while Lloyd's Algorithm 1.15 is simple and it might work well in certain situations, it has no general theoretical guarantees. Some work has been done to make a "wise" choice of the initial points $y^{(1)},\ldots,y^{(k)}$.

So, are there any efficient algorithms with theoretical guarantees? For any $\varepsilon > 0$, there is a $9+\varepsilon$ factor approximation algorithm for the k-means clustering problem [KMN+04] with a polynomial running time (that does not depend on k). This algorithm is based upon [Mat00]. This factor of 9 was recently improved to 6.457 in [ANFSW0] and to 6.12903 in [GOR+21]. It was shown [ACKS15] that there exists some $\varepsilon > 0$ such that approximating the k-means clustering problem with a multiplicative factor of $1+\varepsilon$ for all k is NP-hard. This result was improved recently in [CAS19]. Still, there is a rather large gap between the best general purpose algorithm, and the hardness result. Many algorithms can approximately solve the k-means clustering problem to a multiplicative factor of $1+\varepsilon$, but these algorithms always have an exponential dependence on k for their run times [HPM04]. So, if we try to use $k = 100$, which occurs in many applications, these algorithms seem to be impractical.

It is possible to combine dimension-reduction techniques (such as PCA or the Johnson-Lindenstrass Lemma, Theorem 5.6) with the above algorithms [CEM+15, MMR18], thereby saving time by working in lower dimensions. However, these techniques do not seem to improve the exponential run times in k.

Some streaming algorithms are known for k-means clustering [Che09, FMS07]. For example, the algorithm of [Che09] uses memory of size $O(d^2k^2\varepsilon^{-2}(\log n)^8)$ to approximately solve the k-means problem within a multiplicative factor of $1+\varepsilon$. Note that the points themselves are not actually stored in this algorithm, otherwise the memory requirement would be at least $\Omega(n)$. In fact, only the barycenters of the clusters are typically stored in these streaming algorithms, which drastically reduces the memory requirement.

Since k-means clustering uses unlabelled data (despite the fact that k needs to be specified), it is considered a method in **unsupervised learning**.

Exercise 1.17. Let $n \geq 2$ be a positive integer. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. For any $x, y \in \mathbb{R}^n$, define $\langle x, y\rangle := \sum_{i=1}^n x_i y_i$ and $\|x\| := \langle x, x\rangle^{1/2}$. Let $S^{n-1} := \{x \in \mathbb{R}^n \colon \|x\| = 1\}$ be the sphere of radius 1 centered at the origin. Let $x \in S^{n-1}$ be fixed. Let v be a random vector that is uniformly distributed in S^{n-1}. Prove:

$$\mathbb{E}\,|\langle x, v\rangle| \geq \frac{1}{10\sqrt{n}}.$$

Exercise 1.18 (The Power Method). This exercise gives an algorithm for finding the eigenvectors and eigenvalues of a symmetric matrix. In modern statistics, this is often a useful thing to do. The Power Method described below is not the best algorithm for this task, but it is perhaps the easiest to describe and analyze.

Let A be an $n \times n$ real symmetric matrix. Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the (unknown) eigenvalues of A, and let $v_1, \ldots, v_n \in \mathbb{R}^n$ be the corresponding (unknown) eigenvectors of A such that $\|v_i\| = 1$ and such that $Av_i = \lambda_i v_i$ for all $1 \leq i \leq n$.

Given A, our first goal is to find v_1 and λ_1. For simplicity, assume that $1/2 < \lambda_1 < 1$, and $0 \leq \lambda_n \leq \cdots \leq \lambda_2 < 1/4$. Suppose we have found a vector $v \in \mathbb{R}^n$ such that $\|v\| = 1$ and $|\langle v, v_1\rangle| > 1/n$. (From Exercise 1.17, a randomly chosen v satisfies this property.) Let k be a positive integer. Show that

$$A^k v$$

approximates v_1 well as k becomes large. More specifically, show that for all $k \geq 1$,

$$\left\|A^k v - \langle v, v_1\rangle \lambda_1^k v_1\right\|^2 \leq \frac{n-1}{16^k}.$$

(Hint: use the spectral theorem for symmetric matrices.)

Since $|\langle v, v_1\rangle|\,\lambda_1^k > 2^{-k}/n$, this inequality implies that $A^k v$ is approximately an eigenvector of A with eigenvalue λ_1. That is, by the triangle inequality,

$$\left\|A(A^k v) - \lambda_1(A^k v)\right\| \leq \left\|A^{k+1} v - \langle v, v_1\rangle \lambda_1^{k+1} v_1\right\| + \lambda_1 \left\|\langle v, v_1\rangle \lambda_1^k v_1 - A^k v\right\| \leq 2\frac{\sqrt{n-1}}{4^k}.$$

Moreover, by the reverse triangle inequality,

$$\left\|A^k v\right\| = \left\|A^k v - \langle v, v_1\rangle \lambda_1^k v_1 + \langle v, v_1\rangle \lambda_1^k v_1\right\| \geq \frac{1}{n} 2^{-k} - \frac{\sqrt{n-1}}{4^k}.$$

In conclusion, if we take k to be large (say $k > 10 \log n$), and if we define $z := A^k v$, then z is approximately an eigenvector of A, that is

$$\left\|A\frac{A^k v}{\|A^k v\|} - \lambda_1 \frac{A^k v}{\|A^k v\|}\right\| \leq 4n^{3/2} 2^{-k} \leq 4n^{-4}.$$

And to approximately find the first eigenvalue λ_1, we simply compute

$$\frac{z^T A z}{z^T z}.$$

That is, we have approximately found the first eigenvector and eigenvalue of A.

Remarks. To find the second eigenvector and eigenvalue, we can repeat the above procedure, where we start by choosing v such that $\langle v, v_1\rangle = 0$, $\|v\| = 1$ and $|\langle v, v_2\rangle| > 1/(10\sqrt{n})$. To find the third eigenvector and eigenvalue, we can repeat the above procedure, where we start by choosing v such that $\langle v, v_1\rangle = \langle v, v_2\rangle = 0$, $\|v\| = 1$ and $|\langle v, v_3\rangle| > 1/(10\sqrt{n})$. And so on.

Google's PageRank algorithm uses the power method to rank websites very rapidly. In particular, they let n be the number of websites on the internet (so that n is roughly 10^9). They then define an $n \times n$ matrix C where $C_{ij} = 1$ if there is a hyperlink between websites i and j, and $C_{ij} = 0$ otherwise. Then, they let B be an $n \times n$ matrix such that B_{ij} is 1 divided by the number of 1's in the i^{th} row of C, if $C_{ij} = 1$, and $B_{ij} = 0$ otherwise. Finally, they define

$$A = (.85)B + (.15)D/n$$

where D is an $n \times n$ matrix all of whose entries are 1.

The power method finds the eigenvector v_1 of A, and the size of the i^{th} entry of v_1 is proportional to the "rank" of website i.

Exercise 1.19. Run PCA on a "planted" data set on a computer, consisting of 100 samples in $\mathbb{R}^{10}$ of the random variable $(X, Y, Z_3, \ldots, Z_{10}) \in \mathbb{R}^{10}$ where X, Y are standard Gaussian random variables, Z_i is a mean i Gaussian random variable with variance 10^{-2}, for all $3 \leq i \leq 10$, and $X, Y, Z_3, \ldots, Z_{10}$ are all independent. (You can use your favorite computer program to simulate the random variables.)

Then, run PCA on Airline Safety Information, and try to find out something interesting (this part of the question is intentionally open ended). The data is here, with accompanying article here. (See also here.)

Exercise 1.20. Run a k-means clustering algorithm (e.g. Lloyd's algorithm) on a "planted" data set in $\mathbb{R}^2$ consisting of 50 samples from (X, Y) and another 50 samples from (Z, W) where X, Y, Z, W are all independent Gaussians with variance 1, X, W have mean zero, Y has mean 1 and Z has mean 2. Try at least the values $k = 2, 3, 4, 5$.

Then, run a k-means clustering algorithm on Airline Safety Information, and try to find out something interesting (this part of the question is intentionally open ended).

2. A General Supervised Learning Problem

In this course, one main focus is the following.

Problem 2.1 (Supervised Learning Problem). Let A, B be sets. Let $f\colon A \to B$ be an unknown function. The goal of the learning problem is to determine the function f on all of A using a small number of known values of f on A. Let $x^{(1)}, \ldots, x^{(k)} \in A$ and let $y^{(1)}, \ldots, y^{(k)} \in B$. It is known that

$$f(x^{(i)}) = y^{(i)}, \qquad \forall\, 1 \leq i \leq k.$$

We then want to exactly or approximately determine f on all of A.

The set of ordered pairs $\{(x^{(i)}, y^{(i)})\}_{i=1}^{k}$ is sometimes called the **training set**. The function f is sometimes called a **predictor**, **hypothesis** or **classifier**.

Without some assumptions on f, this problem is impossible to solve, since we could just arbitrarily define f on inputs other than $x^{(1)}, \ldots, x^{(k)}$. So, one must add some assumptions on a class of functions f under consideration. One of the most basic and well studied examples is the following subset of Boolean functions $f\colon \{-1,1\}^n \to \{-1,1\}$.

Definition 2.2 (**Linear Threshold Functions**)**.** A function $f\colon \{-1,1\}^n \to \{-1,1\}$ is called a **linear threshold function** if there exists $w \in \mathbb{R}^n$ and $t \in \mathbb{R}$ such that

$$f(x) = \mathrm{sign}(\langle w, x\rangle - t), \qquad \forall\, x \in \{-1,1\}^n.$$

Here

$$\mathrm{sign}(s) := \frac{s}{|s|} = \begin{cases} 1 & \text{, if } s > 0 \\ 0 & \text{, if } s = 0 \\ -1 & \text{, if } s < 0. \end{cases}$$

A linear threshold function could also be called a **single-layer neural network**.

Here we are assuming that f defined in this way never takes the value 0, and we will always make this assumption unless otherwise stated.

Linear threshold functions can be understood geometrically as two sets of points on either side of a hyperplane.

Definition 2.3 (**Hyperplane**)**.** Let $w \in \mathbb{R}^n, t \in \mathbb{R}$. A **hyperplane** in $\mathbb{R}^n$ is a set of the form

$$\{x \in \mathbb{R}^n \colon \langle w, x\rangle = t\}.$$

That is, if $f\colon \{-1,1\}^n \to \{-1,1\}$ is a linear threshold function, then $\{x \in \{-1,1\}^n \colon f(x) = 1\}$ lies on one side of the hyperplane, and $\{x \in \{-1,1\}^n \colon f(x) = -1\}$ lies on the other side of the hyperplane. For this reason, linear threshold functions are sometimes called **half spaces**.

If we add the constraint that the unknown function f in Problem 2.1 is a linear threshold function, we arrive at the following problem.

Problem 2.4 (**Supervised Learning Problem for Linear Threshold Functions**)**.** Let $f\colon \{-1,1\}^n \to \{-1,1\}$ be an unknown linear threshold function. Let $x^{(1)}, \ldots, x^{(k)} \in \{-1,1\}^n$ and let $y_1, \ldots, y_k \in \{-1,1\}$. It is known that

$$f(x^{(i)}) = y_i, \qquad \forall\, 1 \le i \le k.$$

The goal of the problem is to find $w \in \mathbb{R}^n$ and $t \in \mathbb{R}$ such that

$$\mathrm{sign}(\langle w, x^{(i)}\rangle - t) = y_i, \qquad \forall\, 1 \le i \le k.$$

Remark 2.5 (**Reduction to Homogeneous $t = 0$ Case**)**.** For all $1 \le i \le n$, let $z^{(i)} := (x^{(i)}, 1) \in \{-1,1\}^{n+1}$ and let $w' := (w, t) \in \mathbb{R}^{n+1}$ in Problem 2.4. The goal of Problem 2.4 can then be restated as: find $w' \in \mathbb{R}^{n+1}$ such that

$$\mathrm{sign}(\langle w', z^{(i)}\rangle) = y_i, \qquad \forall\, 1 \le i \le k.$$

For this reason, we will often assume $t = 0$ below when discussing Problem 2.4.

Remark 2.6. Problem 2.4 can be understood geometrically as trying to find a hyperplane that separates the points $\{x \in \{-1,1\}^n \colon f(x) = 1\}$ from $\{x \in \{-1,1\}^n \colon f(x) = -1\}$.

This problem is ill-posed as stated already when $n = 2$ and $k = 4$, since it is impossible to separate pairs of opposite corners of $\{-1,1\}^2$ using a hyperplane.

Example 2.7. If $f(1,1) = f(-1,-1) = 1$ and $f(1,-1) = f(-1,1) = -1$, then there does not exist $w \in \mathbb{R}^2$ and $t \in \mathbb{R}$ such that $\mathrm{sign}(\langle w, x^{(i)}\rangle - t) = y_i \;\forall\; 1 \leq i \leq 4$. To see this, we can look at the "partial derivatives" of f as follows. Let $w = (w_1, w_2) \in \mathbb{R}^2$ and suppose $\langle w, (1,1)\rangle > t$, $\langle w, (-1,-1)\rangle > t$ and $\langle w, (1,-1)\rangle < t$, $\langle w, (-1,1)\rangle < t$. Then

$$2w_2 = \langle w, (1,1) - (1,-1)\rangle > 0, \qquad 2w_1 = \langle w, (1,1) - (-1,1)\rangle > 0,$$

So,

$$\langle w, (1,1)\rangle = w_1 + w_2 > -w_1 - w_2 = \langle w, (-1,-1)\rangle.$$

Therefore, $f(1,1) > f(-1,-1)$, a contradiction.

So, in order for the learning problem 2.4 to be exactly solvable, we must assume that there exists some $w \in \mathbb{R}^n, t \in \mathbb{R}$ such that

$$\mathrm{sign}(\langle w, x^{(i)}\rangle - t) = y_i, \qquad \forall\, 1 \leq i \leq k.$$

If we only want to find a $w \in \mathbb{R}^n, t \in \mathbb{R}$ such that this equality holds for 99% of $1 \leq i \leq k$, then that becomes a much different problem. We will discuss that problem in Section 3.

For a real world example of Problem 2.4, we can think of x as some sequence of letters (encoded in binary) from an email, so that f takes value 1 on a spam email, and f takes value -1 on non-spam emails. The goal is then to find the classifier f that can classify any given email as spam or non-spam, using the "training data" $x^{(1)}, \ldots, x^{(k)} \in \{-1,1\}^n$ and $y_1, \ldots, y_k \in \{-1,1\}$.

2.1. The Perceptron Algorithm.

The Perceptron Algorithm is a basic algorithm for solving Problem 2.4 from the 1960's. In this section, we will consider the homogeneous version of Problem 2.4. Let $x^{(1)}, \ldots, x^{(k)} \in \{-1,1\}^n$ and let $y_1, \ldots, y_k \in \{-1,1\}$ be given. It is known that

$$f(x^{(i)}) = y_i, \qquad \forall\, 1 \leq i \leq k.$$

The goal of the problem is to find $w \in \mathbb{R}^n$ such that

$$\mathrm{sign}(\langle w, x^{(i)}\rangle) = y_i, \qquad \forall\, 1 \leq i \leq k.$$

It is assumed that such a w exists.

Algorithm 2.8 (Perceptron Algorithm).

- Define $w^{(1)} := 0 \in \mathbb{R}^n$ and let $s := 1$
- If there exists some $1 \leq i \leq k$ such that $y_i \neq \mathrm{sign}(\langle w^{(s)}, x^{(i)}\rangle)$, i.e. a mis-classification occurs, define
$$w^{(s+1)} := w^{(s)} + y_i x^{(i)}.$$
- Increase the value of s by one. Repeat the previous step until no such i exists.
- Output $w := w^{(s)}$.

After the first step of the algorithm, we have $w^{(2)} = y_i x^{(i)}$, so we know that $\langle w^{(2)}, x^{(i)}\rangle = y_i$. That is, after one step of the algorithm, there is some $1 \leq i \leq k$ that is classified correctly. Unfortunately, after obtaining $w^{(3)}$ in the next step of the algorithm, if $x^{(i)}$ was the vector obtained from the first step of the algorithm, then $\langle w^{(3)}, x^{(i)}\rangle$ and y_i may not have the same sign. That is, the second step of the algorithm could mis-classify the point we obtained in the first step. So, a priori, Algorithm 2.8 may never terminate. Or at very least, the Algorithm may take a very large number of steps until it terminates. Either situation would be quite bad.

Note that the algorithm can be updated as it receives new data. For this reason, Algorithm 2.8 is called an **online learning** algorithm. However, the algorithm must keep all vectors $x^{(1)}, \ldots, x^{(k)}$ in memory, so it is not a streaming algorithm.

Theorem 2.9 (**Perceptron Algorithm Iteration Bound**). *Let $x^{(1)}, \ldots, x^{(k)} \in \mathbb{R}^n$ and let $y_1, \ldots, y_k \in \{-1, 1\}$ be given. Assume that there exists $\overline{w} \in \mathbb{R}^n$ such that*

$$\operatorname{sign}(\langle \overline{w}, x^{(i)} \rangle) = y_i, \qquad \forall\, 1 \leq i \leq k.$$

Define

$$\beta := \max_{i=1,\ldots,k} \left\| x^{(i)} \right\|, \qquad \theta := \min \Big\{ \|w\| : \forall\, 1 \leq i \leq k,\ y_i \langle w, x^{(i)} \rangle \geq 1 \Big\}.$$

Let $w \in \mathbb{R}^n$ achieve the minimum value in the definition of θ. Then the Perceptron Algorithm 2.8 terminates with a value of s satisfying

$$s \leq (\beta\theta)^2.$$

Proof. We first show that the weight vector $w^{(s)}$ "moves towards" w at each iteration of the algorithm:

$$\langle w^{(s+1)}, w \rangle \geq \langle w^{(s)}, w \rangle + 1, \qquad \forall\, s \geq 1. \qquad (*)$$

To see this, let $1 \leq i \leq k$ such that $x^{(i)}$ is selected during step two in iteration s of the algorithm. By definition of w, note that $y_i \langle w, x^{(i)} \rangle \geq 1$. And by definition of $w^{(s+1)}$ and w,

$$\langle w^{(s+1)}, w \rangle = \langle w^{(s)}, w \rangle + y_i \langle x^{(i)}, w \rangle \geq \langle w^{(s)}, w \rangle + 1.$$

We now show that the weight vector $w^{(s)}$ does not increase too much in length at each iteration:

$$\left\| w^{(s+1)} \right\|^2 \leq \left\| w^{(s)} \right\|^2 + \beta^2, \qquad \forall\, s \geq 1. \qquad (**)$$

Since $1 \leq i \leq k$ was selected in step two of the algorithm, we have $y_i \neq \operatorname{sign}(\langle w^{(s)}, x^{(i)} \rangle)$, i.e. $y_i \langle w^{(s)}, x^{(i)} \rangle < 0$. Using this and the definition of $w^{(s+1)}$,

$$\begin{aligned} \left\| w^{(s+1)} \right\|^2 &= \left\| w^{(s)} + y_i x^{(i)} \right\|^2 = \left\| w^{(s)} \right\|^2 + 2\langle w^{(s)}, y_i x^{(i)} \rangle + \left\| x^{(i)} \right\|^2 \\ &\leq \left\| w^{(s)} \right\|^2 + \left\| x^{(i)} \right\|^2 \leq \left\| w^{(s)} \right\|^2 + \beta^2. \end{aligned}$$

We now conclude the proof. Let $s \geq 1$. Induction and $(*)$ imply that $\langle w^{(s)}, w \rangle \geq s$. Induction and $(**)$ imply that $\left\| w^{(s)} \right\|^2 \leq s\beta^2$. By definition of w, $\|w\| = \theta$. So, the cosine of the angle between w and $w^{(s)}$ satisfies

$$1 \geq \frac{\langle w^{(s)}, w \rangle}{\|w^{(s)}\| \, \|w\|} \geq \frac{s}{\theta\sqrt{s}\beta} = \frac{\sqrt{s}}{\theta\beta}.$$

That is, $\sqrt{s} \leq \beta\theta$. □

Remark 2.10. Instead of going through all points in Algorithm 2.8 until the separating hyperplane is found, one could simply run through all points $x^{(1)}, \ldots, x^{(k)}$ in order. Doing so would speed up the algorithm, but the resulting weight vector w may not separate the points correctly. One might call this effect **underfitting**.

When θ is large, there are some data vectors $x^{(1)}, \ldots, x^{(k)}$ close to the separating hyperplane $\{x \in \mathbb{R}^n : \langle w, x \rangle = 0\}$. More specifically,

Lemma 2.11. *Let $x^{(1)}, \ldots, x^{(k)} \in \mathbb{R}^n$ and let $y_1, \ldots, y_k \in \{-1, 1\}$ be given. Assume that there exists $w \in \mathbb{R}^n$ such that*

$$\mathrm{sign}(\langle w, x^{(i)} \rangle) = y_i, \qquad \forall\, 1 \leq i \leq k.$$

Let $w \in \mathbb{R}^n$ achieve the minimum in $\min\left\{ \|w\| : \forall\, 1 \leq i \leq k,\ y_i\langle w, x^{(i)}\rangle \geq 1 \right\}$. Then $w/\|w\|$ achieves the maximum

$$\max_{v \in \mathbb{R}^n,\, \|v\|=1} \ \min_{i=1,\ldots,k} y_i\langle v, x^{(i)}\rangle. \qquad (*)$$

Proof. Let v achieve the maximum value $a > 0$ in $(*)$. Then $y_i\langle (v/a), x^{(i)}\rangle \geq 1$ by $(*)$. So, by definition of w, $\|w\| \leq \|v/a\| = 1/a$. If $1 \leq i \leq k$, by definition of w,

$$y_i\langle (w/\|w\|), x^{(i)}\rangle = \frac{1}{\|w\|} y_i \langle w, x^{(i)}\rangle \geq \frac{1}{\|w\|} \geq a.$$

So, w also achieves the maximum value in $(*)$. (Consequently, when we take $\min_{i=1,\ldots,k}$, all inequalities on the previous line must be equalities, so that $a = 1/\|w\|$.) □

When $\|v\| = 1$, $y_i\langle v, x^{(i)}\rangle$ is the distance from $x^{(i)}$ to the hyperplane $\{x \in \mathbb{R}^n \colon \langle x, v\rangle = 0\}$. So, in Theorem 2.9, the quantity

$$\theta := \min\left\{ \|w\| : \forall\, 1 \leq i \leq k\ y_i\langle w, x^{(i)}\rangle \geq 1 \right\}.$$

produces a vector w whose perpendicular hyperplane has the largest uniform distance a to the vectors $x^{(1)}, \ldots, x^{(k)}$. Since $a = 1/\|w\| = 1/\theta$, the quantity $1/\theta$ is sometimes called the **margin** of the vectors $x^{(1)}, \ldots, x^{(k)}$, as it represents the "widest" symmetric slab through the origin that can fit between all of the vectors.

2.1.1. *Variants of Learning Linear Threshold Functions.* In the case that the points *cannot* be separated by a hyperplane, suppose we try to find a hyperplane that correctly classifies the largest number of points. Unfortunately, this problem is known to be NP-hard [JP78] [ABSS97, Theorem 4]. In the literature, this problem is called the Open Hemisphere Problem, or Densest Hemisphere Problem.

Also, given that there is some half space that correctly classifies 99% of data points, it is NP-hard to find a half space that correctly classifies 51% of data points [FGKP06]. Note that correctly classifying 50% of data points is easy by just choosing any half space and deciding which side to label +1. In fact, if $\mathbf{NP} \not\subseteq \mathbf{TIME}(2^{(\log n)^{o(1)}})$, no polynomial time algorithm can distinguish between the following cases for learning half spaces with k input points: (i) $1 - 2^{-\Omega(\sqrt{\log k})}$ fraction of points can be correctly classified by some halfspace, or (ii) no more than $(1/2) + 2^{-\Omega(\sqrt{\log k})}$ fraction of points can be correctly classified by any halfspace [FGKP06, Theorem 4].

On the other hand, variants of the Perceptron algorithm do still learn linear threshold functions efficiently when we start with a separating hyperplane and then change each label independently with fixed probability less than 1/2 [BFKV98]. Under these assumptions, efficient learning can then occur in the PAC model (discussed in Section 3) and in the Statistical Query Model. Also, when a small fraction of labels can be arbitrarily corrupted (the so-called agnostic learning model) and some assumptions are made on allowable random samples of inputs, linear threshold functions can be learned efficiently [KKMS08]. With access to Gaussian samples, where a fraction of both samples and labels can be arbitrarily corrupted (the so-called nasty noise model), linear threshold functions can be learned efficiently [DKS18].

Instead of returning the final vector w that is output by the Perceptron algorithm, practitioners use an average of all weight vectors w over all time steps. Supposedly this works better than the Perceptron itself, and it is called the averaged perceptron.

Logistic regression is also said to work well in practice for classification tasks.

Remark 2.12 (Linear Programming). Let $x^{(1)},\ldots,x^{(k)}\in\{-1,1\}^{n}$ and let $y_1,\ldots,y_k\in\{-1,1\}$ be given. Let A be the $k\times n$ matrix whose i^{th} row is $x^{(i)}$. We can rewrite the goal of Problem 2.4 (with $t=0$) as finding $w\in\mathbb{R}^n$

$$\langle w,x^{(i)}\rangle y_i>0,\qquad\forall\, 1\leq i\leq k.$$

Since the $y_1,\ldots,y_k\in\{-1,1\}$ are fixed, this is a set of linear inequalities in w, i.e. finding the existence of such a w is a linear program in n dimensions with k constraints. It is conceivable to use a linear programming algorithm (such as the ellipsoid method or interior point method) to solve Problem 2.4, however it would most likely be slower than the Perceptron Algorithm. On the other hand, the run time of the linear program would not depend on the geometry of the points $x^{(1)},\ldots,x^{(k)}$, while our run time guarantee in Theorem 2.9 does depend on the geometry of the points (and this worst-case guarantee can increase exponentially with n).

Exercise 2.13. Let n be a positive integer. Let c_n be the number of boolean functions $f\colon\{-1,1\}^n\to\{-1,1\}$ that are linear threshold functions. This quantity is of interest since it roughly quantifies the “expressive power” of linear threshold functions for the supervised learning problem. It is known that

$$c_n=2^{n^2(1+o(1))}$$

So, Problem 2.4 asks for the linear threshold function that fits the given data among a family of functions of super-exponential size. For another perspective on the “expressive power” of linear threshold functions, we will look into the VC-dimension in e.g. Proposition 4.10.

Using an inductive argument prove the weaker lower bound

$$c_n\geq 2^{n(n-1)/2}.$$

(Hint: induct on n. If $f\colon\{-1,1\}^n\to\{-1,1\}$, consider $\overline{f}\colon\{-1,1\}^{n+1}\to\{-1,1\}$ defined (partially for now) so that $\overline{f}(x_1,\ldots,x_n,-1):=f(x_1,\ldots,x_n)$ for all $(x_1,\ldots,x_n)\in\{-1,1\}^n$. How many ways can we define $\overline{f}$ on the remaining “half” of the hypercube $\{-1,1\}^{n+1}$ such that $\overline{f}$ is a linear threshold function?)

As we will discuss in Section 8, it is of interest to state the general learning problem for compositions of linear threshold functions (i.e. neural networks). In this case, asymptotics for the number of such functions were recently found in https://arxiv.org/pdf/1901.00434.pdf.

Exercise 2.14. Let $a>0$. Let $X^{(1)},\ldots,X^{(k)}\in\mathbb{R}^n$ be independent identically distributed samples from a Gaussian random vector with mean $(a,0,\ldots,0)$ and identity covariance matrix). Let $X^{(k+1)},X^{(k+2)},\ldots,X^{(2k)}\in\mathbb{R}^n$ be independent identically distributed samples from a Gaussian random vector with mean $(-a,0,\ldots,0)$, where $a>0$ is known. As in our analysis of the perceptron algorithm, define

$$\mathcal{B}:=\max_{i=1,\ldots,2k}\left\|X^{(i)}\right\|$$

$$\Theta := \min\Big\{\, \|w\| : \forall\, 1 \leq i \leq 2k \;\, y_i\langle w, X^{(i)}\rangle \geq 1 \Big\}.$$

Give some reasonable estimates for $\mathbb{E}\mathcal{B}$ and $\mathbb{E}(1/\Theta)$ as a function of a.

2.2. **Embeddings and the "Kernel Trick".** In Remark 2.6, we noticed that exactly learning a linear threshold function amounts to separating data points into two sets by a separating hyperplane. However, some natural data sets may be sortable into two categories that cannot be separated by a hyperplane. For this reason, practitioners using the Perceptron Algorithm often "preprocess" their data so that it is sensibly separable by a hyperplane. The preprocessing can be described as a function $\phi\colon A \to C$, where A, C are sets. (Recall that we phrased Problem 2.1 as determining an unkonwn function $f\colon A \to B$.) The function $\phi\colon A \to C$ is sometimes called a **feature map**.

Problem 2.15 (Supervised Learning Problem for Linear Threshold Functions, with Kernel Trick). Let $f\colon C \to \{-1, 1\}$ be an unknown linear threshold function. Let $x^{(1)}, \ldots, x^{(k)} \in \mathbb{R}^n$ and let $y_1, \ldots, y_k \in \{-1, 1\}$. It is known that

$$f(\phi(x^{(i)})) = y_i, \qquad \forall\, 1 \leq i \leq k.$$

The goal of the problem is to find an inner product space C with inner product $\langle\cdot,\cdot\rangle_C$, an embedding $\phi\colon \mathbb{R}^n \to C$ and $w \in C$ such that

$$\mathrm{sign}(\langle w, \phi(x^{(i)})\rangle_C) = y_i, \qquad \forall\, 1 \leq i \leq k.$$

As above, we assume that Problem 2.15 can be solved when we apply the following modified Perceptron Algorithm.

Algorithm 2.16 (Kernel Perceptron Algorithm, Version 1).

- Define $w^{(1)} := 0 \in C$ and let $s := 1$
- If there exists some $1 \leq i \leq k$ such that $y_i \neq \mathrm{sign}(\langle w^{(s)}, \phi(x^{(i)})\rangle$, i.e. a misclassification occurs, define
$$w^{(s+1)} := w^{(s)} + y_i\phi(x^{(i)}).$$
- Increase the value of s by one. Repeat the previous step until no such i exists.
- Output $w := w^{(s)}$.

Sometimes the embedding ϕ may be difficult to compute or write explicitly. In such cases, it might be desirable to only define ϕ implicitly. In such a case, for all $1 \leq i \leq k$ we only need to define $\phi(x^{(i)})$ to be an element of some inner product space. And if we can rewrite the algorithm so that it only uses the values of the inner products, $\langle\phi(x^{(i)}), \phi(x^{(j)})\rangle_C$ $\forall\; 1 \leq i, j \leq k$, then we need only specify the values of these inner products. We then use the following equivalence.

Exercise 2.17. Let M be a $k \times k$ real symmetric matrix. Then M is positive semidefinite if and only if there exists a real $k \times k$ matrix R such that

$$M = RR^T.$$

In either case, if $r^{(i)}$ denotes the i^{th} row of R, we have

$$m_{ij} = \langle r^{(i)}, r^{(j)}\rangle, \qquad \forall\, 1 \leq i, j \leq k.$$

That is, the values of the inner products $\langle \phi(x^{(i)}), \phi(x^{(j)})\rangle_C$ can be equivalently specified as the values m_{ij} of a real symmetric positive semidefinite matrix. More generally, we have the following infinite-dimensional result from 1909.

Theorem 2.18 (Mercer's Theorem). *Let μ be a Borel measure on $\mathbb{R}^n$ such that the measure of any open set in $\mathbb{R}^n$ is positive. We denote $L_2(\mu) := \{f : \mathbb{R}^n \to \mathbb{R} : \int_{\mathbb{R}^n} |f(x)|^2 d\mu(x) < \infty\}$, and we equip $L_2(\mu)$ with the standard inner product $\int_{\mathbb{R}^n} f(x)g(x)d\mu(x)$ defined for any $f, g \in L_2(\mu)$. Let $m\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a continuous symmetric function ($m(x, y) = m(y, x)$ for all $x, y \in \mathbb{R}^n$) such that, for all $p \geq 1$, for all $z^{(1)}, \ldots, z^{(p)} \in \mathbb{R}^n$, for all $\beta_1, \ldots, \beta_p \in \mathbb{R}$ we have the* ***positive semi-definiteness condition***

$$\sum_{i,j=1}^{p} \beta_i\beta_j m(z^{(i)}, z^{(j)}) \geq 0, \qquad \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |m(x, y)|^2 d\mu(x)d\mu(y) < \infty.$$

Then there exists an orthonormal basis $\{\psi_i\}_{i=1}^{\infty}$ of $L_2(\mu)$ and a sequence of nonnegative real numbers $\{\lambda_i\}_{i=1}^{\infty}$ such that m is equal to the following series, which converges absolutely pointwise:

$$m(x, y) = \sum_{i=1}^{\infty} \lambda_i\psi_i(x)\psi_i(y), \qquad \forall\, x, y \in \mathbb{R}^n.$$

Exercise 2.19. Let μ be a Borel measure on $\mathbb{R}^n$ such that the measure of any open set in $\mathbb{R}^n$ is positive. Let $m\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be continuous with $\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |m(x, y)|^2 d\mu(x)d\mu(y) < \infty$. Show that the following two positive semidefinite conditions on m are equivalent:

- $\forall\ p \geq 1$, for all $z^{(1)}, \ldots, z^{(p)} \in \mathbb{R}^n$, for all $\beta_1, \ldots, \beta_p \in \mathbb{R}$ we have
$$\sum_{i,j=1}^{p} \beta_i\beta_j m(z^{(i)}, z^{(j)}) \geq 0.$$
- $\forall\ f \in L_2(\mu)$, we have
$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)f(y)m(x, y)d\mu(x)d\mu(y) \geq 0.$$

From either condition, we should see that the converse of Mercer's Theorem holds. We should also be able to deduce various properties of positive semidefinite (PSD) kernels. For example, a nonnegative linear combination of PSD kernels is PSD.

So, define $\ell_2 := \{(\beta_i)_{i=1}^{\infty} \colon \sum_{i=1}^{\infty} \beta_i^2 < \infty\}$ with the standard inner product $\langle(\beta_i)_{i=1}^{\infty}, (\gamma_i)_{i=1}^{\infty}\rangle := \sum_{i=1}^{\infty} \beta_i\gamma_i$ and $\phi\colon \mathbb{R}^n \to \ell_2$ by

$$\phi(x) := \{\sqrt{\lambda_i}\,\psi_i(x)\}_{i=1}^{\infty}, \qquad \forall\, x \in \mathbb{R}^n,$$

then

$$\langle \phi(x), \phi(y)\rangle = \sum_{i=1}^{\infty} \lambda_i\psi_i(x)\psi_i(y) = m(x, y), \qquad \forall\, x, y \in \mathbb{R}^n.$$

That is, Theorem 2.18 is an infinite-dimensional generalization of Exercise 2.17. Note that each ψ_i can be nonlinear, and in general the ϕ maps we use will be nonlinear maps.

Recall also that the Spectral Theorem 1.12 for real symmetric matrices $k \times k$ matrices M says there exists a real diagonal D and an orthogonal Q such that

$$M = Q^T DQ$$

We can write this in vector form to more closely match Theorem 2.18. For any $1 \leq p \leq k$, let λ_p denote the p^{th} diagonal entry of D and let $\psi^{(p)} \in \mathbb{R}^k$ denote the k^{th} row of Q. Then

$$m_{ij} = \sum_{p=1}^{k} \lambda_p \psi_i^{(p)} \psi_j^{(p)}, \qquad \forall\, 1 \leq i, j \leq k.$$

The function $m(x, y)$ (or the matrix m_{ij}) is called a **kernel**.

Also, since all countable Hilbert spaces are isometric, we can and will assume that $C = \ell_2$ with the standard inner product, which we denote as $\langle \cdot, \cdot \rangle$. In order to rewrite Algorithm 2.16, we need to express the algorithm using only $\langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$ terms. To this end, for all $1 \leq i \leq k$, let α_i be the number of times that $i \in \{1, \ldots, k\}$ is selected in step two of the algorithm. That is, α_i is the number of times that $x^{(i)}$ is "mis-classified" within the algorithm.

Let $s \geq 1$. Suppose at step s of the algorithm, $\widetilde{x}^{(s)} \in \{x^{(1)}, \ldots, x^{(k)}\}$ is selected with corresponding sign $\widetilde{y}_s \in \{-1, 1\}$. For any $s \geq 1$, induction implies that $w^{(s+1)}$ is a linear combination of $\phi(\widetilde{x}^{(1)}), \ldots, \phi(\widetilde{x}^{(s)})$ in Algorithm 2.16:

$$w^{(s+1)} = \sum_{j=1}^{s} \widetilde{y}_j \phi(\widetilde{x}^{(j)}). \qquad (*)$$

So, at the next step of the algorithm, we can compute

$$\mathrm{sign}(\langle w^{(s+1)}, \phi(\widetilde{x}^{(s+1)}) \rangle) \overset{(*)}{=} \mathrm{sign}\Big(\sum_{j=1}^{s} \widetilde{y}_j \langle \phi(\widetilde{x}^{(j)}), \phi(\widetilde{x}^{(s+1)}) \rangle \Big) =: \mathrm{sign}\Big(\sum_{j=1}^{k} y_j m(\widetilde{x}^{(j)}, \widetilde{x}^{(s+1)}) \Big).$$

Suppose the algorithm terminates after s steps. As mentioned above, for all $1 \leq j \leq k$, let α_j be the number of times that $\phi(x^{(j)})$ appears in the sum $(*)$. Then we can rewrite $(*)$ as

$$w := w^{(s+1)} = \sum_{j=1}^{k} \alpha_j y_j \phi(x^{(j)}).$$

We then plug in any x in the span of $x^{(1)}, \ldots, x^{(k)}$ so that

$$\mathrm{sign}(\langle w^{(k+1)}, \phi(x) \rangle) := \mathrm{sign}(\sum_{j=1}^{k} \alpha_j y_j m(x^{(j)}, x)).$$

In summary, we can rewrite Algorithm 2.16 in the following way in order to solve Problem 2.15.

Algorithm 2.20 (Kernel Perceptron Algorithm, Version 2).

Let $m \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a positive-definite function (kernel) satisfying the assumptions of Mercer's Theorem 2.18.

- Define $\alpha_1 := 0, \ldots, \alpha_k := 0$ and let $s := 1$.
- If there exists some $1 \leq i \leq k$ such that $y_i \neq \mathrm{sign}\Big(\sum_{j=1}^{k} \alpha_j y_j m(x^{(j)}, x^{(i)}) \Big)$, i.e. a mis-classification occurs, then increase the value of α_i by one.
- Increase the value of s by one. Repeat the previous step until no such i exists.
- Output the function $f(x) := \mathrm{sign}(\sum_{j=1}^{k} \alpha_j y_j m(x^{(j)}, x))$, valid for any $x \in \mathbb{R}^n$.

For some rigorous guarantees of a "kernel trick" in the context of a clustering-type problem, see [LOGT12].

Exercise 2.21. For each kernel function $m\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ below, find an inner product space C and a map $\phi\colon \mathbb{R}^n \to C$ such that

$$m(x,y) = \langle \phi(x), \phi(y) \rangle_C, \qquad \forall\, x, y \in \mathbb{R}^n.$$

Conclude that each such m is a positive semidefinite function, in the sense stated in Mercer's Theorem.

- $m(x,y) := 1 + \langle x, y \rangle \;\forall\, x, y \in \mathbb{R}^n$.
- $m(x,y) := (1 + \langle x, y \rangle)^d \;\forall\, x, y \in \mathbb{R}^n$, where d is a fixed positive integer.
- $m(x,y) := \exp(-\|x-y\|^2)$.

Hint: it might be helpful to consider d-fold iterated tensor products of the form $x^{\otimes d} = x \otimes x \otimes \cdots \otimes x$, along with their corresponding inner products.

2.3. **Optional: Proof of Mercer's Theorem.**

Proof of Mercer's Theorem 2.18. Consider the operator on $L_2(\mu)$ defined by

$$T(f)(x) := \int_{\mathbb{R}^n} f(y) m(x,y) d\mu(y), \qquad \forall\, x \in \mathbb{R}^n, \forall\, f \in L_2(\mu)$$

By the Cauchy-Schwarz inequality and by assumption on m, if $f \in L_2(\mu)$, then $T(f) \in L_2(\mu)$, so that $T\colon L_2(\mu) \to L_2(\mu)$, with $\|T\|^2 \leq \int_{\mathbb{R}^n} |m(x,y)|^2 \, d\mu(x) d\mu(y)$, where $\|T\|$ denotes the operator norm. We will show that T is a compact operator, i.e. $T(B(0,1))$ has compact closure in $L_2(\mu)$ with respect to the norm topology, where $B(0,1) := \{f \in L_2(\mu)\colon \|f\| \leq 1\}$, and $\|f\| := (\int_{\mathbb{R}^n} |f(x)|^2 \, d\mu(x))^{1/2}$. When $f, g \in L_2(\mu)$, we denote $f \otimes g\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as the function $(f \otimes g)(x,y) := f(x)g(y)$. Let $\{\phi_i\}_{i=1}^\infty$ be an orthonormal basis of $L_2(\mu)$. Then $\{\phi_i \otimes \phi_j\}_{i,j=1}^\infty$ is an orthonormal basis of $L_2(\mu \times \mu)$. Since $m \in L_2(\mu \times \mu)$, m can be expressed as a linear combination of this orthonormal basis, in the sense that

$$\lim_{p \to \infty} \left\| m - \sum_{i+j \leq p} \langle m, \phi_i \otimes \phi_j \rangle \phi_i \otimes \phi_j \right\| = 0. \qquad (*)$$

So, for any $p \geq 1$, define the following operator on $L_2(\mu)$

$$T_p(f)(x) := \int_{\mathbb{R}^n} f(y) \Big(\sum_{i+j \leq p} \langle m, \phi_i \otimes \phi_j \rangle (\phi_i \otimes \phi_j)(x,y) \Big) d\mu(y), \qquad \forall\, x \in \mathbb{R}^n, \forall\, f \in L_2(\mu).$$

By $(*)$, T is the limit of T_p in the sense that

$$\lim_{p \to \infty} \sup_{f \in L_2(\mu)\colon \|f\| \leq 1} \|(T(f) - T_p(f))\| = 0. \qquad (**)$$

Also, by its definition, each T_p has finite-dimensional range, hence it is a compact operator. The operator T is compact if: given any norm-bounded sequence $\{f_i\}_{i=1}^\infty \in L_2(\mu)$, the sequence $\{Tf_i\}_{i=1}^\infty$ has a convergent subsequence. So, $(**)$ implies that T itself is compact by the following diagonalization argument. Since T_1 is compact, there is a subsequence $\{f_{i,1}\}_{i=1}^\infty$ of $\{f_i\}_{i=1}^\infty$ such that $\{T_1 f_{i,1}\}_{i=1}^\infty$ converges. Since T_2 is compact, there is a subsequence

$\{f_{i,2}\}_{i=1}^{\infty}$ of $\{f_{i,1}\}_{i=1}^{\infty}$ such that $\{T_2 f_{i,2}\}_{i=1}^{\infty}$ converges. And so on. Then let $g_i := f_{i,i}$ for all $i \geq 1$. Fix $i, j \geq 1$ and then let $p \geq 1$.

$$\|T(g_i) - T(g_j)\| \leq \|T(g_i) - T_p(g_i)\| + \|T_p(g_i) - T_p(g_j)\| + \|T_p(g_j) - T(g_j)\|.$$

The first and third terms become small when p is large by $(**)$. The middle term can be made small by choosing p to be the minimum of i and j. In conclusion for all $\varepsilon > 0$, there exists $q \geq 1$ such that, for all $i, j \geq q$, we have $\|T(g_i) - T(g_j)\| < \varepsilon$. That is, $\{T(g_i)\}_{i=1}^{\infty}$ is a Cauchy sequence, hence convergent.

So, T itself is a compact operator, and it is positive semidefinite by Exercise 2.19. The Riesz Theory of Compact operators then applies (Theorem 10.38(4),(6).) That is, there exists a sequence of nonnegative eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ and an orthonormal system of eigenfunctions $\{\psi_i\}_{i=1}^{\infty}$ such that

$$Tf = \sum_{i=1}^{\infty} \lambda_i \langle f, \psi_i \rangle \psi_i, \qquad \forall\, f \in L_2(\mu).$$

We now note that m is nonnegative on its diagonal. We show this by contradiction. Suppose $m(x,x) < 0$ for some $x \in \mathbb{R}^n$. Since m is continuous, there exists an open neighborhood U of x such that m is negative on $U \times U$. By assumption, $\mu(U) > 0$. So, if $f := 1_U$, we would have

$$0 \leq \langle Tf, f \rangle = \int_U \int_U m(x,y) d\mu(x) d\mu(y) < 0,$$

a contradiction. For any $p \geq 1$, define

$$S_p(f) := \sum_{i=1}^{p} \lambda_i \langle f, \psi_i \rangle \psi_i \qquad \forall\, f \in L_2(\mu).$$

And define the corresponding kernel

$$m_p(x,y) := \sum_{i=1}^{p} \lambda_i \psi_i(x) \psi_i(y), \qquad \forall\, x, y \in \mathbb{R}^n.$$

Then $T - S_p$ is also a positive semidefinite operator, so the corresponding kernel $m - m_p$ also has a nonnegative diagonal. That is, for all $p \geq 1$,

$$\sum_{i=1}^{p} \lambda_i (\psi_i(x))^2 \leq m(x,x).$$

The series on the left is therefore absolutely convergent as $p \to \infty$. By the Cauchy-Schwarz inequality, the series defining m_p is also absolutely convergent as $p \to \infty$, for all $x, y \in \mathbb{R}^n$. Denote

$$\overline{m}(x,y) := \lim_{p \to \infty} m_p(x,y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y), \qquad \forall\, x, y \in \mathbb{R}^n.$$

It remains to show that $\overline{m} = m$. To see this, note that the operator with kernel $\overline{m}$ has the same eigenvalues and eigenfunctions as T. Therefore T has kernel $\overline{m}$, as desired. □

3. Probably Approximately Correct (PAC) Learning

Learning involves data from the real world, and real world data is noisy. To account for this, Valiant introduced the PAC learning model in 1984.

Problem 3.1 (Supervised PAC Learning Problem). Let A, B be sets. Let $f\colon A \to B$ be an unknown function. The goal of the learning problem is to determine the function f on all of A using a small number of *randomly drawn* values of f on A. Let $0 < \varepsilon, \delta < 1/2$ and let $\mathbb{P}$ be a probability law on A. We are given access to ε, δ and to random samples from $\mathbb{P}$. (The distribution $\mathbb{P}$ may or may not be known, depending on the problem at hand.) If $x \in A$ is a random sample from $\mathbb{P}$, the pair $(x, f(x))$ is also known.

The goal of the problem is the following. With probability at least $1 - \delta$ (with respect to randomly drawn samples, drawn according to $\mathbb{P}$), output a function $g\colon A \to B$ such that

$$\mathbb{P}(f(x) \neq g(x)) < \varepsilon.$$

That is, the hypothesis g is probably (with probability at least $1 - \delta$) approximately (up to error ε) correct.

We say a subset of functions $\mathcal{F} \subseteq \{f\colon A \to B\}$ is **PAC Learnable** if there exists an algorithm such that, for any probability distribution $\mathbb{P}$ on A, for any $0 < \varepsilon, \delta < 1/2$, the algorithm achieves the above goal.

In the case $A = \{-1, 1\}^n$ and $B = \{-1, 1\}$, we say that $\mathcal{F}$ is **efficiently PAC Learnable** if $\mathcal{F}$ is PAC Learnable, and the algorithm has a run time that is polynomial in $n, 1/\varepsilon, 1/\delta$. When every $f \in \mathcal{F}$ has a positive integer "size" associated to it, we require the algorithm to have run time polynomial in $n, 1/\varepsilon, 1/\delta$ and in the size of the f, where $f \in \mathcal{F}$ is the unknown function being learned by the algorithm.

As discussed in Section 2, Problem 3.1 is too general to be tractable for arbitrary $\mathcal{F}$. In the literature, the family of functions $\mathcal{F}$ is often called a **concept class**.

We include the number of samples of $\mathbb{P}$ in the run time of a PAC learning algorithm. So, being efficiently PAC Learnable requires not taking too many samples from $\mathbb{P}$.

One criticism of PAC learning is that it cannot account for adversarial noise. That is, samples are assumed to independent and identically distributed, so any noise in the data is modeled as i.i.d. noise. On the other hand, one could imagine an adversary corrupts some fraction of the samples arbitrarily. Such noise would be outside the scope of the PAC learning model. For this reason, some contemporary investigations in machine learning have investigated more general adversarial noise models.

3.1. Learning Boolean Conjunctions.

Our first example of a class of PAC learnable functions is boolean conjunctions. For convenience, we use $\{0, 1\}$ instead of $\{-1, 1\}$ in this subsection.

Definition 3.2 (Boolean Conjunctions). Let $I, J \subseteq \{1, \ldots, n\}$. A **boolean conjunction** is a function $f\colon \{0, 1\}^n \to \{0, 1\}$ of the form

$$f(x_1, \ldots, x_n) = \prod_{i \in I} x_i \prod_{j \in J}(1 - x_j), \qquad \forall\, (x_1, \ldots, x_n) \in \{0, 1\}^n.$$

Example 3.3. The function f where $f = 1$ only when $x_1 = 1, x_3 = 0$ and $x_4 = 1$ can be written as

$$f(x_1, \ldots, x_n) := x_1(1 - x_3)x_4, \qquad \forall\, (x_1, \ldots, x_n) \in \{-1, 1\}^n.$$

When f is a boolean conjunction, the set $\{x \in \{0,1\}^n \colon f(x) = 1\}$ can be understood geometrically as an intersection of coordinate halfspaces. Alternatively, if we think of 0 as logical "false" and 1 as "true," a conjunction f is then a combination of logical AND operations on a set of variables together with their negations.

Theorem 3.4. *The class of boolean conjunctions is efficiently PAC learnable.*

Proof. Let f be an unknown boolean conjunction of the form $f(x_1, \ldots, x_n) = \prod_{i \in I} x_i \prod_{j \in J}(1-x_j)$, for some $I, J \subseteq \{1, \ldots, n\}$. We first "guess" a hypothesis g where $I = J = \{1, \ldots, n\}$, so that

$$g(x_1, \ldots, x_n) = \prod_{i=1}^{n} x_i(1 - x_i).$$

Initially, $g = 0$ on $\{-1,1\}^n$. Suppose we sample $y = (y_1, \ldots, y_n) \in \{-1,1\}^n$ from $\mathbb{P}$, and we find that $f(y) = 1$. In such a case, we update g to agree with f. That is, for each $1 \leq i \leq n$, if $y_i = 0$, delete the x_i term from g, and if $y_i = 1$, delete the $(1 - x_i)$ term from g. We repeat this procedure. For any $k \geq 1$, let g_k denote the hypothesis g_k after k iterations of the algorithm, and let $g_0 := g$. Note that each deleted term does not appear in the formula for f (e.g. if $y_i = 0$ and $f(y) = 1$, then y_i cannot appear in the formula of f). So, for any $k \geq 1$, any term in f will be contained in g_k. Let us then try to find the probability that there is a term in g_k not appearing in f.

Let $1 \leq i \leq n$, and consider a term of the form $z := x_i$ or $z := 1 - x_i$. Suppose z appears in g_k but not in f. The term z will be deleted from g_k only when we have a sample where $z = 0$ and $f = 1$. So, the probability that term z is deleted from g_k in one iteration of the algorithm is

$$p(z) := \mathbb{P}(f(y) = 1 \text{ and } z = 0).$$

For any $k \geq 1$, define

$$\varepsilon_k := \mathbb{P}(f(y) \neq g_k(y)).$$

Since $f \neq g_k$ only when $f = 1$, the union bound gives

$$\begin{aligned}\varepsilon_k &= \mathbb{P}(f(y) = 1 \text{ and } g_k(y) = 0)\\ &= \mathbb{P}(f(y) = 1 \text{ and some term in } g_k \text{ is zero}) \leq \sum_{\text{terms } z \text{ in } g_k} p(z). \qquad (*)\end{aligned}$$

Let $\varepsilon > 0$. Suppose $p(z) < \varepsilon/(2n)$ for all terms z in g_k. Since g_k has at most $2n$ terms, we would then have $\varepsilon_k \leq 2n\varepsilon/(2n) = \varepsilon$. Let C_k be the event that $\exists$ a term z in g_k with $p(z) > \varepsilon/(2n)$ and such that z does not appear in f. It remains to bound $\mathbb{P}(C_k)$.

Let z be a term in g_k that is not in f. The probability that z appears in g_k is at most $(1 - p(z))^k$, since each iteration of the algorithm gives us an independent sample from $\mathbb{P}$. Then by the union bound (since at most $2n$ terms appear in g_k)

$$\mathbb{P}(C_k) \leq 2n\Big(1 - \frac{\varepsilon}{2n}\Big)^k. \qquad (**)$$

So, if $\delta > 0$, we choose k such that

$$2n\Big(1 - \frac{\varepsilon}{2n}\Big)^k \leq \delta.$$

Using $1-t\leq e^{-t}$ for all $t\in[0,1]$, we choose k so that $2ne^{-\varepsilon k/(2n)}\leq\delta$, i.e. $-\varepsilon k\leq 2n\log(\delta/(2n))$, i.e.

$$k\geq\frac{2n}{\varepsilon}\log(2n/\delta).$$

In summary, if the algorithm uses more than $\frac{2n}{\varepsilon}\log(2n/\delta)$ samples, then with probability at least $1-\delta$ by $(**)$, the hypothesis g_k of the algorithm satisfies $\mathbb{P}(f\neq g_k)\leq\varepsilon$ by $(*)$. □

Exercise 3.5. Show that the set of conjunctions is contained in the set of linear threshold functions. That is, given a boolean conjunction $f\colon\{0,1\}^n\to\{0,1\}$, find $w\in\mathbb{R}^n, t\in\mathbb{R}$ such that

$$f(x)=1_{\{\langle w,x\rangle>t\}},\qquad\forall\,x=(x_1,\ldots,x_n)\in\{0,1\}^n.$$

3.2. **Learning DNF Formulas.** Surprisingly, the following slightly larger class of natural functions that contains boolean conjunctions is not efficiently PAC learnable, in its own function class.

Definition 3.6 (3-Term DNF Formula). Let $f_1,f_2,f_3\colon\{0,1\}^n\to\{0,1\}$ be boolean conjunctions. A **3-Term DNF Formula** is a function $f\colon\{0,1\}^n\to\{0,1\}$ of the form

$$f(x)=\max(f_1(x),f_2(x),f_3(x)),\qquad\forall\,x\in\{-1,1\}^n.$$

When f is a 3-term DNF formula, the set $\{x\in\{0,1\}^n\colon f(x)=1\}$ can be understood geometrically as a union of 3 conjunctions (which are themselves intersections of coordinate half spaces). Alternatively, if we think of 0 as logical "false" and 1 as "true," f is then a 3-term logical OR operation applied to AND operations, on a set of variables together with their negations.

Theorem 3.7. *If* **RP** $\neq$ **NP***, then the class of* 3*-term DNF formulae is not efficiently PAC learnable, in its own function class.*

Remark 3.8. **P** $\subseteq$ **RP** $\subseteq$ **NP**, so it could technically occur that **P** $\neq$ **NP** while **RP** $=$ **NP**. However it is widely believed that **P** $=$ **RP**.

The above Theorem in [KV94, Section 1.4] says that learning 3-term DNF formulae is not possible to do efficiently, in the class of 3-term DNF formula. However, it is still possible to learn 3-term DNF formulae, when we consider them in a larger class of functions. In fact, this is possible, as we now outline.

Let $a,b,c,d\in\{0,1\}$. Using the distributive property

$$\max(ab,cd)=\max(a,c)\max(a,d)\max(b,c)\max(b,d),$$

we can rewrite any 3-term DNF formula as a 3-term CNF formula:

$$\max(f_1(x),f_2(x),f_3(x))=\prod_{\substack{\text{terms } z_1 \text{ in } f_1\\ \text{terms } z_2 \text{ in } f_2\\ \text{terms } z_3 \text{ in } f_3}}\max(z_1,z_2,z_3).$$

(A 3-term CNF formula is a function on n variables of the form on the right.) Each term in this large product is treated as its own variable (among $(2n)^3$ possible variables), so that the formula on the right is then interpreted as a conjunction on $8n^3$ variables. We then apply the algorithm of Theorem 3.4 to learn the conjunction, and then rewrite the conjunction as a 3-term CNF formula. For more details see [KV94, Section 1.5]. This argument implies the following two things:

Theorem 3.9. *The class of* 3*-term CNF formulae is efficiently PAC learnable.*

Corollary 3.10. *The class of* 3*-term DNF formulae is efficiently PAC learnable when represented as* 3*-term CNF formulae. That is, the DNF function class is considered as a subset of the larger class of* 3*-term CNF formulae.*

This corollary is loosely analogous to the discussion of the Minimum Vertex Cover Problem after 1.11. Even though it might be hard to optimize a function on a small domain, enlarging the domain can make it easier to optimize the function (while changing our interpretation of the meaning of the optimum).

Remark 3.11. In the PAC learning framework, we always assume that the hypothesis class $\mathcal{F}$ is **polynomially evaluatable**. That is, there is an algorithm such that, for any $f \in \mathcal{F}$ and for any $x \in \{0,1\}^n$, the value $f(x)$ is output in time polynomial in n and in the size of f.

It turns out the above argument can be generalized, so that, for any $k \geq 2$, k-term DNF formulae are efficiently PAC learnable within the class of k-CNF formulae. Though, as we can imagine, the number of variables involved in passing between DNF and CNF formula seems to be $(2n)^k$. So, if k itself depends on n (e.g. k is a polynomial in n), then the above argument no longer gives an efficient algorithm.

The following is a well-known open problem since the introduction of PAC learning in 1984: are DNF formula of polynomial size (i.e. with $n^{O(1)}$ terms) PAC learnable (within some larger function class)?

It is possible to PAC learn linear threshold functions, but we will postpone this discussion to Section 4. The story for more general learning models of halfspaces is more complicated. In the **agnostic learning model**, the learner has access to a random sample from $A \times B$, i.e. there is no a priori functional relationship that is assumed between A and B. The goal is then to find a $g \in \mathcal{F}$ minimizing $\mathbb{P}(g(X) \neq Y)$, where (X,Y) is the random sample from $A \times B$. (We can think intuitively think of this model as having a target function f with a fraction of its labels randomly corrupted.) In this setting, which is slightly more general than PAC learning, the agnostic learning of linear threshold functions is possible under some assumptions on the distribution $\mathbb{P}$, when the functions are considered in the larger class of polynomial threshold functions [KKMS08]. However, agnostic learning of linear threshold functions in the class of linear threshold functions is hard [FGRW12]. Moreover, (recalling Exercise 3.5), agnostic learning of conjunctions in the class of linear threshold functions is hard [FGRW12]. With access to Gaussian samples, where a fraction of both samples and labels can be arbitrarily corrupted (the so-called nasty noise model), linear threshold functions can be learned efficiently [DKS18].

Remark 3.12. When a PAC learning algorithm for a class of functions $\mathcal{F}$ outputs a function in the class $\mathcal{F}$, the algorithm is called **proper**. Otherwise, the algorithm is called **improper**.

3.3. **Boosting.** In this section, we investigate a modification of the PAC learning model, where with high probability the algorithm is only guaranteed to classify slightly more than 50% of examples correctly. The main question is: can we somehow "boost" the performance of a "weak" learning algorithm like this to get "stronger" PAC learning algorithm.

Problem 3.13 (**Supervised Weak Learning Problem**)**.** Let A, B be sets. Let $f\colon A \to B$ be an unknown function. The goal of the learning problem is to determine the function f on all of A using a small number of *randomly drawn* values of f on A. Let $0 < \varepsilon, \delta < 1/2$ and let $\mathbb{P}$ be a probability law on A. We are given access to ε, δ and to random samples from $\mathbb{P}$. (The distribution $\mathbb{P}$ may or may not be known, depending on the problem at hand.) If $x \in A$ is a random sample from $\mathbb{P}$, the pair $(x, f(x))$ is also known.

The goal of the problem is the following. With probability at least $1 - \delta$ (with respect to randomly drawn samples, drawn according to $\mathbb{P}$), output a function $g\colon A \to B$ such that

$$\mathbb{P}(x \in A\colon f(x) \neq g(x)) < \frac{1}{2} - \varepsilon.$$

That is, the hypothesis g is probably (with probability at least $1 - \delta$) slightly better than random guessing (since a random assignment g would have probability of mis-classification at most $1/2$). We emphasize that the function g can depend on the random sample.

We say a subset of functions $\mathcal{F} \subseteq \{f\colon A \to B\}$ is ε-**weak learnable** if there exists $\varepsilon > 0$ and an algorithm such that, for any probability distribution $\mathbb{P}$ on A, there exists $\varepsilon > 0$ such that, for any $0 < \delta < 1/2$, the algorithm achieves the above goal.

In the case $A = \{-1, 1\}^n$ and $B = \{-1, 1\}$, we say that $\mathcal{F}$ is **efficiently weak learnable** if $\mathcal{F}$ is weak learnable, and the algorithm has a run time that is polynomial in $n, 1/\varepsilon$ and $1/\delta$.

In the literature, the family of functions $\mathcal{F}$ is often called a **concept class**.

As noted in Section 2.1.1, given that there is some half space that correctly classifies 99% of data points with $B = \{-1, 1\}$, it is hard to find a half space that correctly classifies 51% of data points [FGKP06]. Note that correctly classifying 50% of data points is easy by just choosing any half space and deciding which side to label $+1$. However, this result is not directly relevant in the PAC model, since if we are finding an approximation to a target function (such as a linear threshold function), then it is assumed that all points can be correctly classified. In the case that B has $k > 2$ values, random guessing only correctly classifies $1/k$ sample points, so an ε-weak learning algorithm with ε small is a rather strong assumption when compared to the case $k = 2$.

Suppose that we can weakly learn some concept class. Using our intuition from e.g. the Law of Large Numbers, we know that if we can double our money by betting on a game that has a 51% chance of success, then we can earn money with high probability by betting on independent repetitions of the game. Similarly, we can try to sample the weak learner many times and then take a majority vote of the output, thereby "boosting" the weak learner to a much "stronger" (PAC) learning [Sch90]. For an exposition of this argument, see [KV94, Chapter 4], where a nested sequence of majority votes is used, rather than a single majority function. Here we instead focus on a simpler and more practical boosting procedure, albeit without an explicit guarantee of PAC learning.

In this proof, we denote

$$\Delta_k := \{v \in \mathbb{R}^k\colon \sum_{i=1}^{k} v_i = 1,\, v_i \geq 0,\, \forall\, 1 \leq i \leq k\}.$$

Given the samples $x^{(1)}, \ldots, x^{(k)} \in A$, labels $y_1, \ldots, y_k \in \{-1, 1\}$, and given $v \in \Delta_k$, consider the probability law $\mathbb{P}_v$ on A defined so that $\mathbb{P}_v(x^{(i)}) = v_i$ for all $1 \leq i \leq k$.

Algorithm 3.14 (**Adaptive Boosting (AdaBoost)** [FS97]). The input is $0 < \varepsilon < 1/2$, an ε-weak learning algorithm, a number of iterations t, samples $x^{(1)}, \ldots, x^{(k)} \in A$, and labels $y_1, \ldots, y_k \in \{-1, 1\}$. Initialize $v^{(1)} \in \Delta_k$, with $v^{(1)} := \frac{1}{t}(1, \ldots, 1)$.

For each $1 \leq s \leq t$, do the following.

- Using the ε-weak learning algorithm on input $\mathbb{P}_{v^{(s)}}$, get hypothesis $g_s \colon A \to \{-1, 1\}$. (We assume the random samples given to the weak learning algorithm are independent of each other for all $1 \leq s \leq t$.)
- Let $\gamma_s := \sum_{i=1}^{k} v_i^{(s)} 1_{\{g_s(x^{(i)}) \neq y_i\}}$. (This is the $\mathbb{P}_{v^{(s)}}$ probability that the weak learner mis-classifies its input, so it is at most $1/2 - \varepsilon$ with $\mathbb{P}_{v^{(s)}}$ probability close to 1. Similarly, the following quantity is typically less than one.)
- Let $\beta_s := \frac{\gamma_s}{1-\gamma_s}$. (We may assume $\gamma_s > 0$ since if $\gamma_s = 0$, the learner makes no mistakes, so that $\gamma_1 = 0$, so we do not need to apply the boosting algorithm at all.) (So, β_s is the ratio of the $\mathbb{P}_{v^{(s)}}$ probability of correct classification and the probability of mis-classification.)
- Define $v_i^{(s+1)} := \dfrac{v_i^{(s)} \beta_s^{[1_{\{g_s(x^{(i)})=y_i\}}]}}{\sum_{j=1}^{k} v_j^{(s)} \beta_s^{[1_{\{g_s(x^{(j)})=y_j\}}]}}$, $\forall\, 1 \leq i \leq k$.

Output the hypothesis g that is the "weighted majority vote" of each round of hypothesis:

$$g(x) := \mathrm{sign}\Big(\sum_{s=1}^{t} \log\Big(\frac{1}{\beta_s}\Big) g_s(x) \Big).$$

If the weak learner makes few mistakes, then γ_s is close to 0, as is β_s, so $\log(1/\beta_s)$ is large. So the "weight" of each "voter" is directly related to its number of correct classifications. Also, if the weak learner makes many mistakes, then γ_s is close to 1, β_s is much larger than 1, so $\log(1/\beta_s)$ is negative, i.e. g makes the opposite recommendation of g_s. That is, the boosting algorithm still gains something when the weak learner is very wrong, and this is reflected in the bound below.

Theorem 3.15. *Let g be the output of Algorithm 3.14. Define $\varepsilon_s := \frac{1}{2} - \gamma_s$ for all $1 \leq s \leq t$. Then the average number of mis-classifications of g satisfies*

$$\frac{1}{k} \left|\{1 \leq i \leq k \colon g(x^{(i)}) \neq y_i\}\right| \leq e^{-2\sum_{s=1}^{t} \varepsilon_s^2}.$$

Exercise 3.16. Explain why taking the expected value of this inequality does not guarantee PAC learning.

Proof. For any $\beta \geq 0$ and for any $a \in [0, 1]$, we have $\beta^a \leq 1 - (1 - \beta)a$. (To see this, note that the second derivative in a is positive on the left, so that $a \mapsto \beta^a$ is convex, and both quantities agree when $a = 0$ and when $a = 1$. So convexity implies the inequality.) Now, let

$$z_s := \sum_{j=1}^{k} v_j^{(s)} \beta_s^{[1_{\{g_s(x^{(j)})=y_j\}}]}, \qquad \forall\, 1 \leq s \leq t.$$

Using this inequality and the definition of γ_s we have

$$z_s \leq \sum_{j=1}^{k} v_j^{(s)} [1 - (1 - \beta_s) 1_{\{g_s(x^{(j)})=y_j\}} = 1 - (1 - \beta_s)(1 - \gamma_s). \qquad (\ddagger)$$

By induction, note that, for all $1 \leq i \leq k$,

$$v_i^{(t+1)} = v_i^{(1)} \prod_{s=1}^{t} \frac{\beta_s^{[1_{\{g_s(x^{(i)})=y_i\}}]}}{z_s}. \qquad (*)$$

If the output g mis-classifies $x^{(i)}$ for some $1 \leq i \leq k$, then

$$\begin{aligned} y_i \sum_{s=1}^{t} \log\Big(\frac{1}{\beta_s}\Big) g_s(x^{(i)}) \leq 0 \qquad &\Leftrightarrow \qquad \prod_{s=1}^{t} \beta_s^{-y_i g_s(x^{(i)})} \leq 1 \\ &\Leftrightarrow \qquad \prod_{s=1}^{t} \beta_s^{[1-2\cdot 1_{\{g_s(x^{(i)})=y_i\}}]} \leq 1 \\ &\Leftrightarrow \qquad \prod_{s=1}^{t} \beta_s^{[1_{\{g_s(x^{(i)})=y_i\}}]} \geq \Big(\prod_{s=1}^{t} \beta_s\Big)^{1/2}. \qquad (**) \end{aligned}$$

Using that $w^{(t+1)} \in \Delta_k$,

$$\begin{aligned} \prod_{s=1}^{t} z_s &= \sum_{i=1}^{k} w_i^{(t+1)} \prod_{s=1}^{t} z_s \overset{(*)}{=} \sum_{i=1}^{k} w_i^{(1)} \prod_{s=1}^{t} \beta_s^{[1_{\{g_s(x^{(i)})=y_i\}}]} = \frac{1}{k} \sum_{i=1}^{k} \prod_{s=1}^{t} \beta_s^{[1_{\{g_s(x^{(i)})=y_i\}}]} \\ &\geq \frac{1}{k} \sum_{i\colon\, g(x^{(i)})\neq y_i} \prod_{s=1}^{t} \beta_s^{[1_{\{g_s(x^{(i)})=y_i\}}]} \overset{(**)}{\geq} \frac{1}{k} \sum_{i\colon\, g(x^{(i)})\neq y_i} \Big(\prod_{s=1}^{t} \beta_s\Big)^{1/2} \\ &= \frac{1}{k} \left|\{1 \leq i \leq k \colon g(x^{(i)}) \neq y_i\}\right| \Big(\prod_{s=1}^{t} \beta_s\Big)^{1/2}. \end{aligned}$$

That is,

$$\frac{1}{k} \left|\{1 \leq i \leq k \colon g(x^{(i)}) \neq y_i\}\right| \leq \prod_{s=1}^{t} \frac{z_s}{\sqrt{\beta_s}} \overset{(\ddagger)}{\leq} \prod_{s=1}^{t} \frac{[1-(1-\beta_s)(1-\gamma_s)]}{\sqrt{\beta_s}}.$$

Substituting $\beta_s = \gamma_s/[1-\gamma_s]$ (which happens to minimize the right side over all such choices of β_s) concludes the proof since the s^{th} term is then

$$\beta_s^{-1/2}[1-(1-2\gamma_s)] = \beta_s^{-1/2} 2\gamma_s = 2\sqrt{\gamma_s(1-\gamma_s)} = 2\sqrt{((1/2)-\varepsilon_s)((1/2)+\varepsilon_s)} = \sqrt{1-4\varepsilon_s^2}.$$

The final inequality follows since $1-a \leq e^{-a}$ for all $a \in \mathbb{R}$. □

Remark 3.17. At each iteration of Algorithm 3.14, the weak learner fails with probability at most δ. So,the weak learner succeeds in all t iterations with probability at least $1-\delta t$, by the union bound.

In Theorem 3.15, the quantity

$$\frac{1}{k} \left|\{1 \leq i \leq k \colon g(x^{(i)}) \neq y_i\}\right| = \frac{1}{k} \sum_{i=1}^{k} 1_{\{g(x^{(i)})\neq y_i\}}$$

is called the **empirical risk**, since it gives the average error on the given sample. We would ideally like to bound $\mathbb{P}(g(X) \neq f(X))$, but since we do not have access directly to the random variable X that generates the samples $x^{(1)},\ldots,x^{(k)}$, we cannot compute $\mathbb{P}(g(X)\neq f(X))$ exactly. We will discuss this issue in more detail In Section 6.

In the case that $f\colon A\to B$ is the unknown function with B finite and $|B|>2$, there are a few ways we can modify Algorithm 3.14; here is one way.

Algorithm 3.18 (**Multi-Class Adaptive Boosting (AdaBoost.M1)** [FS97])**.** The input is $0<\varepsilon<1/2$, an ε-weak learning algorithm, a number of iterations t, samples $x^{(1)},\ldots,x^{(k)}\in A$, and labels $y_1,\ldots,y_k\in\{1,\ldots,p\}$. Initialize $v^{(1)}\in\Delta_k$, with $v^{(1)}:=\frac{1}{t}(1,\ldots,1)$.

For each $1\leq s\leq t$, do the following.

- Using the ε-weak learning algorithm on input $\mathbb{P}_{v^{(s)}}$, get hypothesis $g_s\colon A\to\{1,\ldots,p\}$. (We assume the random samples given to the weak learning algorithm are independent of each other for all $1\leq s\leq t$.)
- Let $\gamma_s:=\sum_{i=1}^k v_i^{(s)}1_{\{g_s(x^{(i)})\neq y_i\}}$.
- Let $\beta_s:=\frac{\gamma_s}{1-\gamma_s}$.
- Define $v_i^{(s+1)}:=\dfrac{v_i^{(s)}\beta_s^{[1_{\{g_s(x^{(i)})=y_i\}}]}}{\sum_{j=1}^k v_j^{(s)}\beta_s^{[1_{\{g_s(x^{(j)})=y_j\}}]}}$, $\forall\, 1\leq i\leq k$.

Output the hypothesis g that is the "weighted plurality vote" of each round of hypothesis:

$$g(x):=\mathrm{argmax}_{y\in\{1,\ldots,p\}}\sum_{s=1}^t\log\left(\frac{1}{\beta_s}\right)1_{\{g_s(x)=y\}}.$$

3.4. **Occam's Razor.** Let A be a set, let $\mathcal{F}=\cup_{n=1}^\infty\mathcal{F}_n$ be a set of functions from A to B, and let $\mathcal{G}=\cup_{n=1}^\infty\mathcal{G}_n$ be a set of functions from A to B.

Let $f\in\mathcal{F}$. A sample S with cardinality $m>0$ labelled according to f is a set of ordered pairs

$$\{(x^{(1)},f(x^{(1)}),\ldots,(x^{(m)},f(x^{(m)})\colon x^{(i)}\in A,\,\forall 1\leq i\leq m\}.$$

We assume in this section that the size of $g\in\mathcal{G}$ and $f\in\mathcal{F}$, denoted $\mathrm{size}(\cdot)$, is well-defined to be the bit-length of an encoding of these functions (e.g. $\exists$ a function $\phi_{\mathcal{F}}\colon\mathcal{F}\to\{0,1\}^{\mathbb{N}}$ such that $[\phi_{\mathcal{F}}(f)]_{\mathrm{size}(f)+k}=0$ for all $k\geq 1$. That is, the binary encoding of f only uses the coordinates from 1 to $\mathrm{size}(f)$.) We say that $g\in\mathcal{G}$ is **consistent** with $f\in\mathcal{F}$ on S if $f(x^{(i)})=g(x^{(i)})$ for all $1\leq i\leq m$.

Definition 3.19 (**Occam Algorithm**)**.** Let $\alpha\geq 0$ and let $0\leq\beta<1$ be constants. We say that L is an (α,β)-**Occam algorithm for $\mathcal{F}$ using $\mathcal{G}$** if, for $n\geq 1$ and for any input sample S of cardinality m labelled according to $f\in\mathcal{F}_n$, L outputs a hypothesis $g\in\mathcal{G}$ such that

- g is consistent with f, and
- $\mathrm{size}(g)\leq(n\cdot\mathrm{size}(f))^\alpha m^\beta$.

We say that L is an **efficient** (α,β)-**Occam algorithm** if its run time is at most a polynomial in n,m and $\mathrm{size}(f)$.

Theorem 3.20 (**Occam's Razor**)**.** *Let L be an efficient (α,β)-Occam algorithm for $\mathcal{F}$ using $\mathcal{G}$. Let $0<\varepsilon,\delta<1$, let $f\in\mathcal{F}_n$, and let $\mathbb{P}$ be a probability law on A. Then there is a constant $a>0$ such that, if L is given as input a random sample S of m examples labelled* *according to f (and $\mathbb{P}$), with*

$$m \geq a\Big(\frac{1}{\varepsilon}\log(1/\delta) + \Big(\frac{(n\cdot \mathrm{size}(f))^{\alpha}}{\varepsilon}\Big)^{\frac{1}{1-\beta}}\Big).$$

Then with probability at least $1-\delta$, the output g of L satisfies $\mathbb{P}(f(x)\neq g(x))<\varepsilon$, and the run time of L is at most polynomial in n, $\mathrm{size}(f)$, $1/\varepsilon$ and $1/\delta$.

In the following result, we assume that $\mathcal{G}_n = \cup_{m=1}^{\infty}\mathcal{G}_{n,m}$, and when the algorithm L has input a sample S of cardinality m and $f\in\mathcal{F}_n$, the output is $g\in\mathcal{G}_{n,m}$.

Theorem 3.21 (**Occam's Razor, Cardinality Version**). *Let L be an algorithm such that, for any $n\geq 1$ and for any $f\in\mathcal{F}_n$, if L is given as input a random sample S of m examples labelled according to f, then L runs in time at most polynomial in n, m and $\mathrm{size}(f)$, and outputs $g\in\mathcal{G}_{n,m}$ that is consistent with f on S. Then there is a constant $b>0$ such that, for any n, for any probability law $\mathbb{P}$ on A, and for any $f\in\mathcal{F}_n$, if L is given as input a random sample of m examples drawn according to f (and $\mathbb{P}$), where $|\mathcal{G}_{n,m}|$ satisfies*

$$\log|\mathcal{G}_{n,m}| \leq b\varepsilon m - \log(1/\delta),$$

(or equivalently m satisfies $m\geq (1/(b\varepsilon))(\log|\mathcal{G}_{n,m}|+\log(1/\delta))$), then L is guaranteed to find $g\in\mathcal{G}_{n,m}$ that with probability at least $1-\delta$ satisfies $\mathbb{P}(f(x)\neq g(x))\leq\varepsilon$.

Proof. We say $g\in\mathcal{G}$ is *bad* if $\mathbb{P}(f(x)\neq g(x))>\varepsilon$. Since the random samples of S are each independent, the probability that a fixed bad hypothesis g is consistent with f on a random sample of cardinality m is at most $(1-\varepsilon)^m$. Let $\mathcal{G}'\subseteq\mathcal{G}_{n,m}$ be the subset of all bad hypotheses in $\mathcal{G}_{n,m}$. By the union bound, the probability that there exists $g\in\mathcal{G}'$ that is consistent with f on a random sample of cardinality m is at most $|\mathcal{G}'|(1-\varepsilon)^m$. We want this quantity to be at most δ. Since $\mathcal{G}'\subseteq\mathcal{G}_{n,m}$, it suffices to show that $|\mathcal{G}_{n,m}|(1-\varepsilon)^m\leq\delta$. Taking logarithms and using $\log(1/(1-\varepsilon))=\Theta(\varepsilon)$ concludes the proof. □

Proof of Theorem 3.20. Let $\mathcal{G}_{n,m}$ be the set of hypotheses that can be output when the input $f\in\mathcal{F}_n$ has cardinality m. Since L is an (α,β)-Occam algorithm, every such hypothesis g has bit-length at most $\mathrm{size}(g)\leq(n\cdot\mathrm{size}(f))^{\alpha}m^{\beta}$, so that $|\mathcal{G}_{n,m}|\leq 2^{\mathrm{size}(g)}\leq 2^{(n\cdot\mathrm{size}(f))^{\alpha}m^{\beta}}$. By Theorem 3.21, the output g of L satisfies $\mathbb{P}(f\neq g)<\varepsilon$ with probability at least $1-\delta$, if

$$\log|\mathcal{G}_{n,m}|\leq b\varepsilon m-\log(1/\delta).$$

That is, we want m to satisfy

$$m\geq b^{-1}\varepsilon^{-1}\log|\mathcal{G}_{n,m}|+b^{-1}\varepsilon^{-1}\log(1/\delta).$$

i.e. it suffices to choose m so that $m\geq 2b^{-1}\varepsilon^{-1}\max(\log|\mathcal{G}_{n,m}|,\log(1/\delta))$. Choosing $a=2/b$ concludes the proof. □

In retrospect, the algorithm for learning conjunctions used in Theorem 3.4 was an Occam algorithm. In fact, the bounds from Theorem 3.20 improve upon those in Theorem 3.4.

3.5. Additional Comments.

A slightly different version of the Perceptron Algorithm can be stated as follows

Algorithm 3.22 (**Perceptron Algorithm, Version 2**).

- Define $w^{(1)}:=0\in\mathbb{R}^n$ and let $s:=1$

- If there exists some $1 \leq i \leq k$ such that $y_i \neq \mathrm{sign}(\langle w^{(s)}, x^{(i)} \rangle)$, i.e. a mis-classification occurs, define

$$w^{(s+1)} := w^{(s)} + y_i \frac{x^{(i)}}{\|x^{(i)}\|}.$$

- Increase the value of s by one. Repeat the previous step until no such i exists.
- Output $w := w^{(s)}$.

We can then state a version of Theorem 2.9 for Algorithm 3.22.

Theorem 3.23 (Perceptron Algorithm Run Time, Version 2). *Let $x^{(1)}, \ldots, x^{(k)} \in \mathbb{R}^n$ and let $y_1, \ldots, y_k \in \{-1, 1\}$ be given. Assume that there exists $w \in \mathbb{R}^n$ such that*

$$\mathrm{sign}(\langle w, x^{(i)} \rangle) = y_i, \qquad \forall\, 1 \leq i \leq k.$$

Assume $\|w\| = 1$. Define

$$\sigma := \min_{1 \leq i \leq k} \left| \langle w, \frac{x^{(i)}}{\|x^{(i)}\|} \rangle \right|.$$

Then the Perceptron Algorithm 3.22 terminates with a value of s satisfying

$$s \leq \sigma^{-2}.$$

The Winnow Algorithm is similar to the Perceptron Algorithm, though the data is usually assumed to be vectors from $\{0, 1\}^n$, and multiplicative updates are made to the weight vector w, rather than additive updates.

In our analysis of the Perceptron Algorithm 2.8, we defined

$$\beta := \max_{i=1,\ldots,k} \left\| x^{(i)} \right\|, \qquad \theta := \min \Big\{ \|w\| : \forall\, 1 \leq i \leq k,\ y_i \langle w, x^{(i)} \rangle \geq 1 \Big\}.$$

We remarked after Lemma 2.11 that the margin $1/\theta$ measures how wide a symmetric "slab" through the origin can separate the vectors $x^{(1)}, \ldots, x^{(k)}$ into their two classes. The **support vector machine** is another linear classifier that takes this analysis into account as follows. Let $\lambda > 0$ and suppose we want to find the $w \in \mathbb{R}^n$ and $z_1, \ldots, z_k \in \mathbb{R}$ minimizing

$$\lambda \|w\|^2 + \frac{1}{k} \sum_{i=1}^{k} z_i.$$

subject to the linear constraints

$$y_i \langle w, x^{(i)} \rangle \geq 1 - z_i, \qquad z_i \geq 0, \quad \forall\, 1 \leq i \leq k.$$

This is a quadratic minimization problem subject to linear constraints.

One advantage of the support vector machine over the perceptron is that the support vector machine is still well-defined even when a hyperplane cannot correctly classify all points, whereas the usual perceptron algorithm will never terminate when presented with such a set of points.

4. Vapnis-Chervonenkis (VC) Theory

Definition 4.1 (**Metric Space**)**.** Let A be a set and let $d\colon A\times A\to[0,\infty)$. We say that A is a **metric** on A if:

- $d(x,y)\geq 0$ for all $x,y\in A$, and $d(x,y)=0$ only when $x=y$.
- $d(x,y)=d(y,x)$ for all $x,y\in A$.
- $d(x,y)\leq d(x,z)+d(z,y)$ for all $x,y,z\in A$.

When d is a metric on A, we refer to (A,d) as a **metric space**. For any $x\in A$ and for any $\varepsilon>0$, we denote the **open ball with radius ε centered at** x as

$$B(x,\varepsilon)=B_d(x,\varepsilon):=\{y\in A\colon d(x,y)<\varepsilon\}.$$

Definition 4.2 (**ε-net**)**.** Let A be a set and let $d\colon A\times A\to[0,\infty)$ be a metric on A. An **ε-net** is a subset $\{x^{(i)}\}_{i\in I}$ of A such that

$$\bigcup_{i\in I}B(x^{(i)},\varepsilon)=A.$$

Proposition 4.3. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Let $\varepsilon>0$. Then there exists an ε-net $\mathcal{N}$ in the unit ball $B:=\{x\in\mathbb{R}^n\colon\ \|x\|<1\}$ such that*

$$|\mathcal{N}|\leq(1+2/\varepsilon)^n.$$

Proof. Let $\{x^{(i)}\}_{i=1}^k$ be a set of maximal cardinality such that $\left\|x^{(i)}-x^{(j)}\right\|\geq\varepsilon$ for all $1\leq i,j\leq k$ such that $i\neq j$. It follows that $\{x^{(i)}\}_{i=1}^k$ is an ε-net in B (if $B\cap\Big(\cup_{i=1}^k B(x^{(i)},\varepsilon)\Big)\neq B$, then there exists $x\in B$ such that $\left\|x^{(i)}-x\right\|\geq\varepsilon$ for all $1\leq i\leq k$, contradicting the maximal cardinality of $\{x^{(i)}\}_{i=1}^k$.) By definition of $\{x^{(i)}\}_{i=1}^k$, the open balls $\{B(x^{(i)},\varepsilon/2)\}_{i=1}^k$ are disjoint and their union is contained in $B(0,1+\varepsilon/2)$. So, comparing the volumes of these two sets, we have

$$k(\varepsilon/2)^n\leq(1+\varepsilon/2)^n.$$

That is, $k\leq(1+2/\varepsilon)^n$. □

Exercise 4.4. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Let $\varepsilon>0$. Then any ε-net $\mathcal{N}$ in the unit ball $B:=\{x\in\mathbb{R}^n\colon\ \|x\|<1\}$ satisfies

$$(1/\varepsilon)^n\leq|\mathcal{N}|\leq(1+2/\varepsilon)^n.$$

Definition 4.5 (**Covering Number**)**.** Let (A,d) be a metric space and let $\varepsilon>0$. The **ε-covering number** of (A,d), denoted $\mathcal{N}(A,d,\varepsilon)$ is the smallest cardinality of an ε-net in A.

Definition 4.6 (**VC Dimension**)**.** Let A be a set and let $\mathcal{F}\subseteq\{-1,1\}^A$ be a class of Boolean functions on A. We say a set $B\subseteq A$ is **shattered** by $\mathcal{F}$ if, for any $g\colon B\to\{-1,1\}$, there exists $f\in\mathcal{F}$ such that $f(x)=g(x)$ for all $x\in B$. The **VC-dimension** of $\mathcal{F}$, denoted $\mathrm{VCdim}(\mathcal{F})$ is the largest cardinality of a subset that is shattered by $\mathcal{F}$.

Note that by definition of $\mathrm{VCdim}(\mathcal{F})$, we have $|\mathcal{F}|\geq 2^{\mathrm{VCdim}(\mathcal{F})}$. So, $\mathrm{VCdim}(\mathcal{F})$ is somewhat analogous to the log of the cardinality of $\mathcal{F}$. Let $\mathcal{F}$ be the set of all boolean functions $f\colon\{-1,1\}^n\to\{-1,1\}$. Then $|\mathcal{F}|=2^{2^n}$ and $\mathrm{VCdim}(\mathcal{F})=2^n$, so in this case the trivial lower bound on $|\mathcal{F}|$ is sharp. However, this lower bound can be quite far from exact, as we discuss below.

Lemma 4.7 (**Pajor's Lemma**). *Let A be a finite set and let $\mathcal{F}\subseteq\{-1,1\}^A$ be a class of Boolean functions on A. Then*

$$|\mathcal{F}|\leq|\{\emptyset\subseteq A'\subseteq A\colon A' \text{ is shattered by } \mathcal{F}\}|\,.$$

Proof. We induct on the size of A. The case $|A|=1$ is clear (e.g. if $\mathcal{F}=1$, then the right side is also 1 since $\emptyset$ is counted). Suppose then that the Lemma holds when $|A|=n\geq 1$ and consider the case $|A|=n+1$. Write $A=\{a\}\cup A_0$, where $|A_0|=n$. Let $\mathcal{F}_1:=\{f\in\mathcal{F}\colon f(a)=1\}$, $\mathcal{F}_{-1}:=\{f\in\mathcal{F}\colon f(a)=-1\}$. By the inductive hypothesis,

$$|\mathcal{F}_1|\leq|\{\emptyset\subseteq A'\subseteq A_0\colon A' \text{ is shattered by } \mathcal{F}_1|_{A_0}\}|=:c_1.$$

$$|\mathcal{F}_{-1}|\leq|\{\emptyset\subseteq A'\subseteq A_0\colon A' \text{ is shattered by } \mathcal{F}_{-1}|_{A_0}\}|=:c_{-1}.$$

So, it remains to show that

$$c_1+c_{-1}\leq c_0:=|\{\emptyset\subseteq A'\subseteq A\colon A' \text{ is shattered by } \mathcal{F}\}|\,.$$

If a set $A'\subseteq A_0$ is shattered by $\mathcal{F}_1|_{A_0}$, then A' is also shattered by $\mathcal{F}$, since $\mathcal{F}_1\subseteq\mathcal{F}$. The same goes for $\mathcal{F}_{-1}$. Suppose now that $A'\subseteq A_0$ is in the intersection of the sets defined by c_1 and c_{-1}. Then c_1+c_{-1} will count the set A' twice, while c_0 only counts it once. However, the set $\{a\}\cup A'$ shattered by $\mathcal{F}$, since A' is in both sets corresponding to c_1 and c_{-1}, so this set is also counted by c_0 (but not by c_1 or c_{-1}). The inequality $c_1+c_{-1}\leq c_0$ follows. □

Lemma 4.8 (**Sauer-Shelah**). *Let A be a finite set with $|A|=n$, and let $\mathcal{F}\subseteq\{-1,1\}^A$ be a class of Boolean functions on A. Let $d:=\mathrm{VCdim}(\mathcal{F})$. Then*

$$|\mathcal{F}|\leq\sum_{i=0}^{d}\binom{n}{i}\leq(en/d)^d.$$

Proof. By Pajor's Lemma 4.7,

$$|\mathcal{F}|\leq|\{\emptyset\subseteq A'\subseteq A\colon A' \text{ is shattered by } \mathcal{F}\}|\,.$$

By the definition of $\mathrm{VCdim}(\mathcal{F})$, each set counted on the right has cardinality at most d. So,

$$|\mathcal{F}|\leq|\{\emptyset\subseteq A'\subseteq A\colon\ |A'|\leq d\}|=\sum_{i=0}^{d}\binom{n}{i}.$$

The last inequality follows by the binomial theorem since $d\leq n$, so

$$\sum_{i=0}^{d}\binom{n}{i}(d/n)^d\leq\sum_{i=0}^{d}\binom{n}{i}(d/n)^i\leq\sum_{i=0}^{n}\binom{n}{i}(d/n)^i=(1+(d/n))^n\leq e^d.$$

□

Exercise 4.9. Show that the Sauer-Shelah lemma is sharp for all n,d. That is, find $\mathcal{F}$ with $d:=\mathrm{VCdim}(\mathcal{F})$ such that

$$|\mathcal{F}|=\sum_{i=0}^{d}\binom{n}{i}.$$

(Hint: consider the set of $x\in\{0,1\}^n$ such that x has at most d entries equal to 1.)

Proposition 4.10. *Let $\mathcal{F}$ be the set of linear threshold functions*

$$\mathcal{F} := \{f\colon \mathbb{R}^n \to \{-1,1\} \qquad : \exists\, w \in \mathbb{R}^n,\, t \in \mathbb{R}, \quad f(x) = \mathrm{sign}(\langle w, x\rangle - t),\, \forall\, x \in \mathbb{R}^n\}.$$

Then $\mathrm{VCdim}(\mathcal{F}) = n + 1$.

Proof. Consider the set of vectors $B := \{0, e_1, \ldots, e_n\}$ with exactly $n - 1$ entries equal to 0, together with the zero vector. Then $|B| = n + 1$. We claim that B is shattered by $\mathcal{F}$. Indeed, if $g\colon B \to \{-1,1\}$ is given, denote $w := (g(e_1)-g(0), \ldots, g(e_n)-g(0)) \in \{-2,0,2\}^n$, then define $f(x) := \mathrm{sign}(\langle w, x\rangle + g(0))$. Then for any $1 \leq i \leq n$, $\langle w, e_i\rangle + g(0) = g(e_i)$, and $f(0) = g(0)$, so $f(x) = g(x)$ for all $x \in B$. We conclude that $\mathrm{VCdim}(\mathcal{F}) \geq n + 1$.

We now show that $\mathrm{VCdim}(\mathcal{F}) \leq n + 1$. We argue by contradiction. Suppose $B \subseteq \mathbb{R}^n$ is a set of $n + 2$ vectors that is shattered by $\mathcal{F}$. By Remark 2.5, there is a set $B' \subseteq \mathbb{R}^{n+1}$ of $n + 2$ vectors that is shattered by the set of homogeneous linear threshold functions

$$\mathcal{F}' := \{f\colon \mathbb{R}^{n+1} \to \{-1,1\} \qquad : \exists\, w \in \mathbb{R}^{n+1}, \quad f(x) = \mathrm{sign}(\langle w, x\rangle),\, \forall\, x \in \mathbb{R}^{n+1}\}.$$

Suppose $B' = \{x^{(1)}, \ldots, x^{(n+2)}\} \subseteq \mathbb{R}^{n+1}$. Then there exist constants $\alpha_1, \ldots, \alpha_{n+2} \in \mathbb{R}$ such that $\sum_{i=1}^{n+2} \alpha_i x^{(i)} = 0$. Let $I := \{1 \leq i \leq n+2\colon \alpha_i > 0\}$, $J := \{1 \leq i \leq n+2\colon \alpha_i < 0\}$. Then

$$\sum_{i\in I} \alpha_i x^{(i)} = -\sum_{i\in J} \alpha_i x^{(i)}. \qquad (*)$$

Since B' is shattered by $\mathcal{F}'$, there is a $w \in \mathbb{R}^{n+1}$ such that $\langle w, x^{(i)}\rangle > 0$ for all $i \in I$ and $\langle w, x^{(i)}\rangle < 0$ for all $i \in J$. So, if I is nonempty,

$$0 < \sum_{i\in I} \alpha_i \langle w, x^{(i)}\rangle = \Big\langle w, \sum_{i\in I} \alpha_i x^{(i)}\Big\rangle \overset{(*)}{=} -\Big\langle w, \sum_{i\in J} \alpha_i x^{(i)}\Big\rangle = -\sum_{i\in J} \alpha_i \langle w, x^{(i)}\rangle \leq 0.$$

So, we have a contradiction when I is nonempty. Since either I or J must be nonempty, the proof is complete. □

The following concept, sometimes simply called an ε-net in the literature, is quite different from the ε-nets we discussed for metric spaces.

Definition 4.11 (Measure-Theoretic ε-net). Let $0 < \varepsilon < 1$. Let A be a set and let $\mathbb{P}$ be a probability law on A. Let Ω be a set of subsets of A. A measure theoretic ε**-net** for Ω is a set of points $S \subseteq A$ such that, for every $B \in \Omega$ with $\mathbb{P}(B) > \varepsilon$, there exists $s \in S$ such that $\{s\} \cap B \neq \emptyset$.

That is, every region with large measure contains some point in S. Since no reference to any metric is made in this definition, it is quite different that an ε-net for metric spaces.

Example 4.12. Let $\mathbb{P}$ be the uniform probability law on $A := [0,1]$. Let Ω be the set of closed intervals in $[0,1]$. Then for any $\omega \in \Omega$, $\mathbb{P}(\omega)$ is the usual length of ω. Let $\varepsilon > 0$. Then the points $\{i\varepsilon\}_{i=0}^{\lfloor 1/\varepsilon\rfloor}$ is an ε-net for Ω, since any closed interval in $[0,1]$ of width larger than ε must intersect this set of points.

If on the other hand Ω is the set of *all* subsets of $A := [0,1]$ with the same $\mathbb{P}$ as before, then no finite ε-net exists, since the complement of any finite set has measure 1. However, if $[0,1]$ has the usual metric on it, then a metric ε-net certainly exists.

Exercise 4.13. Show that both our notions of ε-net agree (up to changing the constant ε) in the following case: Ω is a metric space, $\mathbb{P}$ is a probability law on Ω, $A = \{B(x,r)\colon x \in \Omega, r > 0\}$ and there exist $a, b, c_1, c_2 > 0$ such that $c_1 r^a \leq \mathbb{P}(B(x,r)) \leq c_2 r^b$ for all $x \in \Omega, r > 0$.

Below, we consider $\mathcal{F}$ to be a subset of $\{0,1\}$-valued functions on A. Let $\mathbb{P}$ be a probability law on A. Let $f, g \in \mathcal{F}$. Since $f = 1_{\{f=1\}}$, we can identify f with the set where it is 1 and extend set operations to functions in $\mathcal{F}$. For example, $f \cap g := 1_{\{f=1\}\cap\{g=1\}}$ and $f\Delta g := 1_{\{f=1\}\Delta\{g=1\}}$, where Δ denotes symmetric difference. Then $f \cap g = g \cap f$ and $f\Delta g = g\Delta f$. Also, we can define $\mathbb{P}(f) := \mathbb{P}(f = 1) = \mathbb{E}f$, so that $\mathbb{P}$ can be extended to a probability law on $\{0,1\}^A$. The notion of measure-theoretic ε-net for a set of boolean functions is then well-defined using this probability law $\mathbb{P}$. Define now

$$D(f) := \{f\Delta g \colon g \in \mathcal{F}\}.$$

Exercise 4.14. For any $f \in \mathcal{F}$,

$$\mathrm{VCdim}(\mathcal{F}) = \mathrm{VCdim}(D(f)).$$

Lemma 4.15. *Let $\varepsilon > 0$. Suppose $f \in \mathcal{F}$ is a function to be learned by an algorithm and S is a (measure-theoretic) ε-net for $D(f)$. Suppose the algorithm outputs a hypothesis $g \in \mathcal{F}$ such that $f(s) = g(s)$ for all $s \in S$. Then*

$$\mathbb{P}(f \neq g) < \varepsilon.$$

Proof. Since $f(s) = g(s)$ for all $s \in S$, $(g\Delta f)(s) = 0$ for all $s \in S$, so $(g\Delta f) \cap \{s\} = 0$ for all $s \in S$. Since $g\Delta f \in D(f)$, and S is an ε-net for $D(f)$, we conclude by the definition of ε-net that $\mathbb{P}(g\Delta f) < \varepsilon$. That is, $\mathbb{P}(f \neq g) < \varepsilon$. □

So, if $f \in \mathcal{F}$ is the function to be learned, the above lemma shows that creating an ε-net for $D(f)$ with high probability is sufficient to PAC learn $\mathcal{F}$. In retrospect, the Occam's Razor bound in Theorem 3.20 used this fact implicitly.

Proposition 4.16. *Let $\mathbb{P}$ be a probability law on A. Let $\mathcal{F} \subseteq \{0,1\}^A$. Let $d := \mathrm{VCdim}(\mathcal{F})$. Let $0 < \varepsilon, \delta < 1$. Let S be a random sample from A of size m where*

$$m \geq 100(\varepsilon^{-1}\log(1/\delta) + d\varepsilon^{-1}\log(1/\varepsilon)).$$

Then with probability at least $1 - \delta$, S is a (measure-theoretic) ε-net in $\mathcal{F}$.

Proof. Let S_1 be a random sample of size m from A. Let C be the event that S_1 does not form an ε-net in $\mathcal{F}$. Our goal is to upper bound $\mathbb{P}(C)$. If C occurs, then there exists $h \in \mathcal{F}$ with $\mathbb{P}h \geq \varepsilon$ and such that $h \cap S_1 = \emptyset$. Let S_2 be a random size of size m from A that is independent of S_1. Let X be the number of times that S_2 intersects h. Write $X = \sum_{i=1}^m X_i$, where $X_i := 1_{\{i^{th}\text{ sample of } S_2 \text{ intersects } h\}}$ for all $1 \leq i \leq m$. By Chebyshev's inequality, $\forall\, t > 0$,

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}X| > t\mathbb{E}X) &= \mathbb{P}(|X - \mathbb{E}X| > tm\mathbb{E}X_1) \leq t^{-2}m^{-2}(\mathbb{E}X_1)^{-2}\mathrm{Var}(X)\\ &= t^{-2}m^{-2}(\mathbb{E}X_1)^{-2}m\mathrm{Var}(X_1) = t^{-2}m^{-1}((\mathbb{E}X_1)^{-1} - 1) \leq t^{-2}m^{-1}(\varepsilon^{-1} - 1).\end{aligned}$$

Here we used $\mathbb{E}X_1 \geq \varepsilon$ since $\mathbb{P}h \geq \varepsilon$, which also implies

$$\mathbb{P}(X < \varepsilon m/2) \leq \mathbb{P}(X < \mathbb{E}X/2) \leq \mathbb{P}(|X - \mathbb{E}X| > \mathbb{E}X/2) \leq 4m^{-1}(\varepsilon^{-1} - 1)$$

So, if $\varepsilon < 1/2$ and $m > 10/\varepsilon$, $\mathbb{P}(X < \varepsilon m/2) < 1/2$. Let C' be the event that C occurs (so there exists $h \in \mathcal{F}$ with $\mathbb{P}h \geq \varepsilon$, $h \cap S_1 = \emptyset$) and $|S_2 \cap h| > \varepsilon m/2$. We have shown that $\mathbb{P}(C'|C) \geq 1/2$, and since $C' \subseteq C$, we have $\mathbb{P}(C'|C) = \mathbb{P}(C')/\mathbb{P}(C)$, so

$$\mathbb{P}(C') \geq (1/2)\mathbb{P}(C).$$

So, in order to upper bound $\mathbb{P}(C)$ it suffices to upper bound $\mathbb{P}(C')$.

Let S be a random sample from A of size $2m$ and let $h \in \mathcal{F}$ with $\mathbb{P}h \geq \varepsilon$. Suppose $|S \cap h| > \varepsilon m/2$. Then, choose two disjoint $T_1, T_2 \subseteq S$ each of cardinality m, uniformly at random among all such partitions of S into two equal sized sets. Then since (T_1, T_2) is equal in distribution to (S_1, S_2), we have

$$\mathbb{P}(C') = \mathbb{P}(\exists\, \widetilde{h} \in \mathcal{F}\colon \mathbb{P}(\widetilde{h}) \geq \varepsilon,\, \widetilde{h} \cap T_1 = \emptyset,\, |\widetilde{h} \cap T_2| \geq \varepsilon m/2).$$

The probability $\mathbb{P}(\widetilde{h} \cap T_1 = \emptyset, |\widetilde{h} \cap T_2| \geq \varepsilon m/2)$ only depends on $\widetilde{h}|_{T_1 \cup T_2}$. From the union bound and Lemma 4.8,

$$\begin{aligned}\mathbb{P}(C') &\leq \mathbb{E}\Big(\sum_{\widetilde{h} \in \mathcal{F}|_{T_1 \cup T_2}\colon \mathbb{P}(\widetilde{h}) \geq \varepsilon,} 1_{\{\widetilde{h} \cap T_1 = \emptyset,\, |\widetilde{h} \cap T_2| \geq \varepsilon m/2\}} \Big) \\ &\leq (2em/d)^d \mathbb{P}(h \cap T_1 = \emptyset,\, |h \cap T_2| \geq \varepsilon m/2). \qquad (*)\end{aligned}$$

The probability $\mathbb{P}(h \cap T_1 = \emptyset,\, |h \cap T_2| \geq \varepsilon m/2)$ can be computed from the following combinatorial problem. Suppose we have $2m$ cubes sorted into piles U_1, U_2 each of size m, and we label $\ell \geq \varepsilon m/2$ cubes red, uniformly at random. Then $\mathbb{P}(h \cap T_1 = \emptyset,\, |h \cap T_2| \geq \varepsilon m/2)$ is the probability that all red cubes are in U_1. If ℓ is fixed, this probability is

$$\frac{\binom{m}{\ell}}{\binom{2m}{\ell}} = \frac{m!(2m-\ell)!}{(m-\ell)!(2m)!} = \prod_{i=0}^{\ell-1} \frac{m-i}{(2m-i)} \leq \prod_{i=0}^{\ell-1} \frac{1}{2} = 2^{-\ell}.$$

So,

$$\mathbb{P}(h \cap T_1 = \emptyset) \leq \sum_{j=\ell}^{\infty} 2^{-j} \leq 2^{-\ell+1} \leq 2^{1-\varepsilon m/2}. \qquad (**)$$

Combining $(*)$ and $(**)$ gives

$$\mathbb{P}(C') \leq 2 \left(\frac{2em}{d}\right)^d 2^{-\varepsilon m/2}.$$

That is,

$$\log \mathbb{P}(C) \leq \log 4 + d[1 + \log(2m/d)] - \log(2)\varepsilon m/2$$

So, choosing $m \geq 10\varepsilon^{-1}\log(1/\delta) + 10d\varepsilon^{-1}\log(1/\varepsilon)$ means $\log \mathbb{P}(C) \leq \log \delta$, as desired. □

Combining Exercise 4.14 with Propositions 4.15 and 4.16 proves the following Theorem, sometimes called the Fundamental Theorem of Statistical Learning or the Fundamental Theorem of Machine Learning

Theorem 4.17 (**Fundamental Theorem of Statistical Learning, Version 1**). *Let A be a set. Let $\mathcal{F} \subseteq \{0,1\}^A$ be a class of boolean functions. Let $d := \mathrm{VCdim}(\mathcal{F})$. Suppose $f \in \mathcal{F}$ is a function to be learned by an algorithm. Let S be a random sample from A of size m where*

$$m \geq 100(\varepsilon^{-1}\log(1/\delta) + d\varepsilon^{-1}\log(1/\varepsilon)).$$

Then with probability at least $1-\delta$, S is a (measure-theoretic) ε-net for $D(f)$. Suppose the algorithm outputs a hypothesis $g \in \mathcal{F}$ such that $f(s) = g(s)$ for all $s \in S$. Then

$$\mathbb{P}(f \neq g) < \varepsilon.$$

That is, the algorithm can PAC learn $\mathcal{F}$.

Theorem 4.17 immediately implies that, if an algorithm can output a hypothesis g that agrees with the values of f on the random sample of sufficiently large size, then this algorithm can PAC learn. So, knowing that a function class has a relatively small VC-dimension immediately implies that it can be PAC learned. One major caveat of Theorem 4.17 is that **it does not guarantee efficient PAC learnability**. In fact, the task of finding the hypothesis $g \in \mathcal{F}$ that agrees with f might be computationally hard. For example, if $\mathcal{F}$ is the class of 3-term DNF formulae on n variables, then $\mathrm{VCdim}(\mathcal{F}) \leq \log_2 |\mathcal{F}| \leq \log_2(3^{3n}) \leq 6n$, but we know from Theorem 3.7 that $\mathcal{F}$ is not efficiently PAC learnable (in its own function class).

Theorem 4.18 (**Fundamental Theorem of Statistical Learning, Version 2**). *Let A be a set. Let $\mathcal{F}, \mathcal{G} \subseteq \{0,1\}^A$ be two classes of boolean functions. Let $d := \mathrm{VCdim}(\mathcal{G})$. Suppose $f \in \mathcal{F}$ is a function to be learned by an algorithm. Let S be a random sample from A of size m where*

$$m \geq 100(\varepsilon^{-1}\log(1/\delta) + d\varepsilon^{-1}\log(1/\varepsilon)).$$

Then with probability at least $1-\delta$, S is a (measure-theoretic) ε-net for $D(f)$. Suppose the algorithm outputs a hypothesis $g \in \mathcal{G}$ such that $f(s) = g(s)$ for all $s \in S$. Then

$$\mathbb{P}(f \neq g) < \varepsilon.$$

That is, the algorithm can PAC learn $\mathcal{F}$, viewed as a subset of $\mathcal{G}$.

4.1. **Applications of the Fundamental Theorem.** In boosting and other applications, we find

Proposition 4.19. *Let $\mathcal{G}$ be a class of functions on a set A. Let $n \geq 3$. Let $\mathcal{F}_n$ be the set of linear threshold functions in n elements of $\mathcal{G}$:*

$$\mathcal{F}_n := \{f \colon A \to \{-1,1\} \qquad \colon \exists\, w \in \mathbb{R}^n,\, t \in \mathbb{R}, \quad f(x) = \mathrm{sign}(\sum_{i=1}^n w_i g_i(x) - t),\, \forall\, x \in A\}.$$

Assume that $\mathrm{VCdim}(\mathcal{G}) \geq 3$. *Then*

$$\mathrm{VCdim}(\mathcal{F}_n) \leq 10(\mathrm{VCdim}(\mathcal{G})+1)(n+1)\log\Big((\mathrm{VCdim}(\mathcal{G})+1)(n+1)\Big)$$

.

Proof. Let $d := \mathrm{VCdim}(\mathcal{G})$. Let $B := \{x^{(1)}, \ldots, x^{(m)}\} \subseteq A$ be a set that is shattered by $\mathcal{F}_n$. By the Sauer-Shelah Lemma 4.8, there are at most $(em/d)^d$ functions from B to $\{-1,1\}$. Constructing a function in $\mathcal{F}_n$ requires n such functions. Once $x \in B$ is fixed, there are then 2^{n+1} ways of sorting the values $g_1(x), \ldots, g_n(x)$ into two sets. So, the number of functions in $\mathcal{F}_n$ when restricted to B is at most

$$(em/d)^{dn}2^{n+1} \leq m^{(d+1)(n+1)},$$

using $m \geq \mathrm{VCdim}(\mathcal{G}) \geq 3$. By assumption, B is shattered by $\mathcal{F}_n$, so

$$2^m \leq 2^{\mathrm{VCdim}(\mathcal{F}_n)} \leq |\mathcal{F}_n| \leq m^{(d+1)(n+1)}.$$

That is,

$$m\log 2 \leq (d+1)(n+1)\log m$$

Exercise 4.20 concludes the proof. □

Exercise 4.20. Suppose $x, a \geq 1$. Assume that

$$x < a\log(x).$$

Then

$$x < 2a\log a.$$

Iterating Proposition 4.19 leads to sub-optimal bounds. However, the argument there can be adapted to the following more general class of functions.

Definition 4.21 (**Feedforward Neural Network**). A **feedforward neural network** with k layers and boolean activation function is a function f defined as follows. Let $n_0, \ldots, n_{k-1}$ be positive integers. For each $1 \leq i \leq k-1$, assume that

$$f_i \colon \mathbb{R}^{n_{i-1}} \to \{-1,1\}^{n_i},$$
$$f_k \colon \mathbb{R}^{n_{k-1}} \to \{-1,1\}.$$

Assume also that for all $1 \leq i \leq k$, there exists $w^{(i)} \in \mathbb{R}^{d_{i-1}}, t_{ij} \in \mathbb{R}$ such that the j^{th} component of f_i satisfies

$$f_{i,j}(x) = \mathrm{sign}(\langle w^{(i)}, x\rangle - t_{ij}), \qquad \forall\, x \in \mathbb{R}^{n_{i-1}}.$$

Then f is defined to be a function of the form

$$f := f_k \circ f_{k-1} \circ \cdots \circ f_1$$

Corollary 4.22. *Let $\mathcal{F}$ be the class of feedforward neural network with k layers. For all $1 \leq i \leq k$ and for all $1 \leq j \leq n_i$, assume that $f_{ij} \in \mathcal{F}_{ij}$, where $\mathcal{F}_{ij}$ is a class of functions. Denote $d_{ij} := \mathrm{VCdim}(\mathcal{F}_{ij}), d := \sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}$, $n := \sum_{i=1}^{k} n_i$.*

$$\mathrm{VCdim}(\mathcal{F}) \leq 10d\log(den).$$

Proof. Let B be a set of m points in the domain of f that is shattered by $\mathcal{F}$. By the Sauer-Shelah Lemma 4.8, there are at most $(em/d_{ij})^{d_i}$ functions in $\mathcal{F}_{ij}$ from m points to $\{-1,1\}$. Constructing the function f requires choosing each $f_{ij} \in \mathcal{F}_{ij}$. So, the number of such functions f restricted to B is at most

$$\prod_{i=1}^{k}\prod_{j=1}^{n_i}(em/d_{ij})^{d_{ij}},$$

By assumption, B is shattered by $\mathcal{F}$, so

$$2^m \leq 2^{\mathrm{VCdim}(\mathcal{F})} \leq \prod_{i=1}^{k}\prod_{j=1}^{n_i}(em/d_{ij})^{d_{ij}}. \qquad (*)$$

A Lagrange multiplier argument shows that $\sum_{j=1}^{n} -\alpha_j \log\alpha_j \leq \log n$ for any $\alpha_1, \ldots, \alpha_n \geq 0$ with $\sum_{j=1}^{n}\alpha_n = 1$. Letting $\alpha_{ij} := d_{ij}/d$ and exponentiating both sides, we get

$$\prod_{i=1}^{k}\prod_{j=1}^{n_i} d_{ij}^{d_{ij}} \geq (d/n)^d.$$

Combining this with $(*)$,

$$2^m \leq (nem/d)^d.$$

That is,

$$m\log 2 \leq d\log(nem/d).$$

Exercise 4.20 concludes the proof. □

Combining Proposition 4.10 and Theorem 4.17 implies that the set of linear threshold functions can be PAC learned (though not necessarily efficiently). A generalization of Proposition 4.10 says [Ant95]: if $\mathcal{F}_{n,m}$ is the class of polynomial threshold functions of degree at most m, i.e.

$$\mathcal{F}_{n,m} = \Big\{\mathrm{sign}(p)\colon p(x) = \sum_{S\subseteq\{1,\ldots,n\}\colon |S|\leq m} w_S \prod_{i\in S} x_i,\ \forall\, x=(x_1,\ldots,x_n)\in\mathbb{R}^n\Big\},$$

then $\mathrm{VCdim}(\mathcal{F}_{n,m}) = \sum_{i=0}^{m}\binom{n}{i} \leq (en/m)^m$. So, Theorem 4.17 implies that the set of polynomial threshold functions can be PAC learned (though not necessarily efficiently).

Similarly, Corollary 4.22 and Theorem 4.17 implies that the set of feedforward neural networks can be PAC learned (though not necessarily efficiently).

5. Some Concentration of Measure

The mathematical tools contained in this section will be used in Section 6.

5.1. Concentration for Independent Sums.

In certain cases, we can make rather strong conclusions about the distribution of sums of i.i.d. random variables, improving upon the laws of large numbers.

Theorem 5.1 (**Hoeffding Inequality/ Large Deviation Estimate**). *Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$. Let $a_1, a_2, \ldots \in \mathbb{R}$. Then, for any $n \geq 1$,*

$$\mathbb{P}\Big(\sum_{i=1}^{n} a_i X_i \geq t\Big) \leq e^{-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}}, \qquad \forall\, t\geq 0.$$

Consequently,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} a_i X_i\Big| \geq t\Big) \leq 2e^{-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}}, \qquad \forall\, t\geq 0.$$

Proof. By dividing $a_1, \ldots, a_n$ by a constant, we may assume $\sum_{i=1}^{n} a_i^2 = 1$. Let $\alpha > 0$. Using the (exponential) moment method as in Markov's inequality, and $\alpha t \geq 0$,

$$\mathbb{P}(\sum_{i=1}^{n} a_i X_i \geq t) = \mathbb{P}(e^{\alpha\sum_{i=1}^{n} a_i X_i} \geq e^{\alpha t}) \leq e^{-\alpha t}\mathbb{E}e^{\alpha\sum_{i=1}^{n} a_i X_i} = e^{-\alpha t}\prod_{i=1}^{n}\mathbb{E}e^{\alpha a_i X_i}.$$

The last equality used independence of $X_1, X_2, \ldots$. Using an explicit computation and Exercise 5.2,

$$\mathbb{E}e^{\alpha a_i X_i} = (1/2)(e^{\alpha a_i} + e^{-\alpha a_i}) = \cosh(\alpha a_i) \leq e^{\alpha^2 a_i^2/2}, \qquad \forall\, i\geq 1.$$

In summary, for any $t\geq 0$

$$\mathbb{P}(\sum_{i=1}^{n} a_i X_i \geq t) \leq e^{-\alpha t} e^{\alpha^2\sum_{i=1}^{n} a_i^2/2} = e^{-\alpha t + \alpha^2/2}.$$

Since $\alpha > 0$ is arbitrary, we choose α to minimize the right side. This minimum occurs when $\alpha = t$, so that $-\alpha t + \alpha^2/2 = -t^2/2$, giving the first desired bound. The final bound follows by writing $\mathbb{P}(|\sum_{i=1}^n a_i X_i| \geq t) = \mathbb{P}(\sum_{i=1}^n a_i X_i \geq t) + \mathbb{P}(-\sum_{i=1}^n a_i X_i \geq t)$ and then applying the first inequality twice. □

Exercise 5.2. Show that $\cosh(x) \leq e^{x^2/2}$, $\forall\, x \in \mathbb{R}$.

In particular, Hoeffding's inequality implies that

$$\mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq 2e^{-nt^2/2}, \qquad \forall\, t \geq 0.$$

This inequality is much stronger than either Markov's or Cheyshev's inequality, since they only respectively imply that

$$\mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq \frac{1}{t}, \quad \mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq \frac{1}{nt^2}, \qquad \forall\, t \geq 0.$$

Note also that Hoeffding's inequality gives a quantitative bound for any fixed $n \geq 1$, unlike the (non-quantitative) limit theorems which only hold as $n \to \infty$.

Exercise 5.3 (Chernoff Inequality). Let $0 < p < 1$. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbb{P}(X_1 = 1) = p$ and $\mathbb{P}(X_1 = 0) = 1 - p$ for any $i \geq 1$. Then for any $n \geq 1$

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n X_i \geq t\Big) \leq e^{-np}\Big(\frac{ep}{t}\Big)^{tn}, \qquad \forall\, t \geq p.$$

Prove the same estimate for $\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \leq t)$ for any $t \leq p$. (Hint: $1 + x \leq e^x$ for any $x \in \mathbb{R}$, so $1 + (e^\alpha - 1)p \leq e^{(e^\alpha - 1)p}$.)

Exercise 5.4. For any natural number n and a parameter $0 < p < 1$, define an Erdös-Renyi graph on n vertices with parameter p to be a random graph (V, E) on a (deterministic) vertex set V of n vertices (thus (V, E) is a random variable taking values in the discrete space of all $2^{\binom{n}{2}}$ possible undirected graphs one can place on V) such that the events $\{i, j\} \in E$ for unordered pairs with $i, j \in V$ are independent and each occur with probability p.

Suppose we have an Erdös-Renyi random graph $G = (V, E)$ on n vertices with parameter $0 < p < 1$. Define $d := p(n-1)$.

- Show that d is the expected degree of each vertex in G. (The degree of a vertex $v \in V$ is the number of vertices connected to v by an edge in E.)
- Show that there exists a constant $c > 0$ such that the following holds. Assume $p \geq \frac{c \log n}{n}$. Then with probability larger than .9, all vertices of G have degrees in the range $(.9d, 1.1d)$. (Hint: first consider a single vertex, then use the union bound over all vertices.)

5.2. **Concentration for Lipschitz Functions.** One way to phrase the general question in the subject of concentration of measure is: how far is a random variable from its mean value? Hoeffding's Inequality says that linear functions of mean zero ±1 valued independent random variables are exponentially close to their mean value. A similar statement can be made for bounded random variables (see Theorem 5.8 below). In order to answer the general question, we next consider Lipschitz functions of i.i.d. random variables. We focus on the Gaussian setting for simplicity.

For any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we denote $\|x\| := (x_1^2 + \cdots + x_n^2)^{1/2}$.

Theorem 5.5 (Concentration of measure for Gaussians, Lipschitz function form)). *Let $f\colon \mathbb{R}^n \to \mathbb{R}$. Suppose that for all $x, y \in \mathbb{R}^n$, $|f(x) - f(y)| \leq \|x - y\|$, so that f is 1-Lipschitz. Let $X = (X_1, \ldots, X_n)$ be a mean zero Gaussian random vector with identity convariance matrix. Then for all $t > 0$,*

$$\mathbb{P}\left(x \in \mathbb{R}^n \colon\ |f(x) - \mathbb{E}f(X)| \geq t\right) \leq 2e^{-2t^2/\pi^2}.$$

Proof. We assume that f all partial derivatives of f exist and are continuous. Let $Y = (Y_1, \ldots, Y_n)$ be another mean zero Gaussian random vector with identity convariance matrix, such that Y and X are independent. Let $0 \leq \theta \leq \pi/2$ and define

$$Z_\theta := X \sin\theta + Y \cos\theta.$$

By rotation invariance of a Gaussian random vector, Z_θ and $\frac{d}{d\theta} Z_\theta = X\cos\theta - Y\sin\theta$ have the same joint distribution as X and Y (since the vectors $(\sin\theta, \cos\theta)$ and $(\cos\theta, -\sin\theta)$ are orthogonal in $\mathbb{R}^2$.) Let $\phi\colon \mathbb{R} \to [0, \infty)$ be a convex function. Using then Jensen's Inequality, then the Chain Rule, then Jensen's inequality and Fubini's Theorem,

$$\begin{aligned}
&\mathbb{E}\phi(f(X) - \mathbb{E}f(Y)) \leq \mathbb{E}\phi(f(X) - f(Y)) = \mathbb{E}\phi\Big(\int_0^{\pi/2} \frac{d}{d\theta} f(Z_\theta) d\theta\Big) \\
&\quad= \mathbb{E}\phi\Big(\int_0^{\pi/2} \langle(\nabla f)(Z_\theta), \frac{d}{d\theta} Z_\theta\rangle d\theta\Big) = \mathbb{E}\phi\Big(\frac{1}{\pi/2}\int_0^{\pi/2} \frac{\pi}{2}\langle(\nabla f)(Z_\theta), \frac{d}{d\theta} Z_\theta\rangle d\theta\Big) \\
&\quad\leq \mathbb{E}\frac{1}{\pi/2}\int_0^{\pi/2} \phi\Big(\frac{\pi}{2}\langle(\nabla f)(Z_\theta), \frac{d}{d\theta} Z_\theta\rangle\Big) d\theta = \frac{1}{\pi/2}\int_0^{\pi/2} \mathbb{E}\phi\Big(\frac{\pi}{2}\langle(\nabla f)(Z_\theta), \frac{d}{d\theta} Z_\theta\rangle\Big) d\theta \\
&\quad= \frac{1}{\pi/2}\int_0^{\pi/2} \mathbb{E}\phi\Big(\frac{\pi}{2}\langle(\nabla f)(X), Y\rangle\Big) d\theta = \mathbb{E}\phi\Big(\frac{\pi}{2}\langle(\nabla f)(X), Y\rangle\Big)
\end{aligned}$$

Let $\alpha \in \mathbb{R}$ and let $\phi(x) := e^{\alpha x}$ for all $x \in \mathbb{R}$. Then using independence in Y and Fubini's Theorem,

$$\mathbb{E}\exp(\alpha[f(X) - \mathbb{E}f(Y)]) \leq \mathbb{E}\exp\Big(\alpha\frac{\pi}{2}\sum_{i=1}^n \frac{\partial f}{\partial x_i}(X) Y_i\Big) = \mathbb{E}_X \prod_{i=1}^n \mathbb{E}_Y \exp\Big(\alpha\frac{\pi}{2}\frac{\partial f}{\partial x_i}(X) Y_i\Big).$$

Using an explicit computation, for any $s \in \mathbb{R}$ and for any $1 \leq i \leq n$,

$$\mathbb{E}_Y e^{sY_i} = \int_{-\infty}^{\infty} e^{sy} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} = e^{s^2/2}\int_{-\infty}^{\infty} e^{-(y-s)^2/2} \frac{dy}{\sqrt{2\pi}} = e^{s^2/2}.$$

So, applying this inequality with $s = \alpha\frac{\pi}{2}\frac{\partial f}{\partial x_i}(X)$ for each $1 \leq i \leq n$,

$$\mathbb{E}\exp(\alpha[f(X) - \mathbb{E}f(Y)]) \leq \mathbb{E}\exp\Big(\alpha^2\frac{\pi^2}{8}\sum_{i=1}^n \Big(\frac{\partial f}{\partial x_i}(X)\Big)^2\Big) \leq \exp\Big(\alpha^2\frac{\pi^2}{8}\Big).$$

(Since f is 1-Lipschitz, $|\langle \nabla f(x), y\rangle| \leq 1$ for all $x, y \in \mathbb{R}^n$ with $\|y\| \leq 1$. In particular, using $y := \nabla f(x)/\|\nabla f(x)\|$, we get $\|\nabla f(x)\| \leq 1$.) So,

$$\begin{aligned}\mathbb{P}(f(X) - \mathbb{E}f(Y) > t) &= \mathbb{P}(\exp(\alpha[f(X) - \mathbb{E}f(Y)]) > e^{\alpha t})\\ &\leq e^{-\alpha t}\exp\Big(\alpha^2\frac{\pi^2}{8}\Big) = \exp\Big(-\alpha t + \alpha^2\frac{\pi^2}{8}\Big).\end{aligned}$$

The minimum α occurs when $\alpha = 4t/\pi^2$, so making this choice of α, we get

$$\mathbb{P}(f(X) - \mathbb{E}f(Y) > t) \leq \exp(-2t^2/\pi^2).$$

Similarly, $\mathbb{P}(f(X) - \mathbb{E}f(Y) < -t) \leq \exp(-2t^2/\pi^2)$, so that

$$\begin{aligned}\mathbb{P}(|f(X) - \mathbb{E}f(Y)| > t) &= \mathbb{P}(f(X) - \mathbb{E}f(Y) > t) + \mathbb{P}(f(X) - \mathbb{E}f(Y) < -t)\\ &\leq 2\exp(-2t^2/\pi^2).\end{aligned}$$

□

Theorem 5.6 (**Johnson-Lindenstrauss Lemma**). *Let $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^m$. Let $\varepsilon > 0$. Then there exists a linear function $h\colon \mathbb{R}^m \to \mathbb{R}^{O(\varepsilon^{-2}\log n)}$ such that*

$$\left\|x^{(i)} - x^{(j)}\right\| \leq \left\|h(x^{(i)}) - h(x^{(j)})\right\| \leq (1+\varepsilon)\left\|x^{(i)} - x^{(j)}\right\|, \qquad \forall\, 1 \leq i, j \leq n.$$

One proves this via the probabilistic method. By concentration of measure, a random projection does what we require.

Proof. Fix $1 \leq k \leq m$. Let $\Pi\colon \mathbb{R}^m \to \mathbb{R}^m$ be the orthogonal projection such that

$$\Pi(z_1, \ldots, z_m) := (z_1, \ldots, z_k, 0, \ldots, 0), \qquad \forall\,(z_1, \ldots, z_m) \in \mathbb{R}^m.$$

Let $X = (X_1, \ldots, X_m)$ be a standard m-dimensional Gaussian random vector. Define

$$a := \mathbb{E}\|\Pi X\|.$$

We will eventually show that $a \geq 10^{-2}\sqrt{k}$. Observe

$$\mathbb{E}\|\Pi X\|^2 = \mathbb{E}\sum_{i=1}^{k} X_i^2 = k\mathbb{E}X_1^2. = k. \qquad (*)$$

Now, by Theorem 5.5 for the 1-Lipschitz function $x \mapsto \|\Pi x\|$,

$$\begin{aligned}\mathbb{E}\|\Pi X\|^4 &= \int_0^\infty 4u^3\mathbb{P}(\|\Pi X\| \geq u)du\\ &= \int_0^{2a} 4u^3\mathbb{P}(\|\Pi X\| \geq u)du + \int_{2a}^\infty 4u^3\mathbb{P}(\|\Pi X\| \geq u)du\\ &\leq \int_0^{2a} 4u^3 du + \int_{2a}^\infty 4u^3\mathbb{P}(|\,\|\Pi X\| - a| > u/2)du\\ &\leq 16a^4 + 8\int_{2a}^\infty u^3 e^{-u^2/2\pi^2}du = 16a^4 + 8(2\pi^2)(2a^2+\pi^2)e^{-2a^2/\pi^2} \leq 16a^4 + 2\pi^4\\ &\leq 16a^4 + 200k^2 \leq 216\left(\int_{\mathbb{R}^m}\|\Pi x\|^2\gamma_m(x)dx\right)^2 \quad \text{, using Jensen's inequality and } (*).\end{aligned}$$

So, if $Z := \|\Pi X\|$ is a random variable, we have shown that $\mathbb{E}Z^4 < c(\mathbb{E}Z^2)^2$ where $c := 216$. So, using Hölder's Inequality, for $p = 3/2$, $q = 3$,

$$\mathbb{E}Z^2 = \mathbb{E}(Z^{2/3}Z^{4/3}) \leq (\mathbb{E}Z)^{2/3}(\mathbb{E}Z^4)^{1/3} \leq (\mathbb{E}Z)^{2/3}c^{1/3}(\mathbb{E}Z^2)^{2/3}.$$

Using this inequality and $(*)$,

$$\mathbb{E}Z \geq c^{-1/2}\sqrt{\mathbb{E}Z^2} \geq 216^{-1/2}\sqrt{k}. \qquad (**)$$

In summary, $a \geq 2^{-4}\sqrt{k}$ for a defined above.

Let A be an $m \times m$ matrix of i.i.d. standard Gaussian random variables. Fix $x^{(0)} \in \mathbb{R}^m$ with $\|x\| = 1$. By rotation invariance of the Gaussian measure, A and AQ have the same distribution where Q is a fixed $m \times m$ orthogonal matrix, so if we choose Q so that $Q(1,0,\ldots,0)^T = x^{(0)}$, we get

$$\begin{aligned}\mathbb{P}\left(A \in \mathbb{R}^{m\times m} \colon \left|\left\|\Pi A x^{(0)}\right\|_2 - a\right| \geq \varepsilon a\right) &= \mathbb{P}\left(A \in \mathbb{R}^{m\times m} \colon \left|\left\|\Pi A(1,0,\ldots,0)^T\right\|_2 - a\right| \geq \varepsilon a\right)\\ &= \mathbb{P}\left(X \in \mathbb{R}^m \mid \left|\|\Pi X\| - a\right| \geq \varepsilon a\right).\end{aligned}$$

So, by Theorem 5.5 applied to the 1-Lipschitz function $x \mapsto \|\Pi x\|$, and using $a \geq 2^{-4}\sqrt{k}$, for any $\varepsilon > 0$, and for any

$$\mathbb{P}\left(A \in \mathbb{R}^{m\times m} \colon \left|\left\|\Pi A x^{(0)}\right\|_2 - a\right| \geq \varepsilon a\right) \leq 2e^{-2\varepsilon^2 a^2/\pi^2} \leq 2e^{-2^{-10}k\varepsilon^2}.$$

Let $x^{(1)}, \ldots, x^{(n)}$ be n points in $\mathbb{R}^m$. If $k \geq 2^{12}\varepsilon^{-2}\log n$, the union bound shows that

$$\mathbb{P}\left(A \in \mathbb{R}^{m\times m} \colon \exists\, i \neq j \colon \left|\left\|\Pi A\left(\frac{x^{(i)} - x^{(j)}}{\|x^{(i)} - x^{(j)}\|}\right)\right\| - a\right| \geq \varepsilon a\right) \leq \binom{n}{2}2e^{-2^{-10}k\varepsilon^2} < 1.$$

For any $1 \leq i \leq n$, define $y_i := \Pi A x^{(i)}/(a(1-\varepsilon))$. Then $\exists\ A \in \mathbb{R}^{n\times m}$ such that

$$1 \leq \left\|\frac{y^{(i)} - y^{(j)}}{\|x^{(i)} - x^{(j)}\|}\right\| \leq \frac{1+\varepsilon}{1-\varepsilon} \leq 1 + 3\varepsilon, \qquad \forall\, 1 \leq i,j \leq n.$$

So, our required embedding is $h := \frac{\Pi A}{a(1-\varepsilon)}$, so that $h(x^{(i)}) = y^{(i)}$ for all $1 \leq i \leq n$. Note that h is linear and its nonzero entries form a rectangular matrix of i.i.d. Gaussians. Also, we can choose $k := \lceil 2^{12}\varepsilon^{-2}\log n\rceil$. (In fact, if we choose k to be slightly larger, then the probability becomes exponentially small, so essentially all A satisfies our desired property, hence essentially all linear projections $h \colon \mathbb{R}^n \to \mathbb{R}^{O(\varepsilon^{-2}\log n)}$ satisfy our desired property.) □

Exercise 5.7. High-dimensional geometry is much different than low-dimensional geometry, as this exercise demonstrates.

- Show that "most" of the mass of an n-dimensional Gaussian is concentrated on the sphere of radius $\sqrt{n}$ centered at the origin. That is, if $X_1, \ldots, X_n$ are n i.i.d. standard Gaussian random variables, then

  $$\lim_{n\to\infty} \mathbb{P}(\sqrt{X_1^2 + \cdots + X_n^2} \in (n + \sqrt{3n}, n - \sqrt{3n}) \geq 2/3.$$

  In fact, you should be able to compute the limit exactly.
- Generally, "most" of the mass of a high-dimensional convex body is concentrated near the surface of the body. Let Vol_n denote the usual volume in $\mathbb{R}^n$ (so that the volume of a unit square $[0,1]^n$ is 1.) For example, show that, for any $\varepsilon > 0$,

  $$\lim_{n\to\infty} \mathrm{Vol}_n\Big([-\frac{1}{2}(1-\varepsilon), \frac{1}{2}(1-\varepsilon)]^n\Big) = 0.$$

- Let $B_n := \{x \in \mathbb{R}^n \colon \|x\| \leq 1\}$ be the unit ball centered at the origin. Show that
$$\lim_{n\to\infty} \mathrm{Vol}_n(B_n) = 0.$$
- Let $C_n = \{x \in \{[-1/2, 1/2]^n \colon \exists\, y \in \{-1/2, 1/2\}^n \text{ such that } \|x - y\| \leq 1/2\}\}$ be the union of balls of radius $1/2$ centered at the corners of the hypercube $[-1/2, 1/2]^n$. Let $D_n := \{x \in \mathbb{R}^n \colon \|x\| \leq r\}$ be a ball of radius r centered at the origin, where r is chosen to be as large as possible so that D_n does not intersect the interior of C_n. (Put another way, D_n is tangent to the balls C_n.) Find
$$\lim_{n\to\infty} \mathrm{Vol}_n(D_n).$$
Before you do the computation, try to guess what the answer should be.

5.3. **Additional Comments.** Hoeffding's inequality in Theorem 5.1 can be generalized to the following statement.

Theorem 5.8 (**Hoeffding Inequality/ Large Deviation Estimate**). *For all $i \geq 1$, let $a_i < b_i$ be real numbers. Let $X_1, X_2, \ldots$ be independent random variables with $\mathbb{P}(X_i \in [a_i, b_i]) = 1$. Then, for any $n \geq 1$,*
$$\mathbb{P}\Big(\sum_{i=1}^n X_i - \mathbb{E}\big(\sum_{j=1}^n X_j\big) \geq t\Big) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}, \qquad \forall\, t \geq 0.$$
Consequently,
$$\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i - \mathbb{E}\big(\sum_{j=1}^n X_j\big)\Big| \geq t\Big) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}, \qquad \forall\, t \geq 0.$$

The proof of Theorem 5.8 imitates that of Theorem 5.1, while using the following Lemma.

Lemma 5.9 (**Hoeffding's Lemma**). *Let $a < b$ be real numbers. Let X be a random variable with $\mathbb{P}(X \in [a, b]) = 1$. Then for any $\alpha \in \mathbb{R}$,*
$$\mathbb{E}e^{\alpha(X - \mathbb{E}X)} \leq e^{\frac{1}{8}\alpha^2(b-a)^2}.$$

Proof. By replacing X with $X - \mathbb{E}X$, we may assume that $\mathbb{E}X = 0$ and instead prove that
$$\mathbb{E}e^{\alpha X} \leq e^{\frac{1}{8}\alpha^2(b-a)^2}.$$
Since $a \leq X \leq b$ with probability one, there is a random $Y \in [0, 1]$ such that $X =: aY + b(1 - Y)$. That is, this equality holds when $Y := (X - b)/(a - b)$. By convexity of $x \mapsto e^{\alpha x}$,
$$e^{\alpha X} = e^{\alpha(aY + b(1-Y))} \leq Ye^{\alpha a} + (1 - Y)e^{\alpha b} = \frac{X - b}{a - b}e^{\alpha a} + \frac{a - X}{a - b}e^{\alpha b}.$$
Let $\gamma := -b/(a - b)$, $c := \alpha(a - b)$, $f(c) := -\gamma c + \log(1 - \gamma + \gamma e^c)$. Taking expectations of both sides and using $\mathbb{E}X = 0$ we get
$$\mathbb{E}e^{\alpha X} \leq -\frac{b}{a - b}e^{\alpha a} + \frac{a}{a - b}e^{\alpha b} = e^{f(c)}.$$

Note that $f(0)=0$, $f'(c)=-\gamma+\frac{\gamma e^c}{1-\gamma+\gamma e^c}$, $f'(0)=0$, and

$$f''(c)=\frac{\gamma e^c}{1-\gamma+\gamma e^c}-\frac{\gamma^2 e^{2c}}{(1-\gamma+\gamma e^c)^2}=\frac{\gamma e^c}{1-\gamma+\gamma e^c}\Big(1-\frac{\gamma e^c}{1-\gamma+\gamma e^c}\Big)$$

So, if $s:=\frac{\gamma e^c}{1-\gamma+\gamma e^c}$, we have $f''(c)=s(1-s)\leq 1/4$. So, by Taylor's Theorem (with error term), for any $c\in\mathbb{R}$, there exists c_0 between 0 and c such that

$$f(c)=f(0)+cf'(0)+\frac{c^2}{2}f''(c_0)=\frac{c^2}{2}f''(c_0)\leq\frac{c^2}{8}.$$

In conclusion $\mathbb{E}e^{\alpha X}\leq e^{c^2/8}$. □

Theorem 5.5 can be generalized to uniformly log-concave densities on Euclidean space (see Ledoux, "The Concentration of Measure Phenomenon," Proposition 2.18)

Theorem 5.10 (Concentration of measure for Log-Concave Measures, Lipschitz function form)). *Let $f\colon\mathbb{R}^n\to\mathbb{R}$. Suppose that for all $x,y\in\mathbb{R}^n$, $|f(x)-f(y)|\leq\|x-y\|$, so that f is 1-Lipschitz. Let $u\colon\mathbb{R}^n\to\mathbb{R}$ be a function such that $e^{-u(x)}$ is a probability density on $\mathbb{R}^n$. Assume there exists $c>0$ such that the Hessian of u satisfies $\mathrm{Hess}(u)(x)\geq cI$, in the matrix sense. (That is, all eigenvalues of the Hessian of u are bounded below by c, for all $x\in\mathbb{R}^n$.) Let X have distribution e^{-u}. Then, for all $t>0$,*

$$\mathbb{P}\left(x\in\mathbb{R}^n\colon\ |f(x)-\mathbb{E}f(X)|\geq t\right)\leq 2e^{-ct^2/2}.$$

Note that Hoeffding's Inequality 5.8 provides the same bound when $\mathbb{P}(X_1=1)=\mathbb{P}(X_1=-1)=1$, or when $\mathbb{P}(X_1=1)=\varepsilon,\mathbb{P}(X_1=-1)=1-\varepsilon$ with $0<\varepsilon<1$ arbitrary. In the latter case, former case, Hoeffding's inequality is fairly sharp, but in the latter case, Hoeffding's inequality is not quite sharp. Put another way, Hoeffding's inequality does not account for the variance of the random variables. Bennett's inequality below does account for the variance of the random variables, with a potentially worse decay than Hoeffding's inequality for large values of t.

Theorem 5.11 (Bennett's Inequality). *Let $c>0$. Let $X_1,X_2,\ldots$ be independent random variables with $X_i\leq c$. For any $s\geq -1$, define $h(s):=(1+s)\log(1+s)-s$. Define $\sigma^2:=\sum_{i=1}^n\mathrm{Var}X_i^2$ and assume $0<\sigma<\infty$. Then, for any $n\geq 1$,*

$$\mathbb{P}\Big(\sum_{i=1}^n X_i-\mathbb{E}\big(\sum_{j=1}^n X_j\big)\geq t\Big)\leq e^{-\frac{\sigma^2}{c^2}h\left(\frac{ct}{\sigma^2}\right)}\leq e^{-\frac{t^2}{2+2ct/(3\sigma^2)}},\qquad\forall\, t\geq 0.$$

Proof. Without loss of generality, we may assume that $\mathbb{E}X_i=0$ and $X_i\leq 1$ for all $1\leq i\leq n$. For any $s\in\mathbb{R}$, define $\phi(s):=e^s-s-1$. Note that

$$\phi(\alpha)\leq\alpha^2/2,\qquad\text{if}\, s<0,\qquad\phi(\alpha x)\leq\alpha^2\phi(x),\qquad\text{if}\, x>0,\ \alpha\in[0,1].\qquad(*)$$

The second inequality follows e.g. from the power series expansion of ϕ. Using the definition of ϕ and $\mathbb{E}X_i=0$ for all $1\leq i\leq n$, for any $\alpha\in[0,1]$,

$$\begin{aligned}\mathbb{E}e^{\alpha X_i}&=1+\alpha\mathbb{E}X_i+\mathbb{E}\phi(\alpha X_i)=1+\mathbb{E}\phi(\alpha X_i)\\&\overset{(*)}{\leq}1+\mathbb{E}\phi(\alpha\max(X_i,0))+(\alpha^2/2)\mathbb{E}[\max(-X_i,0)]^2\end{aligned}$$

Using now $(*)$ and the bound $X_i \leq 1$ for all $1 \leq i \leq n$, and also $\psi(\alpha) \geq \alpha^2/2$ for $\alpha > 0$,

$$\mathbb{E}e^{\alpha X_i} \leq 1 + \phi(\alpha)\mathbb{E}[\max(X_i, 0)]^2 + (\alpha^2/2)\mathbb{E}[\max(-X_i, 0)]^2 \leq 1 + \phi(\alpha)\mathbb{E}X_i^2 \leq e^{\phi(\alpha)\mathbb{E}X_i^2}.$$

The proof now proceeds as in Hoeffding's inequality, Theorem 5.1. For any $t \geq 0$

$$\mathbb{P}(\sum_{i=1}^n X_i \geq t) \leq e^{-\alpha t}\mathbb{E}e^{\alpha \sum_{i=1}^n X_i} \leq e^{-\alpha t + \phi(\alpha)\sum_{i=1}^n \mathbb{E}X_i^2} = e^{-\alpha t + \phi(\alpha)\sigma^2}.$$

Since $\alpha > 0$ is arbitrary, we choose α to minimize the right side. This minimum occurs when $\frac{t}{\sigma^2} = \phi'(\alpha) = e^\alpha - 1$, so that

$$\alpha = \log\Big(1 + \frac{t}{\sigma^2}\Big).$$

At this value of α, $\phi(\alpha) = \frac{t}{\sigma^2} - \alpha$, so

$$-\alpha t + \sigma^2\phi(\alpha) = -t\log\Big(1 + \frac{t}{\sigma^2}\Big) + t - \sigma^2\log\Big(1 + \frac{t}{\sigma^2}\Big) = \frac{t}{\sigma^2}h(t/\sigma^2).$$

, giving the first desired bound. The final bound follows by writing $\mathbb{P}(|\sum_{i=1}^n a_iX_i| \geq t) = \mathbb{P}(\sum_{i=1}^n a_iX_i \geq t) + \mathbb{P}(-\sum_{i=1}^n a_iX_i \geq t)$ and then applying the first inequality twice. We also use the inequality $h(s) \geq s^2/(2 + 2s/3)$. □

6. Empirical Risk Minimization (ERM) and Concentration

Problem 6.1 (**Statistical Supervised Learning Problem**). Let A, B be sets. Let $f\colon A \to B$ be an unknown function. Let $\mathbb{P}$ be an unknown probability distribution on A. The goal of the learning problem is to approximately determine the function f on all of A using a small number of sample values of f on A. Let $X^{(1)}, \ldots, X^{(k)}$ be a random sample of size k (i.e. a sequence of independent random variables in A distributed according to $\mathbb{P}$) and let $Y^{(1)}, \ldots, Y^{(k)} \in B$. It is known that

$$f(X^{(i)}) = Y^{(i)}, \qquad \forall\, 1 \leq i \leq k.$$

The goal is to output a function $g\colon A \to B$ that minimizes the prediction error

$$\mathbb{P}(f(X^{(1)}) \neq g(X^{(1)})).$$

Since the probability distribution $\mathbb{P}$ is unknown, it is generally impossible to exactly minimize the prediction error. So, the goal is often restated as minimizing the **empirical risk** or **empirical error** defined as

$$\frac{1}{k}\left|i \in \{1, \ldots, k\}\colon g(X^{(i)}) \neq Y^{(i)}\right|. \qquad (*)$$

The task of minimizing the quantity $(*)$ is called **empirical risk minimization (ERM)**. We can equivalently write the empirical error as

$$\frac{1}{k}\sum_{i=1}^k 1_{g(X^{(i)}) \neq Y^{(i)}}.$$

As usual, we try to minimize this quantity over all $g \in \mathcal{F}$, where $\mathcal{F}$ is a specific class of functions from A to B.

Remark 6.2. Since the function G achieving the minimum of the empirical risk depends on the random samples $X^{(1)}, \ldots, X^{(k)}$, G will be a random i.e. *non-deterministic* function.

A basic question is then: how close is the empirical error to the true error? For example, given $t > 0$, can we bound

$$\mathbb{P}\Big(\left|\frac{1}{k}\sum_{i=1}^{k} 1_{g(X^{(i)})\neq Y^{(i)}} - \mathbb{P}(g(X^{(1)}) \neq Y^{(1)})\right| > t\Big)?$$

Since

$$\mathbb{E}\frac{1}{k}\sum_{i=1}^{k} 1_{g(X^{(i)})\neq Y^{(i)}} = \mathbb{P}(g(X^{(1)}) \neq Y^{(1)})$$

, we get a bound from Hoeffding's Inequality 5.8, namely

$$\mathbb{P}\Big(\left|\frac{1}{k}\sum_{i=1}^{k} 1_{g(X^{(i)})\neq Y^{(i)}} - \mathbb{P}(g(X^{(1)}) \neq Y^{(1)})\right| > t\Big) \leq 2e^{-2t^2 k}$$

Therefore:

Proposition 6.3. *For any $\delta > 0$ and for any $g\colon A \to B$, with probability at least $1 - \delta$, we have*

$$\left|\frac{1}{k}\sum_{i=1}^{k} 1_{g(X^{(i)})\neq Y^{(i)}} - \mathbb{P}(g(X^{(1)}) \neq Y^{(1)})\right| \leq \frac{1}{\sqrt{k}}\sqrt{\log(2/\delta)}.$$

This Proposition is however unsatisfactory. If G minimizes the empirical risk, then G is in fact a random function (since it depends on the random variables $X^{(1)}, \ldots, X^{(k)}$). On the other hand, if $\overline{g}$ minimizes the actual risk, then $\overline{g}$ is deterministic. So, in order to compare the minimizers of the empirical and actual risk, we need a bound as in Proposition 6.3 that is *uniform* over all possible minimizers of the empirical risk. For example, suppose that g comes from a class of functions $\mathcal{F}$. In order to compare the empirical and actual risk, we need to bound

$$\mathbb{P}\Big(\sup_{g\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k} 1_{g(X^{(i)})\neq Y^{(i)}} - \mathbb{P}(g(X^{(1)}) \neq Y^{(1)})\right| > t\Big).$$

If we can show this probability is small, then we can in fact conclude that the empirical risk minimum is close to the actual risk minimum. For example:

Proposition 6.4. *Suppose $\mathcal{F}$ is a finite class of functions from A to B. Let $f\colon A \to B$. Let $X^{(1)}, \ldots, X^{(k)}$ be a random sample of size k. Let G be a random element of $\mathcal{F}$ that minimizes the empirical risk $ER_k(g) := \frac{1}{k}\sum_{i=1}^{k} 1_{g(X^{(i)})\neq Y^{(i)}}$ among all $g \in \mathcal{G}$. Let $\overline{g}$ minimize the risk $r(g) := \mathbb{P}(g(X^{(1)}) \neq Y^{(1)})$ among all $g \in \mathcal{G}$. Let $\delta > 0$. Then, with probability at least $1 - \delta$,*

$$r(\overline{g}) \leq r(G) \leq r(\overline{g}) + 2\sqrt{\frac{\log(2\,|\mathcal{F}|) + \log(1/\delta)}{2k}}.$$

Proof. Let $g \in \mathcal{F}$. Using the union bound and Hoeffding's Inequality 5.8,

$$\mathbb{P}(\sup_{g\in\mathcal{F}} |ER_k(g) - r(g)| > t) \leq \sum_{g\in\mathcal{F}} \mathbb{P}(|ER_k(g) - r(g)| > t) \leq |\mathcal{F}|\, 2e^{-2t^2 k}.$$

So, choosing $t := \sqrt{\frac{\log(2|\mathcal{F}|)+\log(1/\delta)}{2k}}$, we get

$$\mathbb{P}(\sup_{g\in\mathcal{F}} |ER_k(g) - r(g)| > t) < \delta.$$

That is, with probability at least $1-\delta$, for all $g\in\mathcal{F}$,

$$|ER_k(g)-r(g)|\leq\sqrt{\frac{\log(2\left|\mathcal{F}\right|)+\log(1/\delta)}{2k}}.$$

In particular, if G minimizes the empirical risk, then with probability at least $1-\delta$,

$$\begin{aligned}r(\overline{g})\leq r(G)\leq ER_k(G)+\sqrt{\frac{\log(2\left|\mathcal{F}\right|)+\log(1/\delta)}{2k}}\\ \leq ER_k(\overline{g})+\sqrt{\frac{\log(2\left|\mathcal{F}\right|)+\log(1/\delta)}{2k}}\leq r(\overline{g})+2\sqrt{\frac{\log(2\left|\mathcal{F}\right|)+\log(1/\delta)}{2k}}.\end{aligned}$$

□

Remark 6.5. More generally, if G satisfies $ER_k(G)\leq\varepsilon+\min_{g\in\mathcal{F}}ER_k(g)$, then with probability at least $1-\delta$,

$$\begin{aligned}r(G)\leq ER_k(G)+\sqrt{\frac{\log(2\left|\mathcal{F}\right|)+\log(1/\delta)}{2k}}\\ \leq ER_k(\overline{g})+\varepsilon+\sqrt{\frac{\log(2\left|\mathcal{F}\right|)+\log(1/\delta)}{2k}}\leq r(\overline{g})+\varepsilon+2\sqrt{\frac{\log(2\left|\mathcal{F}\right|)+\log(1/\delta)}{2k}}.\end{aligned}$$

Proposition 6.4 shows that the empirical risk minimum and the true risk minimum are close to each other, as long as the function class $\mathcal{F}$ is finite. However, we would like to have a similar bound when $\mathcal{F}$ is not finite. Our goal is then to replace the bound in Proposition 6.4 with the VC-dimension. This will be accomplished in Theorem 6.26 below. Before doing so, we discuss some general ways of bounded the expected values of the suprema of random variables.

6.1. Gaussian Processes.

Theorem 6.6. *(**Slepian's Lemma**) Let $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_n)$ be n-dimensional Gaussian random vectors such that $EX_i=EY_i=0$ for $i=1,\ldots,n$, $EX_i^2=EY_i^2=1$ for $i=1,\ldots,n$. Assume that $EX_iX_j\leq EY_iY_j$ for $i,j\in\{1,\ldots,n\}$, $i\neq j$. Then*

$$P(X_1>0,\ldots,X_n>0)\leq P(Y_1>0,\ldots,Y_n>0)$$

More generally, for any $\alpha_1,\ldots,\alpha_n\in\mathbb{R}$,

$$P(X_1>\alpha_1,\ldots,X_n>\alpha_n)\leq P(Y_1>\alpha_1,\ldots,Y_n>\alpha_n)$$

Proof. Let $\{r_{ij}\}_{i,j=1}^n$ be a symmetric positive definite matrix. Define

$$(2\pi)^{-n/2}\left|\det r\right|^{-1/2}\int_0^\infty\cdots\int_0^\infty e^{-x^Tr^{-1}x/2}dx=:\int_0^\infty\cdots\int_0^\infty g(x,r)dx=:f(r)$$

In the special case that $r_{ij}=EX_iX_j$, we have from the definition of the mulvariate normal,

$$P(X_1>0,\ldots,X_n>0)=(2\pi)^{-n/2}\left|\det r\right|^{-1/2}\int_0^\infty\cdots\int_0^\infty e^{-x^Tr^{-1}x/2}dx$$

From Theorem 6.7,

$$\int e^{-i\langle y,x\rangle}e^{-y^Try/2}dy=\left|\det r\right|^{-1/2}(2\pi)^{n/2}e^{-x^Tr^{-1}x/2}$$

So, $g(x, r) = (2\pi)^n \int e^{-i\langle y,x\rangle} e^{-y^T ry/2} dy$. And for $i \neq j$, $\partial g/\partial r_{ij} = \partial^2 g/\partial x_i \partial x_j$. So, by differentiating under the integral sign and then integrating by parts,

$$\begin{aligned}\frac{\partial f}{r_{12}} &= \int_0^\infty \cdots \int_0^\infty \frac{\partial^2 g}{\partial x_1 \partial x_2} dx \\ &= \int_0^\infty \cdots \int_0^\infty g(0, 0, x_3, \ldots, x_n) dx_3 \cdots dx_n \geq 0 \qquad (*)\end{aligned}$$

For $\lambda \in [0, 1]$ and $\varepsilon > 0$, let $r_{ij} = \lambda EY_iY_j - (1 - \lambda)EX_iX_j + \varepsilon\delta_{ij}$. Then r is symmetric positive semidefinite, so $(*)$ and our assumption implies that

$$\frac{\partial f}{\partial \lambda} = \sum_{i \neq j} \frac{\partial f}{\partial r_{ij}} \frac{\partial r_{ij}}{\partial \lambda} = \sum_{i \neq j} \frac{\partial f}{\partial r_{ij}} (EY_iY_j - EX_iX_j) \geq 0$$

Integrating this inequality for $\lambda \in [0, 1]$ and then letting $\varepsilon \to 0$ concludes the theorem. □

Theorem 6.7. ***(Gaussian Fourier Transform)*** $\mathcal{F}(e^{-\pi|x|^2}) = e^{-\pi|\xi|^2}$

Proof. It suffices to prove the one dimensional identity $\int_{\mathbb{R}} e^{-\pi(x^2+2ixy)} dx = e^{-\pi y^2}$. In this case we write $\int_{\mathbb{R}} e^{-\pi(x^2+2ixy)} dx = e^{-\pi y^2} \int_{\mathbb{R}} e^{-\pi(x+iy)^2} dx$ and then shift the contour. □

Theorem 6.8. ***(Slepian's Inequality)*** *Let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ be n-dimensional Gaussian random vectors such that $EX_i = EY_i = 0$ for $i = 1, \ldots, n$, and $EX_i^2 = EY_i^2 = 1$ for $i = 1, \ldots, n$. Assume that $E(X_i - X_j)^2 \leq E(Y_i - Y_j)^2$ for $i, j \in \{1, \ldots, n\}$, $i \neq j$. Then for all $\alpha \in \mathbb{R}$,*

$$P(\sup_{i=1,\ldots,n} X_i > \lambda) \leq P(\sup_{i=1,\ldots,n} Y_i > \lambda)$$

In particular, $E \sup_{i=1,\ldots,n} X_i \leq E \sup_{i=1,\ldots,n} Y_i$.

Proof. By our assumption, $EX_iX_j \geq EY_iY_j$ for $i, j \in \{1, \ldots, n\}$, $i \neq j$. For $i = 1, \ldots, n$ let $f_i \colon \mathbb{R} \to [0, \infty)$ be a non-increasing smooth bounded function. Let $h(x) = \prod_{i=1}^n f_i(x_i)$. For $i \neq j$, $\partial^2 h/\partial x_i \partial x_j = f_i'(x_i) f_j'(x_j) \geq 0$. Let $(Z_1, \ldots, Z_n)$ be a mean zero Gaussian random vector with unit variances, and with positive definite covariance matrix r. Let $f(r) = Eh(Z_1, \ldots, Z_n)$. As in the proof of Slepian's Lemma, Theorem 6.6, for $i \neq j$,

$$\frac{\partial f}{\partial r_{ij}} = \int h(x) \frac{\partial g}{\partial r_{ij}} dx = \int \frac{\partial^2 h}{\partial x_i \partial x_j} g \geq 0.$$

We therefore conclude that $E \prod_{i=1}^n f_i(X_i) \geq E \prod_{i=1}^n f_i(Y_i)$. Fix $\lambda \in \mathbb{R}$. Let f_i approach $1_{(-\infty,\lambda]}$, so that $P(\sup_{i=1,\ldots,n} X_i < \lambda) \geq P(\sup_{i=1,\ldots,n} Y_i < \lambda)$, proving the theorem. □

Theorem 6.9. ***(Sudakov-Fernique Inequality*** [Cha05]***)*** *Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ be n-dimensional Gaussian random vectors such that $EX_i = EY_i = 0$ for $i = 1, \ldots, n$. Assume that $E(X_i - X_j)^2 \leq E(Y_i - Y_j)^2$ for $i, j \in \{1, \ldots, n\}$, $i \neq j$. Then*

$$E \sup_{i=1,\ldots,n} X_i \leq E \sup_{i=1,\ldots,n} Y_i$$

Proof. Let $g, X_1, \ldots, X_n$ be Gaussians with $Eg^2 = \tau^2$.

$$\begin{aligned} EgF(g) &= \frac{1}{\sqrt{2\pi}\tau}\int_{\mathbb{R}} te^{-t^2/(2\tau^2)}F(t)dt = \frac{\tau^2}{\sqrt{2\pi}\tau}\int_{\mathbb{R}}\left(-\frac{d}{dt}\right)e^{-t^2/(2\tau^2)}F(t)dt \\ &= \frac{\tau^2}{\sqrt{2\pi}\tau}\int_{\mathbb{R}} F'(t)e^{-t^2/(2\tau^2)}dt = Eg^2EF'(g) \end{aligned}$$

Let $X_i' := X_i - g\frac{EgX_i}{Eg^2}$. Then $EX_i'g = 0$ for $i = 1, \ldots, n$, so $(X_1', \ldots, X_n')$ is independent of g. So, conditioned on $(X_1', \ldots, X_n')$, we have

$$\begin{aligned} E(gF(X)|X_1', \ldots, X_n') &= Eg^2E\left(\frac{dF}{dg}|X_1', \ldots, X_n'\right) \\ &= \sum_{i=1}^n Eg^2E\left(\frac{\partial F}{\partial x_i}\frac{\partial x_i}{\partial g}|X_1', \ldots, X_n'\right) \\ &= \sum_{i=1}^n EgX_iE\left(\frac{\partial F}{\partial x_i}|X_1', \ldots, X_n'\right) \end{aligned}$$

So, by integrating out the conditioning,

$$EgF(X) = \sum_{i=1}^n EgX_iE\left(\frac{\partial F}{\partial x_i}\right)$$

In particular,

$$E(X_jF(X)) = \sum_{i=1}^n EX_iX_jE\left(\frac{\partial F}{\partial x_i}\right) \qquad (*)$$

Let $F(x) = F_\beta(x) := \frac{1}{\beta}\log(\sum_{i=1}^n e^{\beta x_i})$. We may assume that X and Y are independent. For $0 \le t \le 1$, let $Z := \sqrt{1-t}X + \sqrt{t}Y$, and let $f(t) := E(F_\beta(Z))$. By the chain rule, we have

$$f'(t) = E\sum_{i=1}^n \frac{\partial F}{\partial x_i}(Z)\left(\frac{Y_i}{2\sqrt{t}} - \frac{X_i}{2\sqrt{1-t}}\right)$$

Using $(*)$ and the chain rule,

$$E\left(\frac{\partial F}{\partial x_i}(Z)X_i\right) = \sqrt{1-t}\sum_{j=1}^n E(X_iX_j)E\left(\frac{\partial^2 F}{\partial x_i \partial x_j}(Z)\right)$$

$$E\left(\frac{\partial F}{\partial x_i}(Z)Y_i\right) = \sqrt{t}\sum_{j=1}^n E(Y_iY_j)E\left(\frac{\partial^2 F}{\partial x_i \partial x_j}(Z)\right)$$

So, combining these equalities,

$$f'(t) = \frac{1}{2}\sum_{i,j=1}^n E\left(\frac{\partial^2 F}{\partial x_i \partial x_j}(Z)\right)[E(Y_iY_j) - E(X_iX_j)]$$

Now, by the definition of F, $\partial F/\partial x_i =: p_i(x) = e^{\beta x_i}/(\sum_{i=1}^n e^{\beta x_i})$. So, for fixed x, the numbers $p_i(x)$, $i = 1, \ldots, n$ are nonnegative and sum to 1. Observe that

$$\frac{\partial^2 F}{\partial x_i \partial x_j}(x) = \begin{cases} \beta(p_i(x) - p_i(x)^2) & , i = j \\ -\beta p_i(x) p_j(x) & , i \neq j \end{cases}$$

So,

$$\sum_{i,j=1}^n \frac{\partial^2 F}{\partial x_i \partial x_j}(x)[E(Y_i Y_j) - E(X_i X_j)]$$
$$= \beta \sum_{i=1}^n p_i(x)[E(Y_i Y_i) - E(X_i X_i)] - \beta \sum_{i,j=1}^n p_i(x) p_j(x)[E(Y_i Y_j) - E(X_i X_j)]$$

Also, since $\sum_{i=1}^n p_i(x) = 1$, we have

$$\sum_{i=1}^n p_i(x)[E(Y_i Y_i) - E(X_i X_i)]$$
$$= \frac{1}{2} \sum_{i,j=1}^n p_i(x) p_j(x)[E(Y_i Y_i) - E(X_i X_i) + E(Y_j Y_j) - E(X_j X_j)]$$

So, combining these equalities,

$$\sum_{i,j=1}^n \frac{\partial^2 F}{\partial x_i \partial x_j}(x)[E(Y_i Y_j) - E(X_i X_j)]$$
$$= \frac{\beta}{2} \sum_{i,j=1}^n p_i(x) p_j(x)[(E(Y_i Y_i) + E(Y_j Y_j) - 2E(Y_i Y_j))$$
$$- (E(X_i X_i) + EX_j X_j - 2EX_i X_j)]$$
$$= \frac{\beta}{2} \sum_{i,j=1}^n p_i(x) p_j(x)[E(Y_i - Y_j)^2 - E(X_i - X_j)^2]$$

That is, $f'(t) \geq 0$ for all $0 \leq t \leq 1$. So,

$$E(F(Y)) = f(1) \geq f(0) = E(F(X)) \qquad (**)$$

Now,

$$\max_{i=1,\ldots,n} x_i = \beta^{-1} \log e^{\beta \max_{i=1,\ldots,n} x_i} \leq \beta^{-1} \log e^{\beta \sum_{i=1}^n x_i}$$
$$\leq \beta^{-1} \log(n e^{\beta \max_{i=1,\ldots,n} x_i}) = \beta^{-1} \log n + \max_{i=1,\ldots,n} x_i$$

That is, $\max_{i=1,\ldots,n} x_i \leq F_\beta(x) \leq \beta^{-1} \log n + \max_{i=1,\ldots,n} x_i$. So, letting $\beta \to \infty$ in $(**)$ concludes the theorem. □

Theorem 6.10. *(**Talagrand's Majorizing Measures Theorem**) Suppose* $(Y_1, \ldots, Y_n)$ *is a Gaussian random vector. Let* $X = \{1, \ldots, n\}$*. For* $i, j \in X$*, define* $d(i,j) := \sqrt{E(Y_i - Y_j)^2}$*. Let* $\mathcal{P}_X$ *be the set of Borel probability measures on* X*. Define*

$$\gamma_2(X,d) := \inf_{\mu \in \mathcal{P}_X} \sup_{x \in X} \int_0^\infty \sqrt{\log\left(\frac{1}{\mu(B(x,r))}\right)} dr$$

Then $c\gamma_2(X,d) \leq E\sup_{i \in X} Y_i \leq C\gamma_2(X,d)$*.*

6.2. Sub-Gaussian Processes.

Definition 6.11 (**Sub-Gaussian Random Variable**)**.** A real-valued random variable X is said to be **sub-Gaussian** if there exist $c_1, c_2 > 0$ such that

$$\mathbb{P}(|X| > t) \leq c_1 e^{-c_2 t^2}, \qquad \forall\, t > 0.$$

Exercise 6.12. Let X be a real-valued random variable with mean zero. Then the following are equivalent

- $\exists\ a > 0$ such that, for all $t \in \mathbb{R}$, $\mathbb{E}e^{tX} \leq e^{t^2 a^2/2}$.
- $\exists\ b > 0$ such that, for all $t > 0$, $\mathbb{P}(|X| > t) \leq 2e^{-bt^2}$.
- $\exists\ c > 0$ such that $\mathbb{E}e^{cX^2} \leq 2$.
- $\exists\ d > 0$ such that $(\mathbb{E}\,|X|^p)^{1/p} \leq d\sqrt{p}$, $\forall\ p \geq 1$.

(If you need hints look at Proposition 2.5.2 in Vershynin's book.)

For all $t \in \mathbb{R}$, define

$$\psi_2(t) := e^{t^2} - 1.$$

For a random variable X define

$$\|X\|_{\psi_2} := \inf\{t > 0 \colon \mathbb{E}\psi_2(|X|/t) \leq 1\}.$$

Exercise 6.13. Show that $\|\cdot\|_{\psi_2}$ is a norm on the set of sub-Gaussian random variables.

Proposition 6.14.

$$\|X - \mathbb{E}X\|_{\psi_2} \leq 3.2\,\|X\|_{\psi_2}.$$

Proof. From Exercise 6.13, $\|X - \mathbb{E}X\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E}X\|_{\psi_2}$. So, it remains to bound $\|\mathbb{E}X\|_{\psi_2}$. Since $\|1\|_{\psi_2} = 1/\sqrt{\log 2}$, Exercise 6.13 implies that $\|\mathbb{E}X\|_{\psi_2} = |\mathbb{E}X|/\sqrt{\log 2}$. So, by e.g. Jensen's inequality,

$$\|\mathbb{E}X\|_{\psi_2} \leq (\log 2)^{-1/2}\mathbb{E}\,|X| \leq \|X\|_{\psi_2}$$

The last inequality used $\mathbb{E}\psi_2(|X|/\|X\|_{\psi_2}) \leq 1$, implying that $\mathbb{P}(|X| > t) = \mathbb{P}(e^{cX^2} \leq e^{ct^2}) \leq e^{-ct^2}\mathbb{E}e^{cX^2} \leq e^{-ct^2}2$ when $c = 1/\|X\|_{\psi_2}^2$, so that

$$\mathbb{E}\,|X| = \int_0^\infty \mathbb{P}(|X| > t)dt \leq \int_0^\infty e^{-ct^2}2dt = c^{-1/2}\int_0^\infty e^{-t^2}2dt = \|X\|_{\psi_2}\sqrt{\pi}.$$

Finally, note that $1 + \sqrt{\pi/\log 2} \leq 3.2$ □

Hoeffding's Inequality 5.1 readily generalizes to the following.

Theorem 6.15 (General Hoeffding Inequality). *Let $X_1, X_2, \ldots$ be independent sub-Gaussian random variables. Then, for any $n \geq 1$,*

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i - \mathbb{E}\big(\sum_{j=1}^{n} X_j\big) \geq t\Big) \leq e^{-\frac{t^2}{16\sum_{i=1}^n \|X_i - \mathbb{E}X_i\|_{\psi_2}^2}} \leq e^{-\frac{t^2}{160\sum_{i=1}^n \|X_i\|_{\psi_2}^2}}, \qquad \forall\, t \geq 0.$$

Consequently,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} X_i - \mathbb{E}\big(\sum_{j=1}^{n} X_j\big)\Big| \geq t\Big) \leq 2e^{-\frac{t^2}{16\sum_{i=1}^n \|X_i - \mathbb{E}X_i\|_{\psi_2}^2}} \leq 2e^{-\frac{t^2}{160\sum_{i=1}^n \|X_i\|_{\psi_2}^2}}, \qquad \forall\, t \geq 0.$$

Proof.

$$\mathbb{P}(\sum_{i=1}^{n}[X_i - \mathbb{E}X_i] \geq t) = \mathbb{P}(e^{\alpha\sum_{i=1}^n [X_i - \mathbb{E}X_i]} \geq e^{\alpha t}) \leq e^{-\alpha t}\mathbb{E}e^{\alpha\sum_{i=1}^n [X_i - \mathbb{E}X_i]} = e^{-\alpha t}\prod_{i=1}^{n}\mathbb{E}e^{\alpha[X_i - \mathbb{E}X_i]}.$$

Using $e^x \leq x + e^{x^2}$,

$$\mathbb{E}e^{\alpha[X_i - \mathbb{E}X_i]} \leq \mathbb{E}[X_i - \mathbb{E}X_i] + \mathbb{E}e^{\alpha^2[X_i - \mathbb{E}X_i]^2} = \mathbb{E}e^{\alpha^2[X_i - \mathbb{E}X_i]^2}.$$

As in Proposition 6.14, $\mathbb{P}(|X_i - \mathbb{E}X_i| > t) \leq 2e^{-c_i t^2}2$ where $c_i = 1/\|X_i - \mathbb{E}X_i\|_{\psi_2}^2$, so if $\alpha \leq \sqrt{c_i}$,

$$\begin{aligned}\mathbb{E}e^{\alpha[X_i - \mathbb{E}X_i]} &\leq \mathbb{E}e^{\alpha^2[X_i - \mathbb{E}X_i]^2} = 1 + \int_0^\infty 2t\alpha^2 e^{\alpha^2 t^2}\mathbb{P}(|X_i - \mathbb{E}X_i| > t)dt \\ &\leq 1 + \int_0^\infty 4t\alpha^2 e^{\alpha^2 t^2} 2e^{-c_i t^2} dt = 1 + 2\alpha^2[c_i - \alpha^2]^{-1}\end{aligned}$$

So, if $\alpha^2 < c_i/2$, we have,

$$\mathbb{E}e^{\alpha[X_i - \mathbb{E}X_i]} \leq 1 + 2\frac{\alpha^2 2}{c_i} \leq 1 + \frac{4\alpha^2}{c_i} \leq e^{4\alpha^2/c_i}.$$

If $\alpha^2 > c_i/2$, then using $2\lambda x \leq \lambda^2/c + x^2 c$, we have by the first assertion

$$\mathbb{E}e^{\alpha[X_i - \mathbb{E}X_i]} \leq e^{2\alpha^2/c_i}\mathbb{E}e^{[X_i - \mathbb{E}X_i]^2 c_i/8} \leq e^{2\alpha^2/c_i}(1 + \frac{4c_i}{8c_i}) \leq e^{4\alpha^2/c_i},$$

since $2\alpha^2/c_i > 1$, so $2 \leq e$. In summary, recalling the definition of c_i,

$$\mathbb{P}(\sum_{i=1}^{n}[X_i - \mathbb{E}X_i] \geq t) \leq e^{-\alpha t}\prod_{i=1}^{n} e^{4\alpha^2\|X_i - \mathbb{E}X_i\|_{\psi_2}^2} = e^{-\alpha t}e^{4\alpha^2\sum_{i=1}^n \|X_i - \mathbb{E}X_i\|_{\psi_2}^2}$$

We then choose $\alpha := t/[8\sum_{i=1}^n \|X_i - \mathbb{E}X_i\|_{\psi_2}^2]$ to get

$$\mathbb{P}(\sum_{i=1}^{n}[X_i - \mathbb{E}X_i] \geq t) \leq e^{-\frac{t^2}{16\sum_{i=1}^n \|X_i - \mathbb{E}X_i\|_{\psi_2}^2}}.$$

Proposition 6.14 then implies the second inequality. □

Proposition 6.16. *Let $X_1, \ldots, X_k$ be independent mean zero sub-Gaussian random variables. Then*

$$\left\|\sum_{i=1}^{k} X_i\right\|_{\psi_2}^2 \leq 100 \sum_{i=1}^{k} \|X_i\|_{\psi_2}^2 .$$

In the case that X_i takes at most two values for all $1 \leq i \leq k$, we can replace 100 *with* $6 \log 2$.

Proof. Let $t > 0$. Using independence, $\mathbb{E}e^{t\sum_{i=1}^{k} X_i} = \prod_{i=1}^{k} \mathbb{E}e^{tX_i}$. Exercise 6.12 then concludes the proof, since $\prod_{i=1}^{k} \mathbb{E}e^{tX_i} \leq e^{ct^2 \sum_{i=1}^{k} \|X_i\|_{\psi_2}^2}$.

In the case that X_i takes at most two values for all $1 \leq i \leq k$, we have $X_i = a_i Z_i$ for some $a_i \in \mathbb{R}$ where $\mathbb{P}(Z_i = 1) = \mathbb{P}(Z_i = -1) = 1/2$ for all $1 \leq i \leq k$ so that by Exercise 5.2

$$\|X_i\|_{\psi_2}^2 = a_i^2 \|1\|_{\psi_2}^2 = a_i^2 / \log(2).$$

$$\mathbb{E}e^{\alpha X_i} = (1/2)(e^{\alpha a_i} + e^{-\alpha a_i}) = \cosh(\alpha a_i) \leq e^{\alpha^2 a_i^2/2}$$

$$\mathbb{E}e^{\alpha(\sum_{i=1}^{k} X_i)} = \prod_{i=1}^{k} \mathbb{E}e^{\alpha X_i} \leq e^{\alpha^2 \sum_{i=1}^{k} a_i^2/2}.$$

$$\mathbb{P}(\sum_{i=1}^{k} X_i > t) = \mathbb{P}(e^{\alpha(\sum_{i=1}^{k} X_i)} > e^{\alpha t}) \leq e^{-\alpha t} e^{\alpha^2 \sum_{i=1}^{k} a_i^2/2}$$

Choosing $\alpha := t/[\sum_{i=1}^{k} a_i^2]$ gives

$$\mathbb{P}(\sum_{i=1}^{k} X_i > t) \leq e^{-t^2/[2\sum_{i=1}^{k} a_i^2]}, \qquad \mathbb{P}(|\sum_{i=1}^{k} X_i| > t) \leq 2e^{-t^2/[2\sum_{i=1}^{k} a_i^2]}$$

$$\begin{aligned}\mathbb{E}e^{\alpha(\sum_{i=1}^{k} X_i)^2} &= 1 + \int_0^\infty 2t\alpha e^{\alpha t^2} \mathbb{P}(|\sum_{i=1}^{k} X_i| > t) dt \leq 1 + \int_0^\infty 4t\alpha e^{\alpha t^2} e^{-t^2/[2\sum_{i=1}^{k} a_i^2]} dt \\ &= 1 + \int_0^\infty 4t\alpha e^{t^2[\alpha - 1/[2\sum_{i=1}^{k} a_i^2]]} dt = 1 + 2\frac{-\alpha}{\alpha - 1/[2\sum_{i=1}^{k} a_i^2]}.\end{aligned}$$

This quantity is equal to 2 when $1/\alpha = 6\sum_{i=1}^{k} a_i^2$. So,

$$\left\|\sum_{i=1}^{k} X_i\right\|_{\psi_2} = \inf\{\beta > 0 \colon \mathbb{E}e^{\beta^{-2}(\sum_{i=1}^{k} X_i)^2} \leq 2\} \leq \sqrt{6\sum_{i=1}^{k} a_i^2}.$$

$$\left\|\sum_{i=1}^{k} X_i\right\|_{\psi_2}^2 \leq 6\sum_{i=1}^{k} a_i^2 = 6\log 2 \sum_{i=1}^{k} \|X_i\|_{\psi_2}^2 .$$

□

Definition 6.17 (Sub-Gaussian Increments). Let (A, d) be a metric space and let $\{X_a\}_{a\in A}$ be a random process. We say that $\{X_a\}_{a\in A}$ has **sub-Gaussian increments** if there exists $c > 0$ such that

$$\|X_a - X_b\|_{\psi_2} \leq c d(a, b), \qquad \forall\, a, b \in A.$$

Example 6.18. Let A be a set, let $\{X_a\}_{a\in A}$ be a Gaussian process, and *define* a metric d on the set A by

$$d(a,b) := \|X_a - X_b\|_{\psi_2}.$$

Exercise 6.19. Let $Y_1, Y_2, \ldots$ be a sequence of sub-Gaussian random variables. (These random variables are not assumed to be independent.) Prove that

$$\mathbb{E}\max_{i\geq 1}\frac{|Y_i|}{\sqrt{1+\log(i+1)}} \leq 100\sup_{i\geq 1}\|Y_i\|_{\psi_2}.$$

Conclude that, for any integer $n \geq 2$, we have

$$\mathbb{E}\max_{1\leq i\leq n}|Y_i| \leq 100\sqrt{\log n}\cdot\max_{1\leq i\leq n}\|Y_i\|_{\psi_2}.$$

(Hint: there are a few related ways to do this. Your first step could use the union bound of the form $\mathbb{P}(\max_{i\geq 1} X_i > t) \leq \sum_{i\geq 1}\mathbb{P}(X_i > t)$.)

(Optional: Show that you can replace 100 with 2 in both inequalities above.)

Recall the covering number $\mathcal{N}(A,d,\varepsilon)$ defined in Definition 4.5.

Theorem 6.20 (Dudley's Inequality). *Let (A,d) be a metric space and let $\{X_a\}_{a\in A}$ be a random process with sub-Gaussian increments and $\mathbb{E}X_a = 0$, $\forall$ $a \in A$. Then*

$$\mathbb{E}\sup_{a\in A}X_a \leq 12\sqrt{2}\,c\int_0^\infty\sqrt{\log\mathcal{N}(A,d,\varepsilon)}d\varepsilon.$$

Here $c := \sup_{a\in A}\|X_a - X_b\|_{\psi_2}/d(a,b)$.

Proof. Without loss of generality, we may assume that A is finite. Also, by scaling, we may assume that $c = 1$. For any integer k, let $\mathcal{N}_k$ be a 2^{-k}-net of A such that

$$|\mathcal{N}_k| = \mathcal{N}(A,d,2^{-k})$$

For any $a \in A$ and $\forall$ $k \in \mathbb{Z}$, let $b_k(a) \in \mathcal{N}_k$ be a point that is close to a, so that

$$d(a, b_k(a)) \leq 2^{-k}. \qquad (*)$$

(Such a $b_k(a)$ exists by definition of a 2^{-k}-net.)

Since A is finite, there exists an integers m, n and some $a_0 \in A$ such that $\mathcal{N}_m = \{a_0\}$ and $\mathcal{N}_n = A$. Consequently, for any $a \in A$, we have

$$b_m(a) = a_0, \qquad b_n(a) = a. \qquad (**)$$

Since $\mathbb{E}X_{a_0} = 0$ by assumption, we have

$$\mathbb{E}\sup_{a\in A}X_a = \mathbb{E}\sup_{a\in A}(X_a - X_{a_0}).$$

So, it suffices to bound the right side. To do so, we use a telescoping sum

$$X_a - X_{a_0} \overset{(**)}{=} X_{b_n(a)} - X_{b_m(a)} = \sum_{k=m}^{n-1}(X_{b_{k+1}(a)} - X_{b_k(a)}).$$

We then have

$$\mathbb{E}\sup_{a\in A}(X_a - X_{a_0}) = \mathbb{E}\sup_{a\in A}\sum_{k=m}^{n-1}(X_{b_{k+1}(a)} - X_{b_k(a)}) \leq \sum_{k=m}^{n-1}\mathbb{E}\sup_{a\in A}(X_{b_{k+1}(a)} - X_{b_k(a)}).$$

We now bound each individual term on the right. By Definition 6.17 with $c = 1$, and the triangle inequality,

$$\left\|X_{b_{k+1}(a)} - X_{b_k(a)}\right\|_{\psi_2} \leq d(b_{k+1}(a), b_k(a)) \leq d(b_{k+1}(a), a) + d(a, b_k(a)) \overset{(*)}{\leq} 2^{-(k+1)} + 2^{-k} = \frac{3}{2}2^{-k}.$$

The term $\sup_{a\in A}(X_{b_{k+1}(a)} - X_{b_k(a)})$ is a supremum over all pairs of points in $\mathcal{N}_{k+1} \times \mathcal{N}_k$, and

$$|\mathcal{N}_{k+1} \times \mathcal{N}_k| = |\mathcal{N}_{k+1}| \cdot |\mathcal{N}_k| \leq |\mathcal{N}_{k+1}|^2 .$$

Therefore, Exercise 6.19 implies that

$$\mathbb{E}\sup_{a\in A}(X_{b_{k+1}(a)} - X_{b_k(a)}) \leq (2) \cdot \frac{3}{2}2^{-k}\sqrt{\log|\mathcal{N}_{k+1}|^2} = 3 \cdot 2^{-k}\sqrt{2\log|\mathcal{N}_{k+1}|}.$$

In summary,

$$\begin{aligned}\mathbb{E}\sup_{a\in A} X_a &\leq 3\sqrt{2}\sum_{k=m}^{n-1} 2^{-k}\sqrt{\log|\mathcal{N}_{k+1}|} \\ &= 6\sqrt{2}\sum_{k=m}^{n-1} 2^{-k-1}\sqrt{\log|\mathcal{N}(A, d, 2^{-k-1})|} \leq 6\sqrt{2}\sum_{k\in\mathbb{Z}} 2^{-k-1}\sqrt{\log|\mathcal{N}(A, d, 2^{-k-1})|}.\end{aligned}$$

To turn the sum into an integral, note that $\left|\mathcal{N}(A, d, 2^{-k-1})\right| \leq |\mathcal{N}(A, d, \varepsilon)|$ for any $\varepsilon \leq 2^{-k-1}$, and $2^{-k-1} = 2\int_{2^{-k-2}}^{2^{-k-1}} d\varepsilon$, so

$$\begin{aligned}\sum_{k\in\mathbb{Z}} 2^{-k-1}\sqrt{\log|\mathcal{N}(A, d, 2^{-k-1})|} &= \sum_{k\in\mathbb{Z}} 2\int_{2^{-k-2}}^{2^{-k-1}} \sqrt{\log|\mathcal{N}(A, d, 2^{-k-1})|}d\varepsilon \\ &\leq \sum_{k\in\mathbb{Z}} 2\int_{2^{-k-2}}^{2^{-k-1}} \sqrt{\log|\mathcal{N}(A, d, \varepsilon)|}d\varepsilon = 2\int_0^\infty \sqrt{\log|\mathcal{N}(A, d, \varepsilon)|}d\varepsilon.\end{aligned}$$

□

Remark 6.21. The above proof of Dudley's inequality also proves the following. For a fixed $a_0 \in A$, we have

$$\mathbb{E}\sup_{a\in A}|X_a - X_{a_0}| \leq 12\sqrt{2}\,c\int_0^\infty \sqrt{\log\mathcal{N}(A, d, \varepsilon)}d\varepsilon.$$

Here we do not need to make a mean zero assumption as in Theorem 6.20.

Exercise 6.22. Using the argument for Dudley's inequality, deduce the following concentration inequality. For any $u > 0$,

$$\mathbb{P}\Big(\sup_{a\in A} X_a \leq 12\sqrt{2}\,c\int_0^\infty \sqrt{\log\mathcal{N}(A, d, \varepsilon)}d\varepsilon + u \cdot \mathrm{diam}(A)\Big) \geq 1 - 2e^{-u^2}.$$

Here $\mathrm{diam}(A) := \sup_{a,a'\in A} d(a, a')$. (Hint: show $\sup_{a\in A}\left|X_{b_{k+1}(a)} - X_{b_k(a)}\right| \leq 2^{-k}\sqrt{\log|\mathcal{N}_{k+1}|} + u_k$ with high probability at least $1 - 2e^{-u_k^2}$. Then sum over k, use the union bound, and choose the u_k appropriately, e.g. try $u_k = u + \sqrt{k-m}$.)

Exercise 6.22 is not optimal, since it does not demonstrate concentration of the empirical process around its mean. For such a concentration result, see Theorems 6.28 and 6.31 below.

In order to relate Dudley's inequality 6.20 back to statistical learning, we would like to replace the covering number by the VC-dimension, as in the following theorem.

For any $f, g \in \{-1, 1\}^A$, define the standard L_2-metric

$$d_2(f, g) = d_{2,\mathbb{P}}(f, g) := \|f - g\|_2 = (\mathbb{E}\,|f - g|^2)^{1/2}.$$

Theorem 6.23 (**Covering Number and VC-Dimension**). *Let $\mathcal{F} \subseteq \{0, 1\}^A$. Then, for every $\varepsilon > 0$,*

$$|\mathcal{N}(\mathcal{F}, d_2, \varepsilon)| \leq (2/\varepsilon)^{18\mathrm{VCdim}(\mathcal{F})}.$$

Lemma 6.24 (**Dimension-Reduction**). *Let $\mathbb{P}$ be a probability law on a set A. Let $\mathcal{F} \subseteq \{0, 1\}^A$ be a set of n functions. Assume that*

$$d_{2,\mathbb{P}}(f, g) > \varepsilon, \quad \forall\, f, g \in \mathcal{F},\ f \neq g.$$

Then $\exists\ m \leq 330\varepsilon^{-4}\log n$ and $\exists\ \{a_1, \ldots, a_m\} \in A$ such that, if $\mathbb{P}_m$ denotes the uniform probability law on $\{a_1, \ldots, a_m\}$, then

$$d_{2,\mathbb{P}_m}(f, g) > \varepsilon/2, \quad \forall\, f, g \in \mathcal{F},\ f \neq g.$$

(If we replace $> \varepsilon/2$ with > 0 in this inequality, we can instead choose $m \leq 200\varepsilon^{-4}\log n$.)

Proof. Let $X, X_1, \ldots, X_m$ be independent random variables taking values in A, each with distribution $\mathbb{P}$. Let $f, g \in \mathcal{F}$ with $f \neq g$. Let $h := (f - g)^2$. We will bound

$$\frac{1}{m}\sum_{i=1}^{m}(h(X_i) - \mathbb{E}h(X))$$

Recalling that $\|1\|_{\psi_2} = 1/\sqrt{\log 2}$,

$$\|h(X_i) - \mathbb{E}h(X)\|_{\psi_2} \leq \frac{1}{\sqrt{\log 2}}\|h(X_i) - \mathbb{E}h(X)\|_{\infty} \leq \frac{2}{\sqrt{\log 2}}.$$

The last inequality used $h \leq 2$. We now apply the General Hoeffding Inequality 6.15 to get

$$\mathbb{P}\Big(|\frac{1}{m}\sum_{i=1}^{m}(h(X_i) - \mathbb{E}h(X))| > 3\varepsilon^2/4\Big) \leq 2e^{-9m\varepsilon^4[\log 2]/1024}.$$

Consider the random subset of A defined to be $\{X_1, \ldots, X_m\}$ and let $\mathbb{P}_m$ denote the uniform probability law on this random subset. Then

$$d_{2,\mathbb{P}_m}(f, g)^2 - d_{2,\mathbb{P}}(f, g)^2 = \frac{1}{m}\sum_{i=1}^{m}(h(X_i) - \mathbb{E}h(X))$$

Then with probability at least $1 - 2e^{-9m\varepsilon^4[\log 2]/1024}$ (with respect to $\mathbb{P}$), we have

$$d_{2,\mathbb{P}_m}(f, g)^2 \geq d_{2,\mathbb{P}}(f, g)^2 - \frac{3\varepsilon^2}{4} \geq \varepsilon^2 - \frac{3\varepsilon^2}{4} = \frac{\varepsilon^2}{4}.$$

Taking the union bound over at most n^2 pairs of $(f, g) \in \mathcal{F} \times \mathcal{F}$, with probability at least $1 - 2n^2e^{-9m\varepsilon^4[\log 2]/1024}$ (with respect to $\mathbb{P}$), we have for *all* $f, g \in \mathcal{F}$ with $f \neq g$,

$$d_{2,\mathbb{P}_m}(f, g)^2 \geq \frac{1}{4}\varepsilon^2.$$

So, choosing $m \geq 1024[9\log 2]^{-1}\varepsilon^{-4}\log(2n^2)$ ensures that this probability is positive, i.e. there exist some points $a_1,\ldots,a_m$ satisfying our desired conclusion. $\square$

Proof of Theorem 6.23. Recalling the proof of Proposition 4.3, there exists a set $\mathcal{F}$ of $n \geq \mathcal{N}(A,d,\varepsilon)$ functions such that $d_2(f,g) \geq \varepsilon$ for all $f,g \in \mathcal{F}$ with $f \neq g$. Applying Lemma 6.24 to $\mathcal{F}$, we then get a set of points $A_m \subseteq A$ with $m := |A_m| \leq 200\varepsilon^{-4}\log n$ as in that Lemma. Denote $\mathcal{F}_m := \mathcal{F}|_{A_m}$. Since all $f,g \in \mathcal{F}_m$ have positive distance between each other, all functions in $\mathcal{F}_m$ distinct. By the Sauer-Shelah Lemma 4.8, if $d_m := \mathrm{VCdim}(\mathcal{F}_m)$,

$$n \leq (em/d_m)^{d_m} \leq (200\varepsilon^{-4}\log n/d_m)^{d_m} \leq (400\varepsilon^{-4}\log(n^{1/(2d_m)}))^{d_m} \leq (400\varepsilon^{-4})^{d_m}n^{1/2}.$$

That is,

$$n \leq (400\varepsilon^{-4})^{2d_m} \leq (2^9\varepsilon^{-4})^{2\mathrm{VCdim}(\mathcal{F})}.$$

$\square$

Lemma 6.25 (Symmetrization). *Let $Z_1,\ldots,Z_k$ be i.i.d. random variables independent of $X_1,\ldots,X_k$ with $\mathbb{P}(Z_1=1)=\mathbb{P}(Z_1=-1)=1/2$. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^k f(X_i)-\mathbb{E}f(X_1)\right| \leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^k Z_if(X_i)\right|.$$

Proof. Let $X_1',\ldots,X_k'$ be i.i.d. random variables, independent of all other random variables $X_1,\ldots,X_k,Z_1,\ldots,Z_k$ such that X_1 and X_1' have the same distribution. Note that moving the supremum inside an expected value can only increase its value. So, using Jensen's inequality

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^k f(X_i)-\mathbb{E}f(X_1)\right| &= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^k[f(X_i)-\mathbb{E}f(X_i)]\right|\\ &= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^k[f(X_i)-\mathbb{E}f(X_i)-\mathbb{E}[f(X_i')-\mathbb{E}f(X_i')]]\right|\\ &\leq \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^k[f(X_i)-\mathbb{E}f(X_i)-f(X_i')+\mathbb{E}f(X_i')]\right|\\ &= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^k[f(X_i)-f(X_i')]\right|\end{aligned}$$

For any $1 \leq i \leq k$, $f(X_i)-f(X_i')$ and $Z_i[f(X_i)-f(X_i')]$ have the same distribution, since $\mathbb{P}(f(X_i)-f(X_i')>t)=\mathbb{P}(f(X_i)-f(X_i')<-t)$, so

$$\begin{aligned}\mathbb{P}(Z_i[f(X_i)-f(X_i')]>t) &= (1/2)\mathbb{P}(f(X_i)-f(X_i')>t)+(1/2)\mathbb{P}(f(X_i)-f(X_i')<-t)\\ &= \mathbb{P}(f(X_i)-f(X_i')>t).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k}f(X_i)-\mathbb{E}f(X_1)\right| &\leq \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k}Z_i[f(X_i)-f(X_i')]\right|\\ &\leq \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k}Z_if(X_i)\right|+\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k}Z_if(X_i')\right| = 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k}Z_if(X_i)\right|.\end{aligned}$$

□

We can finally combine all of the tools above to give bounds for empirical processes in terms of VC-dimension.

Theorem 6.26 (**Empirical Risk and VC-Dimension**). *Let A be a set. Let $\mathbb{P}$ be a probability law on A. Let $\mathcal{F}\subseteq\{0,1\}^A$. Assume that $\mathrm{VCdim}(\mathcal{F})\geq 1$. Let $X_1,\ldots,X_k$ be independent random variables with distribution $\mathbb{P}$. Then for all $k\geq 1$.*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k}f(X_i)-\mathbb{E}f(X_1)\right|\leq 610\sqrt{\frac{1+\mathrm{VCdim}(\mathcal{F})}{k}}.$$

Proof. By Lemma 6.25

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k}f(X_i)-\mathbb{E}f(X_1)\right|\leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{k}\sum_{i=1}^{k}Z_if(X_i)\right|.$$

For any $f\in\mathcal{F}$, define $Y_f:=\frac{1}{\sqrt{k}}\sum_{i=1}^{k}Z_if(X_i)$.

For the remainder of the proof, we condition on $X_1,\ldots,X_k$ without explicitly denoting this conditioning. We would like to apply Dudley's inequality for the process $\{Y_f\}_{f\in\mathcal{F}}$. To do so, we need to check for sub-Gaussian increments. Proposition 6.16 and $\|Z_1\|_{\psi_2}=(\log 2)^{-1/2}$ implies that, for any $f,g\in\mathcal{F}$,

$$\|Y_f-Y_g\|_{\psi_2}=\left\|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}Z_i[f(X_i)-g(X_i)]\right\|_{\psi_2}\leq 6\Big(\frac{1}{k}\sum_{i=1}^{k}[f(X_i)-g(X_i)]^2\Big)^{1/2}.$$

Consider the set $A=\{1,\ldots,k\}$ and let $\mathbb{P}_k$ be uniform on A. We then have

$$\|Y_f-Y_g\|_{\psi_2}\leq 6\cdot d_{2,\mathbb{P}_k}(f,g),\qquad \forall\, f,g\in\mathcal{F}.$$

Dudley's Inequality, Theorem 6.20 (the sum version), and Theorem 6.23 imply that

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}Y_f &\leq 6\cdot 6\sqrt{2}\sum_{j=0}^{\infty}2^{-j-1}\sqrt{\log\mathcal{N}(\mathcal{F},d_{2,\mathbb{P}_k},2^{-j-1})}\\ &\leq 6\cdot 6\sqrt{2}\sum_{j=0}^{\infty}2^{-j-1}\sqrt{18\mathrm{VCdim}(\mathcal{F})\log\left(\frac{2}{2^{-j-1}}\right)}\\ &\leq 3\cdot 12\cdot 2\cdot\sqrt{9}\cdot\sqrt{\log 2}\cdot\sqrt{\mathrm{VCdim}(\mathcal{F})}\sum_{j=0}^{\infty}2^{-j-1}\sqrt{j+2}\\ &\leq 3\cdot 12\cdot 2\cdot\sqrt{9}\cdot\sqrt{\log 2}\cdot(1.7)\cdot\sqrt{\mathrm{VCdim}(\mathcal{F})}.\end{aligned}$$

The sum starts at $j = 0$ since $\sup_{f,g\in\mathcal{F}} d_{2,\mathbb{P}_k}(f,g) \leq 1$ so $\mathcal{N}(\mathcal{F}, d_{2,\mathbb{P}_k}, 2^{-j-1}) = 1$ for all $j \leq -1$. Also, we used $\sum_{j=0}^{\infty} 2^{-j-1}\sqrt{j+2} \leq 1.695$. The result follows (without the absolute value sign), using $3 \cdot 12 \cdot 2 \cdot \sqrt{9} \cdot \sqrt{\log 2} \cdot (1.695) \leq 305$. To get the same result with the absolute value sign, we use Remark 6.21 to get

$$\mathbb{E}\sup_{f\in\mathcal{F}} |Y_f| \leq \mathbb{E}\sup_{f\in\mathcal{F}\cup\{0\}} |Y_f| = \mathbb{E}\sup_{f\in\mathcal{F}\cup\{0\}} |Y_f - Y_0| \leq 305\sqrt{\mathrm{VCdim}(\mathcal{F}\cup\{0\})}.$$

We then note that $\mathrm{VCdim}(\mathcal{F}\cup\{0\}) \leq 1 + \mathrm{VCdim}(\mathcal{F})$, since adding a single function to $\mathcal{F}$ can only increase the VC-dimension by at most 1. □

Remark 6.27. By using Exercise 6.22 instead of Dudley's Inequality, we can turn Theorem 6.26 into a high probability statement: For any $u > 0$,

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} |Y_f| \leq 305\sqrt{1 + \mathrm{VCdim}(\mathcal{F})} + u\Big) \geq 1 - 2e^{-u^2}.$$

Therefore,

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} \left|\frac{1}{k}\sum_{i=1}^{k} f(X_i) - \mathbb{E}f(X_1)\right| \leq 610\frac{\sqrt{1 + \mathrm{VCdim}(\mathcal{F})}}{\sqrt{k}} + u\Big) \geq 1 - 2e^{-u^2 k}.$$

So, as in Proposition 6.4, if G is a random element of $\mathcal{F}$ that minimizes the empirical risk $ER_k(g) := \frac{1}{k}\sum_{i=1}^{k} 1_{g(X_i)\neq f(X_i)}$ among all $g \in \mathcal{F}$, and if $\overline{g}$ minimize the risk $r(g) := \mathbb{P}(g(X_1) \neq f(X_1))$ among all $g \in \mathcal{F}$, then with probability at least $1 - \delta$,

$$r(\overline{g}) \leq r(G) \leq r(\overline{g}) + 610\frac{\sqrt{\mathrm{VCdim}(\mathcal{F})}}{\sqrt{k}} + 2\frac{\sqrt{\log(2/\delta)}}{\sqrt{k}}.$$

6.3. **General Empirical Processes.** Bennett's Inequality 5.11

Theorem 6.28 (**Talagrand's Inequality**, [Tal96]). *There exists a constant $b > 0$ such that the following holds. Let $X_1, \ldots, X_k$ be independent random variables taking values in a measurable space A. Let $\mathcal{F}$ be a countable family of real-valued measurable functions on A. Define $u := \sup_{f\in\mathcal{F}} \|f\|_\infty$ and $v := \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{k}(f(X_i))^2$. Define*

$$Z := \sup_{f\in\mathcal{F}} \sum_{i=1}^{k} f(X_i)$$

Then, for any $t > 0$,

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq be^{-\frac{t}{bu}\log\left(1+\frac{tu}{v}\right)}.$$

Remark 6.29. Replacing $\mathcal{F}$ with $\mathcal{F} \cup (-\mathcal{F})$ in Theorem 6.28, we get the same inequality for

$$Z := \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^{k} f(X_i)\right|.$$

Also replacing $\mathcal{F}$ with $\{f - \mathbb{E}f(X_1) \colon f \in \mathcal{F}\}$, we get the same inequality for

$$Z := \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^{k}[f(X_i) - \mathbb{E}f(X_i)]\right|,$$

where now $v := \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{k}(f(X_i) - \mathbb{E}f(X_i))^2$.

Example 6.30. Suppose $\mathcal{F} \subseteq \{0,1\}^A$ or $\mathcal{F} \subseteq \{-1,1\}^A$ and we use $\widetilde{f} := \frac{1}{k} f$ in the Theorem. Then $u \leq 1/k$, $v \leq 1/k$, so for

$$Z := \sup_{f \in \mathcal{F}} \left| \frac{1}{k} \sum_{i=1}^{k} f(X_i) - \mathbb{E} f(X_i) \right|,$$

we have

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq b e^{-\frac{tk}{b} \log(1+t)}.$$

So, as in Proposition 6.4, if G is a random element of $\mathcal{F}$ that minimizes the empirical risk $ER_k(g) := \frac{1}{k} \sum_{i=1}^{k} 1_{g(X_i) \neq f(X_i)}$ among all $g \in \mathcal{F}$, and if $\overline{g}$ minimize the risk $r(g) := \mathbb{P}(g(X_1) \neq f(X_1))$ among all $g \in \mathcal{F}$, then with probability at least $1 - b e^{-\frac{tk}{b} \log(1+t)}$,

$$r(\overline{g}) \leq r(G) \leq r(\overline{g}) + 2t.$$

Conclusion. If the number of samples k is significantly larger than $\mathrm{VCdim}(\mathcal{F})$, then Empirical Risk Minimization is a reasonable thing to do. Recall that we made a similar observation in the Fundamental Theorem 4.17.

6.4. **Additional Comments.** Sharpened form of Talagrand's inequality, Theorem 6.28.

Theorem 6.31 (**Bousquet's Inequality**, [BLM13, Chapter 12], [Bou02])**.** *Let $X_1, \ldots, X_n$ be independent random vectors. Let $\mathcal{F}$ be a countable family of real-valued measurable functions on A. Assume that $\mathbb{E} f(X_1) = 0$ for all $f \in \mathcal{F}$ and $f(X_1) \leq 1$ for all $f \in \mathcal{F}$. Define*

$$Z := \sup_{f \in \mathcal{F}} \sum_{i=1}^{k} f(X_i)$$

Define $\sigma^2 := \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbb{E}(f(X_i))^2$, $w := 2\mathbb{E}Z + \sigma^2$. For any $s \geq -1$ define $h(s) := (1+s)\log(1+s) - s$ and $\phi(s) := e^s - s - 1$. Then, for any $t > 0$,

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq e^{-wh(t/w)}$$

In particular, $\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq e^{-t^2/(2(w+t/3))}$. Also, $\mathbb{E} e^{t(Z - \mathbb{E}Z)} \leq w\phi(t)$.

7. Sparse Recovery

7.1. **Geometric Inequalities. Notation**. In this section, when $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ and when $0 < p < \infty$, we denote

$$\|y\|_p := \Big(\sum_{i=1}^{n} |y_i|^p \Big)^{1/p}.$$

We also denote $\|y\|_\infty := \max_{1 \leq i \leq n} |y_i|$ and $\|y\| := \|y\|_2$. Note that if $0 < p < 1$, $\|\cdot\|_p$ is not a norm since the triangle inequality does not hold (i.e. the unit ball $\{y \in \mathbb{R}^n \colon \|y\|_p \leq 1\}$ is not convex.)

Remark 7.1. Recall that if $1 \leq p \leq \infty$ and if $1 \leq p' \leq \infty$ satisfies $1/p + 1/p' = 1$ (where we interpret $1/\infty$ as 0), then

$$\|y\|_p = \sup_{a \in \mathbb{R}^n \colon \|a\|_{p'} \leq 1} \langle a, y \rangle.$$

The right side is at most the left side by Hölder's inequality, and the right side can be seen to be equal to the left by choosing $a = (a_1, \ldots, a_n)$ so that $a_i := \mathrm{sign}(y_i) \left|y_i\right|^{p-1} \left\|y\right\|_p^{1-p}$ for all $1 \leq i \leq n$ and noting that $p' + p = p'p$, so $p'(p-1) = p$ and $1 + p/p' - p = 0$, so

$$\left\|a\right\|_{p'} = \left\|y\right\|_p^{1-p} (\sum_{i=1}^{n} \left|y_i\right|^{p'(p-1)})^{1/p'} = \left\|y\right\|_p^{1-p} (\sum_{i=1}^{n} \left|y_i\right|^{p})^{1/p'} = \left\|y\right\|_p^{1-p} \left\|y\right\|_p^{p/p'} = \left\|y\right\|_p^{1-p+p/p'} = 1.$$

Theorem 7.2 (**Hölder's Inequality**). *Let $x, y \in \mathbb{R}^n$. Then*

$$\left|\langle x, y\rangle\right| \leq \left\|x\right\|_p \left\|y\right\|_{p'}$$

Proof. By scaling, we may assume $\left\|x\right\|_p = \left\|y\right\|_{p'} = 1$ (zeros and infinities being trivial). Also, the case $p = 1, p' = \infty$ follows from the triangle inequality, so we assume $1 < p < \infty$. From concavity of the log function, we have the pointwise inequality

$$\left|x_i y_i\right| = (\left|x_i\right|^p)^{1/p} (\left|y_i\right|^q)^{1/q} \leq \frac{1}{p} \left|x_i\right|^p + \frac{1}{q} \left|y_i\right|^q, \qquad \forall\, 1 \leq i \leq n$$

which upon summation gives the result. □

Theorem 7.3 (**Triangle Inequality**). *Let $x, y \in \mathbb{R}^n$ and let $1 \leq p \leq \infty$. Then*

$$\left\|x + y\right\|_p \leq \left\|x\right\|_p + \left\|y\right\|_p.$$

Proof. The case $p = \infty$ follows from the scalar triangle inequality, so assume $1 \leq p < \infty$. By scaling, we may assume $\left\|x\right\|_p = 1 - t$, $\left\|y\right\|_p = t$, for some $t \in (0, 1)$ (zeros and infinities being trivial). Define $v := x/(1-t)$, $w := y/t$. Then by convexity of $z \mapsto \left|z\right|^p$ on $\mathbb{R}$,

$$\left|(1-t)v_i + t w_i\right|^p \leq (1-t) \left|v_i\right|^p + t \left|w_i\right|^p, \qquad \forall\, 1 \leq i \leq n$$

which upon summation completes the proof. □

A random (column) vector $X \in \mathbb{R}^n$ is called **isotropic** if $\mathbb{E}XX^T$ is the $n \times n$ identity matrix. A random vector $X \in \mathbb{R}^n$ is called **sub**-Gaussian if, for any $x \in \mathbb{R}^n$, the random variable $\langle x, X\rangle$ is sub-Gaussian. We then define

$$\left\|X\right\|_{\psi_2} := \sup_{x \in \mathbb{R}^n \colon \left\|x\right\| = 1} \left\|\langle X, x\rangle\right\|_{\psi_2}.$$

Theorem 7.4 (**Matrix-Deviation Inequality**). *Let M be an $m \times n$ random matrix. For any $1 \leq i \leq m$, let $M^{(i)}$ denote the i^{th} row of M, and assume that $M_1, \ldots, M_m$ are independent, isotropic and sub-Gaussian. Then for any $A \subseteq \mathbb{R}^n$,*

$$\mathbb{E} \sup_{x \in A} \left|\left\|Mx\right\| - \sqrt{m} \left\|x\right\|\right| \leq c (\max_{1 \leq i \leq m} \left\|M^{(i)}\right\|_{\psi_2})^2 \gamma(A),$$

where $\gamma(A) := \mathbb{E} \sup_{x \in A} \left|\langle x, G\rangle\right|$, where G is a standard Gaussian random vector in $\mathbb{R}^n$. Moreover, for any $u > 0$

$$\mathbb{P}(\sup_{x \in A} \left|\left\|Mx\right\| - \sqrt{m} \left\|x\right\|\right| \leq c (\max_{1 \leq i \leq m} \left\|M^{(i)}\right\|_{\psi_2})^2 [w(A) + u \cdot r(A)]) \geq 1 - 2e^{-u^2},$$

where $r(A) := \sup_{x \in A} \left\|x\right\|$ and $w(A) := \mathbb{E} \sup_{x \in A} \langle x, G\rangle$.

Remark 7.5. In general, $w(A) \leq \gamma(A)$. If A is a single nonzero point then $w(A) = 0$ and $\gamma(A) > 0$. If $A = -A$, then $w(A) = \gamma(A)$. For any $A, B \subseteq \mathbb{R}^n$, we have $w(A+B) = w(A) + w(B)$. Also, for any A we have

$$\begin{aligned} w(A) &= ((1/2)A + (1/2)A) = (1/2)w(A) + (1/2)w(A) \\ &= (1/2)w(A) + (1/2)w(-A) = (1/2)w(A-A) \end{aligned}$$

So, since $-(A-A) = A-A$, we have $\gamma(A-A) = w(A-A) = 2w(A) \leq 2\gamma(A)$.

Example 7.6. Suppose $\{m_{ij}\}_{1\leq i\leq m, 1\leq j\leq n}$ is a set of standard i.i.d. Gaussian random variables. Then for any $x \in \mathbb{R}^n$, we have

$$\mathbb{E}\left\|Mx\right\|^2 = \mathbb{E}\sum_{i=1}^{m}\sum_{j=1}^{n}(m_{ij}x_j)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}x_j^2\mathbb{E}m_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}x_j^2 = \sum_{i=1}^{m}\left\|x\right\|^2 = m\left\|x\right\|^2.$$

So, we anticipate that typically $\|Mx\| \approx \sqrt{m}\,\|x\|$, and this is the content of Theorem 7.4.

A Different Proof of Theorem 5.6. Let $B := \{x^{(1)}, \ldots, x^{(k)}\} \subseteq \mathbb{R}^n$. Let

$$A := \left\{\frac{x^{(i)} - x^{(j)}}{\left\|x^{(i)} - x^{(j)}\right\|} : 1 \leq i < j \leq k\right\}.$$

From Exercise 6.19, $w(A) \leq \gamma(A) \leq 10^3\sqrt{\log k}$. Also, $r(A) = 1$. Let M be an $m \times n$ matrix of i.i.d. standard Gaussian random variables. Theorem 7.4 implies that, with high probability,

$$\sup_{x,y\in B}\left|\frac{\|Mx - My\|}{\|x-y\|} - \sqrt{m}\right| \leq c \cdot 10^4\sqrt{\log k}.$$

That is, with high probability,

$$\left|\left\|\frac{M}{\sqrt{m}}x - \frac{M}{\sqrt{m}}y\right\| - \|x-y\|\right| \leq c\cdot 10^4\sqrt{\frac{\log k}{m}}\left\|x-y\right\|, \qquad \forall\, x,y \in B.$$

Choosing $m \geq c^{-1}10^{-4}\varepsilon^{-2}\log k$, we have, with high probability,

$$\left|\left\|\frac{M}{\sqrt{m}}x - \frac{M}{\sqrt{m}}y\right\| - \|x-y\|\right| \leq \varepsilon\left\|x-y\right\|, \qquad \forall\, x,y \in B.$$

So, there exists a matrix $M/\sqrt{m}$ satisfying the conclusion of Theorem 5.6. □

Theorem 7.7 (M^* **Bound**). *Let $A \subseteq \mathbb{R}^n$. Let M be a random $m \times n$ matrix with independent, isotropic and sub-Gaussian rows. Then the random subspace* $\ker(M)$ *satisfies*

$$\mathbb{E}\,\mathrm{diam}(A \cap \ker(M)) \leq \frac{c(\max_{1\leq i\leq m}\left\|M^{(i)}\right\|_{\psi_2})^2 w(A)}{\sqrt{m}}.$$

More generally,

$$\mathbb{E}\sup_{z\in\mathbb{R}^n}\mathrm{diam}(A \cap [z + \ker(M)]) \leq \frac{c(\max_{1\leq i\leq m}\left\|M^{(i)}\right\|_{\psi_2})^2 w(A)}{\sqrt{m}}.$$

Proof. Let $B := A - A \in \mathbb{R}^n$. Theorem 7.4 for the set B says that

$$\mathbb{E}\sup_{x,y\in A}\big|\|Mx - My\| - \sqrt{m}\,\|x-y\|\big| \leq c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^2\gamma(A{-}A) = 2c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^2\gamma(A),$$

the last equality following from Remark 7.5. Restricting the supremum on the left to a smaller subset can only decrease the expected value. If we restrict to $x, y \in A\cap\ker(M)$, then $Mx = My = 0$, so we get

$$\mathbb{E}\sup_{x,y\in A\cap\ker(M)}\sqrt{m}\,\|x-y\| \leq 2c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^2\gamma(A).$$

To get the more general statement, note: if $x, y \in A\cap(z+\ker(M))$, then $M(x-y) = 0$. □

Example 7.8. Suppose $A = \{x \in \mathbb{R}^n\colon \|x\| \leq 1\}$ is the unit ball centered at the origin. If $m < n$, then M will have nontrivial null space, so $\mathrm{diam}(A\cap\ker(M)) = 2$. In the case that m is close to n, the right side of Theorem 7.7 is not so far from the left side since $w(A) = \mathbb{E}_{\sup_{a\in A}}\langle G, a\rangle = \mathbb{E}\,\|G\| \approx \sqrt{n}$, where G is a standard n-dimensional Gaussian random vector, using also Remark 7.1. That is,

$$\frac{c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^2 w(A)}{\sqrt{m}} \approx \sqrt{\frac{n}{m}}.$$

On the other hand, if m is much smaller than n, then this quantity is quite far from the right side's value of 2.

Theorem 7.9 (**Escape Theorem**). *Let $A \subseteq S^{n-1} = \{x \in \mathbb{R}^n\colon \|x\| = 1\}$. Let M be a random $m\times n$ matrix with independent, isotropic and sub-Gaussian rows. If*

$$m > 4c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^4[w(A)]^2,$$

then

$$\mathbb{P}(A\cap\ker(M) = \emptyset) \geq 1 - 2\exp\Big(-\frac{cm}{4\max_{1\leq i\leq m}\|M^{(i)}\|_{\psi_2}^4}\Big).$$

Proof. Note that $r(A) \leq 1$ and $\|x\| = 1$ for all $x \in A$. Theorem 7.4 implies that, with probability at least $1 - 2e^{-u^2}$,

$$\sup_{x\in A}\big|\|Mx\| - \sqrt{m}\big| \leq c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^2(w(A) + u).$$

If this event occurs and $A\cap\ker(M) \neq \emptyset$, let $x \in A\cap\ker(M)$. Then $Mx = 0$, so the inequality reduces to

$$\sqrt{m} \leq c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^2(w(A) + u).$$

If we choose $u := \sqrt{m}[2c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^2]^{-1}$, we then get

$$\sqrt{m} \leq c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^2 w(A) + \frac{\sqrt{m}}{2}.$$

That is,

$$\sqrt{m} \leq 2c(\max_{1\leq i\leq m}\big\|M^{(i)}\big\|_{\psi_2})^2 w(A)$$

But this inequality contradicts the assumption of the Theorem. To avoid the contradiction, we conclude that $A\cap\ker(M) = \emptyset$ with probability at least $1 - 2e^{-u^2}$. □

Example 7.10. Suppose $A = \{x \in \mathbb{R}^n \colon \|x\| = 1\}$ is the unit ball centered at the origin. As in the previous example,

$$w(A) = \mathbb{E}_{\sup_{a \in A}} \langle G, a \rangle = \mathbb{E}\,\|G\| \approx \sqrt{n},$$

where we again used Remark 7.1, so

$$4c(\max_{1 \le i \le m} \left\|M^{(i)}\right\|_{\psi_2})^4 [w(A)]^2 \approx n.$$

So, in this case the Theorem has little content, since $m > n$ implies that M should have trivial nullspace with high probability.

If $m < n$, then M will have nontrivial null space, so the Theorem becomes interesting when A is not all of S^{n-1}. And in order to allow $m < n$, we also need $w(A)$ to be fairly small compared to n.

7.2. Applications.

Problem 7.11 (Signal Recovery Problem). Suppose we would like to find an unknown signal $x \in \mathbb{R}^n$. We do not have access to x, but we do have access to

$$Mx + w,$$

where M is a known $m \times n$ real matrix, and w is an unknown vector. The vector w is called the noise vector.

The goal is to recover x either exactly or approximately.

Consider the noiseless case when $w = 0$. Assume also that a set A is given, and it is known that $x \in A$. A first approximation to a solution of Problem 7.11 is

Algorithm 7.12. Input: Let $x \in \mathbb{R}^n$ be unknown and let $A \subseteq \mathbb{R}^n$ be a known set such that $x \in A$. Let M be a known $m \times n$ matrix.

Goal: Solve Problem 7.11 with $w = 0$.

Output: Let $x' \in A$ be any solution to the equation

$$Mx = Mx'.$$

If A is a convex set, this problem can be solved efficiently using convex programming. (For example, if A is a polytope, then this problem can be solved by linear programming.)

In case M is random, Algorithm 7.12 performs unexpectedly well in solving Problem 7.11 (when $w = 0$).

Theorem 7.13. *Let M be a random $m \times n$ matrix with independent, isotropic and sub-Gaussian rows. Let X be the (random) output of Algorithm 7.12. Then*

$$\mathbb{E}\,\|x - X\| \le \frac{c(\max_{1 \le i \le m} \left\|M^{(i)}\right\|_{\psi_2})^2 w(A)}{\sqrt{m}}$$

Proof. Since $X \in A$ and $Mx = MX$, we have $X \in A \cap (x + \mathrm{ker} M)$. Also, $x \in A \cap (x + \mathrm{ker} M)$, so $\|x - X\| \le \mathrm{diam}(A \cap [z + \ker(M)])$, and by Theorem 7.7,

$$\mathbb{E}\,\|x - X\| \le \mathbb{E} \sup_{z \in \mathbb{R}^n} \mathrm{diam}(A \cap [z + \ker(M)]) \le \frac{c(\max_{1 \le i \le m} \left\|M^{(i)}\right\|_{\psi_2})^2 w(A)}{\sqrt{m}}.$$

□

Remark 7.14. The bound in Theorem 7.13 is best when $w(A)$ is small.

In sparse recovery, we make the additional assumption that $x \in \mathbb{R}^n$ has many nonzero entries. Define

$$\|x\|_0 := |\{1 \leq i \leq n \colon x_i \neq 0\}| \,.$$

This quantity is not a norm, since for any $t > 0$, $\|tx\|_0 = \|x\|$. We say $x \in \mathbb{R}^n$ is **s-sparse** if $\|x\|_0 \leq s$. It is often realistic to assume that a signal x is s-sparse for a reasonably small s.

Unfortunately, since $A := \{y \in \mathbb{R}^n \colon \ \|y\|_0 \leq s\}$ is a non-convex set, Algorithm 7.12 is often computationally hard. So, in order to solve Problem 7.11 with such an A, we instead replace $\|\cdot\|_0$ with $\|\cdot\|_1$ in the definition of A. Since $\|x\|_0 = \lim_{p \to 0^+} \|x\|_p^p$, this is not unreasonable.

Note that, if $\|x\|_0 \leq s$, then if $z \in \mathbb{R}^n$ is defined by $z_i := \mathrm{sign}(x_i)$ for all $1 \leq i \leq n$, then

$$\|x\|_1 = \langle z, x \rangle \leq \|z\| \, \|x\| = \sqrt{\|x\|_0} \, \|x\| \leq \sqrt{s} \, \|x\| \,.$$

We therefore let

$$A := \{y \in \mathbb{R}^n \colon \ \|y\|_1 \leq \sqrt{s} \, \|y\|\}.$$

Corollary 7.15 (Approximate Sparse Recovery). *Let $s \geq 1$. Let $x \in \mathbb{R}^n$ with $\|x\| \leq 1$. Let M be a random $m \times n$ matrix with independent, isotropic and sub-Gaussian rows. Let X be the (random) output of Algorithm 7.12, where*

$$A := \{y \in \mathbb{R}^n \colon \ \|y\|_1 \leq \sqrt{s}\}.$$

Then

$$\mathbb{E} \, \|x - X\| \leq c(\max_{1 \leq i \leq m} \left\|M^{(i)}\right\|_{\psi_2})^2 \sqrt{\frac{s \log n}{m}}.$$

Proof. Apply Theorem 7.13 and Exercise 6.19 with Remark 7.1 to get

$$\begin{aligned} w(A) &= \mathbb{E} \sup_{a \in A} \langle G, a \rangle \leq \mathbb{E} \sup_{a \in \mathbb{R}^n \colon \, \|a\|_1 \leq \sqrt{s}} \langle G, a \rangle \\ &= \sqrt{s} \, \mathbb{E} \sup_{a \in \mathbb{R}^n \colon \, \|a\|_1 \leq 1} \langle G, a \rangle = \sqrt{s} \, \mathbb{E} \, \|G\|_\infty \leq 10 \sqrt{s \log n}. \end{aligned}$$

Here G is a standard n-dimensional Gaussian random vector. □

If m is much larger than $s \log n$, Algorithm 7.12 approximately solves Problem 7.11 well. It is important here that m (the "number of measurements") can be much smaller than n.

In fact, Corollary can be improved to guarantee *exact* recovery (with high probability), if we modify Algorithm 7.12 in the following way.

Algorithm 7.16. Input: Let $x \in \mathbb{R}^n$ be an unknown vector. Let M be a known $m \times n$ matrix.

Goal: Solve Problem 7.11 with $w = 0$.

Output: $x' \in \mathbb{R}^n$ minimizing $\|x'\|_1$ among all solutions $z \in \mathbb{R}^n$ of $Mx = Mz$.

Theorem 7.17 (Exact Sparse Recovery). *Let M be a random $m \times n$ matrix with independent, isotropic and sub-Gaussian rows. Let $k := \max_{1 \leq i \leq m} \left\|M^{(i)}\right\|_{\psi_2}$. Then with probability at least $1 - 2e^{-cm/k^4}$, the following holds.*

Let $s \geq 1$. Let $x \in \mathbb{R}^n$ be s-sparse, and suppose

$$m \geq ck^4 s \log n.$$

Let X be the (random) output of Algorithm 7.16. Then

$$X = x.$$

Proof. Let $h := X - x$. Let $S := \{1 \leq i \leq n \colon x_i \neq 0\}$. We first show that

$$\|h1_{S^c}\|_1 \leq \|h1_S\|_1 . \qquad (*)$$

By definition of X in Algorithm 7.16, $\|X\|_1 \leq \|x\|_1$. On the other hand, by the triangle inequality, and using $x1_S = x$ and $x1_{S^c} = 0$,

$$\begin{aligned}\|x\|_1 \geq \|X\|_1 &= \|x+h\|_1 = \|(x+h)1_S\|_1 + \|(x+h)1_{S^c}\|_1 \\ &\geq \|x1_S\|_1 - \|h1_S\|_1 + \|(x+h)1_{S^c}\|_1 = \|x\|_1 - \|h1_S\|_1 + \|h1_{S^c}\|_1 .\end{aligned}$$

Rearranging then proves $(*)$.

We now show that

$$\|h\|_1 \leq 2\sqrt{s}\,\|h\| . \qquad (**)$$

Indeed, if $z \in \mathbb{R}^n$ is defined by $z_i := \mathrm{sign}(h_i)1_S$ for all $1 \leq i \leq n$, then

$$\|h\|_1 = \|h1_S\|_1 + \|h1_{S^c}\|_1 \overset{(*)}{\leq} 2\,\|h1_S\|_1 = 2\langle h, z\rangle \leq 2\,\|h\|\,\|z\| \leq 2\sqrt{s}\,\|h\| .$$

Now, consider the event that $h \neq 0$. From $(*)$, we have

$$\frac{h}{\|h\|} \in A_s := \{y \in \mathbb{R}^n \colon\ \|y\| = 1,\ \|y\|_1 \leq 2\sqrt{s}\}.$$

By definition of h and X, $Mh = MX - Mx = 0$. That is,

$$\frac{h}{\|h\|} \in A_s \cap \ker(M).$$

Theorem 7.9 says that $A_s \cap \ker(M) = \emptyset$ with high probability, if $m \geq ck^4 w(A_s)^2$. As in the proof of Corollary 7.15,

$$w(A_s) = \mathbb{E} \sup_{a \in A_s} \langle G, a\rangle \leq 2\sqrt{s}\mathbb{E}\,\|G\|_\infty \leq 10\sqrt{s\log n}.$$

That is, if $m \geq ck^4 s\log n$, the intersection is empty with high probability, i.e. $h \neq 0$ with small probability, as desired. □

Remark 7.18. If $\|y\| = 1$, then $\|y\|_1 \leq \sqrt{n}\,\|y\|$ by Hölder's inequality, so A_s gives an interesting definition of sparsity when s is significantly less than n, and this is reflected in the assumption of the theorem.

7.3. **Additional Comments.** Similar analysis can be performed on the LASSO algorithm. See section 10.6 in [Ver18].

Algorithm 7.19 (LASSO). Input: Let $x \in \mathbb{R}^n$ be an unknown vector. Let $r > 0$. Let M be a known $m \times n$ matrix.

Goal: Solve Problem 7.11.

Output: $x' \in \mathbb{R}^n$ minimizing $\|Mx + w - Mx'\|$ among all $z \in \mathbb{R}^n$ with $\|z\|_1 \leq r$.

8. Deep Learning

Definition 8.1 (Feedforward Neural Network). A **feedforward neural network** with k layers and activation function $h\colon \mathbb{R} \to \mathbb{R}$ is a function f defined as follows. Let $n_0, \ldots, n_{k-1}$ be positive integers. For each $1 \leq i \leq k-1$, assume that

$$f_i\colon \mathbb{R}^{n_{i-1}} \to \mathbb{R}^{n_i},$$

$$f_k\colon \mathbb{R}^{n_{k-1}} \to \mathbb{R}.$$

Assume also that for all $1 \leq i \leq k$, there exists $w^{(i)} \in \mathbb{R}^{d_{i-1}}, t_{ij} \in \mathbb{R}$ such that the j^{th} component of f_i satisfies

$$f_{i,j}(x) = h(\langle w^{(i)}, x\rangle - t_{ij}), \qquad \forall\, x \in \mathbb{R}^{n_{i-1}}.$$

Then f is defined to be a function of the form

$$f := f_k \circ f_{k-1} \circ \cdots \circ f_1.$$

We refer to $\max_{0 \leq i \leq k-1} n_i$ as the **width** of the neural network. We also refer to k as the **depth** or **number of layers** of the neural network.

Note that if h is itself a linear function, then f is also a linear function. So, it is most sensible to choose a nonlinear activation function h.

Common examples of activation functions include:

- $h(x) = \mathrm{sign}(x)$ or $(1 + \mathrm{sign}(x))/2$, $\forall\, x \in \mathbb{R}$ (Boolean activation function).
- $h(x) = \max(x, 0)$, $\forall\, x \in \mathbb{R}$ (Rectified Linear Unit) (ReLU).
- $h(x) = (1 + e^{-2x})^{-1}$, $\forall\, x \in \mathbb{R}$ (Sigmoid/Logistic Function)
- $h(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, $\forall\, x \in \mathbb{R}$.

In the last case, note that $(1/2)[1 + \tanh(x)]$ is equal to the sigmoid function.

By Corollary 4.22, if $f_{ij} \in \mathcal{F}_{ij}$, if $\mathcal{F}_{ij}$ is a class of boolean functions, and if $d := \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathrm{VCdim}(\mathcal{F}_{ij})$, then

$$\mathrm{VCdim}(\mathcal{F}) \leq 10d \log(den).$$

So, Theorem 4.17 gives a (possibly inefficient) PAC learning algorithm for learning feedforward neural networks with boolean activation function. Similarly, Remark 6.27 demonstrates that Empirical Risk Minimization is a (possibly inefficient) algorithm for learning this function class.

8.1. Depth Separation.

Define $f\colon [0,1] \to [0,1]$ by

$$f(x) := \begin{cases} 2x & \text{, if } 0 \leq x \leq 1/2 \\ 2 - 2x & \text{, if } 1/2 \leq x \leq 1. \end{cases}$$

Let $f_1 := f$ and for any integer $m \geq 2$, define

$$f_m := f_{m-1} \circ f.$$

That is, f_m is a composition of m copies of f Then f_m consists of 2^{m-1} "spikes," or "sawteeth." In particular,

$$f_m(i2^{-m}) = \begin{cases} 0 & \text{, if } 1 \leq i \leq 2^m \text{ is even} \\ 1 & \text{, if } 1 \leq i \leq 2^m \text{ is odd.} \end{cases}$$

Let h denote the ReLU function, $h(x) := \max(x, 0)$ for all $x \in \mathbb{R}$. We can write f as a two-layer ReLU network with four nodes:

$$f(x) = h(2h(x) - 4h(x - 1/2)), \qquad \forall\, x \in [0, 1]$$

Consequently, for any $m \geq 1$, we can write f_m as a $2m$-layer ReLU network with $3m + 1$ nodes and width 2.

A natural question is: can we represent f_m using a ReLU network with small depth and width? The answer is: no.

Let $m \geq 1$. Suppose we are given $x_1, \ldots, x_m \in [0, 1]$ and $y_1, \ldots, y_m \in \{0, 1\}$. For any $g\colon [0, 1] \to [0, 1]$, define

$$ER_m(g) := \frac{1}{m}\sum_{i=1}^{m} 1_{1_{g(x_i)>1/2} \neq y_i}.$$

Theorem 8.2 ([Tel15]). *Let $(y_1, \ldots, y_{2^m}) := (0, 1, 0, 1, 0, \ldots)$ and let $x_i := i2^{-m}$ for all $1 \leq i \leq 2^m$. Note that $ER_{2^m}(f_m) = 0$. On the other hand, if a network g has k layers and the width w satisfying*

$$w < 2^{\frac{m-\ell}{k} - 1},$$

for some $\ell \geq 1$, then

$$ER_{2^m}(g) \geq \frac{1}{2} - \frac{1}{2^\ell}.$$

Proof. For a function $g\colon \mathbb{R} \to \mathbb{R}$ that is piecewise-affine, let $a(g)$ denote the (minimal) number of pieces needed to define g. For any piecewise-affine functions $g_1, g_2\colon \mathbb{R} \to \mathbb{R}$, we then have

$$a(g_1 + g_2) \leq a(g_1) + a(g_2), \qquad a(g_1 \circ g_2) \leq a(g_1) \cdot a(g_2). \qquad (*)$$

Suppose now that $g\colon [0, 1] \to [0, 1]$ is a depth k width w ReLU network. Since each layer of g involves at most w additions of the ReLU function, and the ReLU function is piecewise-affine with 2 pieces, we have

$$a(g) \leq (2w)^k \leq 2^{m-\ell}.$$

Suppose g is affine on a set of at most $2^{m-\ell}$ disjoint intervals $I_1, I_2, \ldots \subseteq [0, 1]$. For any $j \geq 1$, suppose the interval I_j contains p_j of the points $x_1, \ldots, x_m$. Since $y_1, y_2, \ldots$ oscillates between 0 and 1, the inequality $1_{g>1/2}(x_i) \neq y_i$ occurs for at least $\lceil (p_j - 2)/2 \rceil \geq (p_j - 2)/2$ of the points $x_1, \ldots, x_m$ in that interval. We can find the total number of times this inequality occurs by summing over j. The inequality $1_{g>1/2}(x_i) \neq y_i$ occurs in $[0, 1]$ at least

$$\sum_{j=1}^{2^{m-\ell}} (p_j - 2)/2 = 2^{m-1} - 2^{m-\ell} = 2^{m-1}[1 - 2^{-\ell+1}].$$

That is, $ER_{2^m}(g) \geq \frac{1}{2} - \frac{1}{2^\ell}$. □

Example 8.3. Suppose $k = \sqrt{m}$ and $\ell = 2$. Then any $\sqrt{m}$-layer network of width $w < 2^{\sqrt{m}-2}$ cannot approximate f_m well. That is, super-polynomial width is required to approximate f_m with $\sqrt{m}$ layers, even though f_m can be written exactly with a depth m network of width 2.

A logical circuit is a function $f\colon \{0, 1\}^n \to \{0, 1\}$ consisting of a nested composition of the following three functions:

- NOT: $\{0, 1\} \to \{0, 1\}$, defined by NOT(0) := 1 and NOT(1) := 0.

- AND: $\{0,1\}^2 \to \{0,1\}$, defined by $\mathrm{AND}(x,y) := 1_{x=y=1}$.
- OR: $\{0,1\}^2 \to \{0,1\}$, defined by $\mathrm{OR}(x,y) := 1 - 1_{x=y=0}$.

The notions of width and depth are defined for logical circuits as in the case of neural networks.

Theorem 8.4 ([HRST17]). *Fix $k \geq 1$. Then there exists a logical circuit $f\colon \{0,1\}^n \to \{0,1\}$ of depth k and size $O(n)$ such that any depth $k-1$ circuit that agrees with f on at least $(1/2)+\delta$ fraction of inputs from $\{0,1\}^n$ must have size $O(e^{n^{\Omega(1/k)}})$.*

Theorem 8.5 ([ES16]). *Suppose the activation function h of a neural network satisfies the following. There exists $c_0, \alpha > 0$ such that*

- *$|h(x)| \leq c_0(1+|x|)^\alpha$ for all $x \in \mathbb{R}$, and*
- *For any $\ell > 0, r > 0$, for any ℓ-Lipschitz $f\colon \mathbb{R} \to \mathbb{R}$ that is constant in $[-r,r]^c$, for any $\delta > 0$, there exists real numbers $w \leq cr\ell/\delta$, a, $\{\alpha_i, \beta_i, \gamma_i\}_{i=1}^w$ such that*

$$\left\|f(x) - [a + \sum_{i=1}^{w} \alpha_i h(\beta_i x - \gamma_i)]\right\|_\infty < \delta.$$

That is, a two-layer network can approximate Lipschitz functions well.

Then there exist universal constants $c, c' > 0$ such that the following holds. For any $n \geq 1$, there exists a probability measure $\mathbb{P}$ on $\mathbb{R}^n$ and there exists $f\colon \mathbb{R}^n \to [-1,1]$ supported in $B(0, c'\sqrt{n})$ such that

- *f can be written as a 3-layer neural network of width $O(n^5)$, and*
- *Any 2-layer neural network g of width at most ce^{cn} satisfies*

$$\mathbb{E}\,|f-g|^2 \geq c.$$

Remark 8.6. The activation functions mentioned above (boolean, ReLU, and sigmoid) all satisfy the hypothesis of the Theorem [ES16].

Rough Sketch of the Proof. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a smooth function with compact support. For any $y \in \mathbb{R}^n$, define the Fourier transform of f by

$$\widehat{f}(y) := \int_{\mathbb{R}^n} f(x) e^{-2\pi i \langle x,y\rangle} dx.$$

In the case $n=1$, note that the constant function $f := 1$ satisfies $\widehat{f}(y) = 0$ for all $y \neq 0$ (in the distributional sense). More specifically, $\widehat{f}$ is equal (in the distributional sense) to the probability measure that assigns mass 1 to the origin 0. (That is, for any smooth compactly supported function $\phi\colon \mathbb{R} \to \mathbb{R}$, we have $\langle \widehat{f}, \phi\rangle := \langle f, \widehat{\phi}\rangle = \int_{\mathbb{R}} \widehat{\phi}(y)dy = \phi(0)$; the last equality follows from the Fourier Inversion formula $\phi(x) = \int_{\mathbb{R}^n} \widehat{f}(y) e^{2\pi i\langle x,y\rangle} dy$.) More generally, if $f\colon \mathbb{R}^n \to \mathbb{R}$, and f can be written as $f(x_1, \ldots, x_n) = f_1(x_1)$ for some $f_1\colon \mathbb{R} \to \mathbb{R}$, for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$, then (in the distributional sense), $\widehat{f}(y) = 0$ for all $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ that do not lie on the y_1-axis.

More generally, if $g\colon \mathbb{R}^n \to \mathbb{R}$, and if g can be written as $g(x) = \sum_{i=1}^{j} g_i(\langle v_i, x\rangle)$ for some $g_1, \ldots, g_j\colon \mathbb{R} \to \mathbb{R}$, $v_1, \ldots, v_j \in \mathbb{R}^n$, for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$, then (in the distributional sense), $\widehat{g}(y) = 0$ except on the union of the lines parallel to each of $v_1, \ldots, v_j$, i.e.

$\cup_{i=1}^{j}\mathrm{span}(v_i)$. From Plancherel's Formula, if $\mathbb{P}$ has density ϕ^2 on $\mathbb{R}^n$, then

$$\begin{aligned}\mathbb{E}\,|f-g|^2 &= \int_{\mathbb{R}^n} |f(x)-g(x)|^2\,|\phi(x)|^2\,dx = \int_{\mathbb{R}^n} |f(x)\phi(x)-g(x)\phi(x)|^2\,dx\\ &= \int_{\mathbb{R}^n} |\widehat{f\phi}(x)-\widehat{g\phi}(x)|^2 dx.\end{aligned}$$

The proof uses ϕ such that $\widehat{\phi} := 1_{B(0,r_n)}$ where $r_n > 0$ is chosen so that $\int_{B(0,r_n)} dx = 1$. Since $\widehat{g}$ is supported in the lines $\cup_{i=1}^{j}\mathrm{span}(v_i)$, and since $\widehat{g\phi} = \widehat{g} * \widehat{\phi}$, we have that $\widehat{g\phi}$ is supported in the tubes

$$\cup_{i=1}^{j}[\mathrm{span}(v_i) + B(0,r_n)].$$

So, in order to complete the proof, it suffices to find a function f such that $\widehat{f}$ (and $\widehat{f\phi} = \widehat{f} * \widehat{\phi}$) have most of their mass supported away from *any* tubes of the form $\cup_{i=1}^{j}[\mathrm{span}(v_i)] + B(0,r_n)$. This is accomplished by choosing f to be radial (i.e. f is only a function of $\|x\|$), so that $\widehat{f}$ is also radial. Since f is radial, one can show that a 3-layer neural network can closely approximate it. □

A similar result with weaker separation was proven in [KW16] for the boolean activation function $h(x) := 1_{x\geq 0}$, $x \in \mathbb{R}$. In the following statement, we identify a neural network with its representation as a directed graph.

Theorem 8.7 ([KW16])**.** *For any $n \geq 1$, there exists a function $f_n\colon \{0,1\}^n \to \{0,1\}$ such that*

- *f_n can be written as a 3-layer neural network with $O(n)$ nodes and boolean activation function, and*
- *for any 2-layer neural network g with boolean activation function that agrees with f_n on $(.5+\varepsilon)$ fraction of $\{0,1\}^n$ for some $\varepsilon \geq \Omega(\sqrt{\log n/n})$, g must either have: more than $\Omega(\varepsilon^3 n^{3/2}/\log^3 n)$ nodes, or more than $\Omega(\varepsilon^3 n^{5/2}/\log^{7/2} n)$ edges.*

Rough Sketch of the Proof. For any $k \geq 1$, define bin: $\{0,1\}^k \to \{0,1\}$ by

$$\mathrm{bin}(a_1,\ldots,a_k) := \sum_{j=1}^{k} a_{j-1}2^{j-1}, \qquad \forall\,(a_1,\ldots,a_k) \in \{0,1\}^k.$$

Define also the multiplexer function

$$M_{2^k}(x_0,\ldots,x_{2^k-1},a_1,\ldots,a_k) := x_{\mathrm{bin}(a_1,\ldots,a_k)}, \quad \forall\,(x_0,\ldots,x_{2^k-1}) \in \{0,1\}^{2^k},\ (a_1,\ldots,a_k) \in \{0,1\}^k.$$

Now, fix k that is a positive power of 2, let $n := 2^k$, and define

$$f_n(x,\{a_{i,j}\}_{1\leq i\leq k, 1\leq j\leq 2^k/k}) := M_{2^k}\Big(x, \sum_{j=1}^{2^k/k} a_{1,j} \bmod 2, \ldots, \sum_{j=1}^{2^k/k} a_{k,j} \bmod 2\Big),$$

$\forall\, x \in \{0,1\}^{2^k}$, $\{a_{i,j}\}_{1\leq i\leq k, 1\leq j\leq 2^k/k} \in \{0,1\}^{2^k}$.

Note that $1_{\sum_{j=1}^n a_j \geq k} - 1_{\sum_{j=1}^n a_j \geq k+1} = 1_{\sum_{j=1}^n a_j = k}$, so the parity function

$$(a_1,\ldots,a_n) \mapsto \sum_{j=1}^{n} a_j \bmod 2$$

can be written as

$$\sum_{k\geq 0\text{ even}} 1_{\sum_{j=1}^{n} a_j = k} = \sum_{k\geq 0\text{ even}} h(\sum_{j=1}^{n} a_j - k) - h(\sum_{j=1}^{n} a_j - k + 1) \qquad (*)$$

where $h(y) = 1_{y\geq 0}$ for all $y \in \mathbb{R}$. In particular, the function $\sum_{k\geq 0\text{ even}} h(\sum_{j=1}^{n} a_j - k) - h(\sum_{j=1}^{n} a_j - k + 1)$ takes only 0 or 1 values.

In the case $k = 1$, $M_2(x_0, x_1, a)$ can be written (in logical notation) as

$$M_2(x_1, x_2, a) = (x_0 \wedge (1 - a)) \vee (x_1 \wedge a).$$

That is, M_2 can be written using 2 AND functions, one OR function and one NOT function. In the case $k = 2$, $M_4(x_0, x_1, x_2, x_3, a_1, a_2)$ can be written (in logical notation) as

$$\begin{aligned}&M_4(x_0, x_1, x_2, x_3, a_1, a_2)\\ &\quad = (x_0 \wedge (1 - a_1) \wedge (1 - a_2)) \vee (x_1 \wedge a_1 \wedge (1 - a_2)) \vee (x_2 \wedge (1 - a_1) \wedge a_2) \vee (x_3 \wedge a_1 \wedge a_2).\end{aligned}$$

That is, M_4 can be written using 8 AND functions, three OR function and four NOT functions.

More generally, M_{2^k} can be written using $k2^k$ AND functions, at most k OR functions and at most $k2^k$ NOT functions. That is, M_{2^k} can be computed using a depth 2 circuit of at most $3k2^k$ AND, OR and NOT functions.

Note that the AND of m variables can be written itself as a one-layer neural network. For example:

$$x_1 \wedge \cdots \wedge x_m = 1_{x_1+\cdots+x_m > m-1}, \qquad \forall\, x_1, \ldots, x_m \in \{0, 1\}.$$

Similarly, the OR of m variables can be written as a one-layer neural network. For example:

$$x_1 \vee \cdots \vee x_m = 1_{x_1+\cdots+x_m > 0}, \qquad \forall\, x_1, \ldots, x_m \in \{0, 1\}.$$

So, using also $(*)$, the composition of M_{2^k} and k parity functions can be written as a 3-layer neural network with $O(k2^k) = O(n \log n)$ gates. The first item is proven.

We now prove the second item. Suppose we take x, we randomly sort the indices of x into $k = \log n$ subsets of equal size, and in each of these subsets, we fix all but one entry in that subset uniformly at random. A Lemma shows that the probability that an LTF is not a constant function after this procedure is $O(k/\sqrt{n}) = O(\log n/\sqrt{n})$. (For example, consider a majority function on n variables; randomly fixing many entries will typically cause $x_1 + \cdots + x_n$ to be $\Omega(\sqrt{n})$ plus the unconstrained variables.) So, if a two layer neural network has $O(\varepsilon^3 n^{3/2}/\log^3 n)$ nodes, then the expected number of bottom-layer nodes that are not fixed to be constant by this restriction procedure is $O(\varepsilon^3 n/\log^2 n)$. So, by Markov's inequality, with probability at least $1-\varepsilon/3$, when g is restricted randomly in this way, all but $O(\varepsilon^2 n/\log^2 n)$ of its bottom level gates are restricted to be constant. That is, the restricted g is equivalent to a two-layer neural network with $O(\varepsilon^2 n/\log^2 n)$ gates.

On the other hand, if this random restriction procedure is applied to $\{a_{i,j}\}_{1\leq i\leq k, 1\leq j\leq 2^k/k}$ so e.g. all but one of $a_{i,1}, \ldots, a_{i,2^k/k}$ is fixed (for every $1 \leq i \leq j$), then each parity function such as $\sum_{j=1}^{2^k} a_{1,j} \bmod 2$ is not constant, but it is a function of the single un-fixed variable.

We can then think of f_n, for fixed x, as a function of $a_1, \ldots, a_k$. Moreover, f_n is interpreted as a random function of $a_1, \ldots, a_k$ (where the randomness comes from a random choice of x). In this way, f_n is interpreted as a random function. But a random function on $k = \log_2 n$

boolean inputs cannot agree with a two-layer neural network with $O(\varepsilon^2 n/\log^2 n)$ gates on $(1/2)+\varepsilon/3$ fraction of its inputs (this is another Lemma which we omit).

□

9. Appendix: Basics of Complexity Theory

A Turing machine is a standard model of computation introduced by Turing in 1937. Informally, a one-tape Turing machine is a computing device with a finite state control device, a tape (with cells indexed by the positive integers), and a tape head that points to (or scans) a given cell on the tape. At a given time (indexed by the positive integers), the machine changes its state, writes a symbol on the cell to which is currently pointing, and then moves the tape head one cell to the left or one cell to the right. Time is then increased by one. The action of the machine at a given time is a function of its current state and of the symbol that is currently scanned by the tape head. More formally:

Definition 9.1 (**Turing Machine**)**.** A **one-tape Turing machine** is defined by

- $\Sigma :=$ a finite set ("alphabet"), together with a blank symbol $\{\Box\}$.
- $Q :=$ a finite set of "control states" $\cup\ \{q_{\text{accept}}, q_{\text{reject}}, q_{\text{start}}\}$.
- $\delta\colon \Sigma\times Q\to\Sigma\times Q\times\{\leftarrow,\rightarrow\}$, the "transition function."

A **configuration** of a Turing machine consists of

- the symbols on the tape up to and including its rightmost non-blank symbol,
- the current state of the control device (an element of Q), and
- the position of the tape head (an element of the positive integers).

So, if $i+1, n$ are positive integers, if $\sigma_1,\ldots,\sigma_n\in\Sigma$, and if $q\in Q$, the string

$$\sigma_1\sigma_2\cdots\sigma_i q\sigma_{i+1}\sigma_{i+2}\cdots\sigma_n$$

represents the configuration of the Turing machine where the finite state control is in state q, the tape head is currently scanning position $i+1$ on the tape, and the contents of the tape are $\sigma_1\cdots\sigma_n$.

Let $x_1,\ldots,x_n\in\Sigma\setminus\{\Box\}$ and let $x=x_1x_2\cdots x_n$ be the corresponding string. A Turing machine then **computes** on the input x by doing the following.

- The Turing machine is initialized to the initial configuration $C_0 := q_{\text{start}}x_1x_2\cdots x_n$.
- At any time step $i\geq 1$, the Turing machine applies the transition function δ to the previous configuration C_{i-1} to obtain the next configuration C_i. So, if $C_{i-1}=\sigma_1\sigma_2\cdots\sigma_i q\sigma_{i+1}\sigma_{i+2}\cdots\sigma_n$, we have $\delta(\sigma_{i+1},q)=:(\sigma',q',a)$ for some $\sigma'\in\Sigma$, $q'\in Q$ and $a\in\{\leftarrow,\rightarrow\}$. We then define
$$C_i := \begin{cases}\sigma_1\sigma_2\cdots\sigma_{i-1}q'\sigma_i\sigma'\sigma_{i+2}\sigma_{i+3}\cdots\sigma_n & ,\text{ if } a=\leftarrow\\ \sigma_1\sigma_2\cdots\sigma_{i-1}\sigma_i\sigma' q'\sigma_{i+2}\sigma_{i+3}\cdots\sigma_n & ,\text{ if } a=\rightarrow.\end{cases}$$
- If at any time step $k\geq 1$ the Turing machine enters a halting state (q_{accept} or q_{reject}), we say the machine **halts** and **accepts** (or **rejects**) input x in k steps.

If the machine halts in k steps, the computation of the machine on input x can be described as a sequence of configurations $C_0, C_1,\ldots,C_k$.

Remark 9.2. Modern computers do not directly implement Turing machines. Modern computers more closely resemble RAM machines, or von Neumann machines.

Remark 9.3. A multi-tape Turing Machine is defined similarly to the above, except it has one single input tape (where only the input state is written), one single output tape, a constant number of other ("work") tapes, the function δ is then a function of all of the tapes and the current state of the finite control. When the configuration changes, each work tape head overwrites the cell it is currently scanning, and each work tape head moves left or right one cell.

Let Σ be a finite set (or "alphabet") and let Σ^* denote the set of finite strings of elements of Σ.

Definition 9.4 (Decided). A set (or "language") $S \subseteq \Sigma^*$ is **recognized** (or **decided**) by a Turing Machine M if:

- If $x \in S$, then $M(x)$ accepts.
- If $x \notin S$, then $M(x)$ rejects.

Definition 9.5 (Accepted). A set (or "language") $S \subseteq \Sigma^*$ is **accepted** by a Turing Machine M if:

- If $x \in S$, then $M(x)$ accepts.
- If $x \notin S$, then $M(x)$ does not accept.

Let $\mathbb{N} := \{0, 1, 2, \ldots\}$. For any $x \in \Sigma^*$, let $|x|$ denote the length of the string x.

Definition 9.6 (Time Complexity). Let $f\colon \mathbb{N} \to \mathbb{N}$. We say a set $S \subseteq \Sigma^*$ satisfies $S \in \mathrm{TIME}(f(n))$ if there exists a multitape Turing Machine M such that

- M recognizes S, and
- For all $x \in S$, $M(x)$ halts within $f(|x|)$ steps.

Definition 9.7 (Space Complexity). Let $f\colon \mathbb{N} \to \mathbb{N}$. We say a set $S \subseteq \Sigma^*$ satisfies $S \in \mathrm{SPACE}(f(n))$ if there exists a multitape Turing Machine M such that

- M recognizes S, and
- For all $x \in S$, $M(x)$ uses no more than $f(|x|)$ squares of its *work tapes*. That is, $M(x)$ scans at most $f(|x|)$ cells of its work tapes.

Definition 9.8 (Time Complexity for Functions). Let $f\colon \mathbb{N} \to \mathbb{N}$. We say a function $g\colon \Sigma^* \to \Sigma^*$ satisfies $g \in \mathrm{FTIME}(f(n))$ if there exists a multitape Turing Machine M such that

- For all $x \in \Sigma^*$, $M(x)$ writes $g(x)$ on its output tape, and
- $M(x)$ halts within $f(|x|)$ steps.

Definition 9.9 (Space Complexity for Functions). Let $f\colon \mathbb{N} \to \mathbb{N}$. We say a function $g\colon \Sigma^* \to \Sigma^*$ satisfies $g \in \mathrm{FSPACE}(f(n))$ if there exists a multitape Turing Machine M such that

- For all $x \in \Sigma^*$, $M(x)$ writes $g(x)$ on its output tape, and
- $M(x)$ uses no more than $f(|x|)$ squares of its *work tapes*.

Definition 9.10 (Complexity Class P). We define

$$\mathbf{P} := \cup_{k \geq 1} \mathrm{TIME}(n^k).$$

Definition 9.11 (**Complexity Class NP**). **NP** is the class of sets $S \subseteq \Sigma^*$ such that there exists a Turing Machine V (a "verifier") and a polynomial $p\colon \mathbb{N} \to \mathbb{N}$ such that $x \in S$ if and only if: $\exists$ a string $y \in \Sigma^*$ of length at most $p(|x|)$ such that $V(x, y)$ accepts in time at most $p(|x|)$.

Definition 9.12 (**NP-hard and NP-complete**). A set $S \subseteq \Sigma^*$ is **NP**-hard if for all $T \in$ **NP**, there exists a logspace function $g\colon \Sigma^* \to \Sigma^*$ such that $x \in S$ if and only if $g(x) \in T$. And S is **NP**-complete if $S \in$ **NP** and S is **NP**-hard.

The SAT problem is the following. Let n

Definition 9.13 (**Satisfiability Problem (SAT)**). Let $f\colon \{-1,1\}^n \to \{-1,1\}$ be an unknown function with query access . Decide whether or not there exists $x_1, \ldots, x_n \in \{-1,1\}$ such that $f(x_1, \ldots, x_n) = 1$.

The Satisfiability problem is sometimes stated as follows.

Definition 9.14 (**CNF Satisfiability Problem (CNF SAT)**). Let $y_1, \ldots, y_n$ be variables. Suppose we are given a CNF formula

$$(z_1 \vee z_2 \vee \cdots \vee z_{m_1}) \wedge (z_{m_1+1} \vee \cdots \vee z_{m_2}) \wedge \cdots .$$

where $z_1, z_2, \ldots$ are either variables $y_1, \ldots, y_n$ or their negations, $\vee$ represents a logical "or" operation, and $\wedge$ represents a logical "and" operation. (That is, if $a, b \in \{0,1\}$, we have $a \vee b := \min(a, b)$ and $a \wedge b := \max(a, b)$.)

Decide whether or not there exists $x_1, \ldots, x_n \in \{0,1\}$ such that the CNF formula evaluates to 1 when $y_i = x_i$ for all $1 \leq i \leq n$.

Theorem 9.15 (**Cook-Levin, 1971**). SAT *is* **NP**-*complete.* CNFSAT *is* **NP**-*complete.*

NP can be equivalently defined using non-deterministic Turing Machines. A **non-deterministic Turing machine** is defined by repeating the definition of a Turing machine, except the function δ is allowed to take multiple values. Then the computation of a non-deterministic Turing machine can be viewed as a tree of configurations, rather than a sequence of configurations, since one configuration can transition to multiple different configurations. More specifically, we associate a directed edge from configuration C to configuration C' if δ maps configuration C to configuration C'. A non-deterministic Turing machine accepts $x \in \Sigma^*$ in time $t > 0, t \in \mathbb{Z}$ if there exists a path from the initial configuration to an accepting configuration of length at most t. A non-deterministic Turing machine M **accepts** a set $S \subseteq \Sigma^*$ if: $x \in S$ if and only if M accepts x.

Intuitively, a non-deterministic Turing machine is a Turing machine that is allowed to make arbitrary choices at each step of its computation.

A nondeterministic Turing machine can solve SAT is linear time using a binary tree of configurations (Exercise).

Definition 9.16 (**Non-deterministic Time Complexity**). Let $f\colon \mathbb{N} \to \mathbb{N}$. We say a set $S \subseteq \Sigma^*$ satisfies $S \in \mathrm{NTIME}(f(n))$ if there exists a nondeterministic Turing Machine M such that

- M recognizes S, and
- For all $x \in S$, when M has input x, the computation of M has no path longer than $f(|x|)$.

An equivalent definition of **NP** is then

$$\mathbf{NP} := \cup_{k\geq 1}\mathrm{NTIME}(n^k).$$

Definition 9.17 (Complexity Class #P). A function $f\colon \{0,1\}^* \to \mathbb{N}$ is in **#P** if $\exists$ a polynomial $p\colon \mathbb{N}\to\mathbb{N}$ and a polynomial time Turing Machine M such that, for all $x\in\{0,1\}^*$,

$$f(x) = \left|\left\{y\in\{0,1\}^{p(|x|)}\colon M(x,y)=1\right\}\right|.$$

Equivalently, f is in **#P** if there is a polynomial time non-deterministic Turing machine M such that, for all $x\in\{0,1\}^*$, $f(x)$ is equal to the number of accepting paths of M on input x.

Informally, **#P** problems count the number of solutions to problems in **NP**. Here is an example of a problem (i.e. a function) in this class.

Definition 9.18 (#SAT). Let $g\colon\{-1,1\}^n\to\{-1,1\}$ be an unknown function with query access. Find the number of $(x_1,\ldots,x_n)\in\{-1,1\}^n$ such that $g(x_1,\ldots,x_n)=1$.

Definition 9.19 (Complexity Class RP). RP is the class of sets $S\subseteq\Sigma^*$ such that there exists a non-deterministic polynomial time Turing Machine M such that

- M accepts S, and
- If $x\in S$, then at least $1/2$ of all computation paths of $M(x)$ accept.

Informally, **RP** is a Turing machine that can use randomness at each of its computation steps.

10. Appendix: Some Functional Analysis

Below, we consider either $\mathbb{R}$ or $\mathbb{C}$ as scalars. That is, we will be dealing with vector spaces over the fields $\mathbb{R}$ or $\mathbb{C}$, and we will only distinguish them where necessary. A **topological vector space** is a Hausdorff topological space X that is also a vector space, such that: the map $(x,y)\mapsto x-y$ from $X\times X\to X$ is continuous , and the map $(\alpha,x)\mapsto\alpha x$ from $\{\text{scalars}\}\times X\to X$ is continuous.

A **normed linear space** X is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) with a **norm** $\|\cdot\|$. A norm is a function $\|\cdot\|: X\to[0,\infty)$ such that $\|x\|\geq 0$ with equality if and only if $x=0$, $\|\alpha x\| = |\alpha|\,\|x\|$ for α a scalar, and $\|x+y\|\leq\|x\|+\|y\|$. Using the norm, we see that $d(x,y):=\|x-y\|$ is a metric, whose open balls define the metric topology on X. We refer to this topology as the norm topology, or strong topology. Using the triangle inequality, one can show that a normed linear space is also a topological vector space. A **Banach space** is a normed linear space that is complete with respect to the norm topology. A vector subspace $Y\subseteq X$ is simply called a **subspace**, and will be closed unless otherwise stated. Also, unless otherwise stated, $\overline{A}$ denotes the closure of A with respect to the norm topology.

An **inner product** is a function $(\cdot,\cdot)\colon X\times X\to\{\text{scalars}\}$ that is linear in the first argument, conjugate linear in the second argument, Hermitian symmetric, and positive definite. An inner product space is a vector space with an inner product. Defining $\|x\|:=(x,x)^{1/2}$ shows that an inner product space is a normed linear space (using the Cauchy-Schwarz inequality). A **Hilbert space** is an inner product space that is complete in the norm topology. Cauchy-Schwartz shows that the inner product is continuous with respect to the norm topology, since

$$|(u,v)-(u_0,v_0)| \le |(u-u_0,v)| + |(u_0,v-v_0)| \le \|u-u_0\| \, \|v\| + \|v-v_0\| \, \|u_0\|$$

One can show that $\|x-y\|^2 = \|x\|^2 - 2\Re(x,y) + \|y\|^2$. In Hilbert space we have the **parallelogram law** $\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$, and the **polarization identity** $(u,v) = \frac{1}{4}\sum_k i^k \left\|u + i^k v\right\|^2$, where $k = 0,2$ for X over $\mathbb{R}$, and $k = 0,1,2,3$ for X over $\mathbb{C}$. Actually, polarization holds in a Banach space if and only if it is a Hilbert space.

A continuous linear function $L\colon (X, \|\cdot\|_X) \to (Y, \|\cdot\|_Y)$ is called a **linear operator**. One can show that L is uniformly continuous if and only if it is continuous if and only if it continuous at zero if and only if it is **bounded**, i.e. $\|L(x)\|_Y \le M \|x\|_X$. Showing that continuity at zero implies boundedness involves a scaling argument. Showing that boundedness implies uniform continuity follows by linearity.

The least such M such that $\|L(x)\|_Y \le M\|x\|_X$ is called the **operator norm** of L, and it is denoted by $\|L\|$. Note that the set of bounded linear maps $B(X,Y)$ from X to Y is a normed linear space with respect to the operator norm topology. If Y is complete, then $B(X,Y)$ is a Banach space. (Fix $x \in X$, observe $L_n(x) \to L(x)$, L achieves linearity using subsequential arguments, etc.) Essentially by definition, we have $\|L\| := \sup_{\|x\|\le 1} \|L(x)\|$. If X, Y are Hilbert spaces, $\|L\| = \sup_{\|x\|\le 1, \|y\|\le 1} |\langle L(x), y\rangle|$. For $L\colon H \to H$ a linear operator on a Hilbert space H, we can apply Thm. 10.2 to define the **adjoint** L^*. Here $L^*(u) = v$ is the unique v such that $(L(w), u) = (w, L^*(u))$. Moreover, $\|L\| = \|L^*\|$ (using the formula for $\|L\|$ from a few previous lines). More directly, we can observe that $\|A^*y\|^2 = (AA^*y, y) \le \|A\| \, \|A^*y\| \, \|y\|$, so $\|A^*\| \le \|A\|$. Then, using that $A^{**} = A$ (take complex conjugates of the definition of A^*) we see that $\|A\| \le \|A^*\|$.

A **linear functional** on X is a bounded map from X to scalars ($\mathbb{C}$ or $\mathbb{R}$). The space of linear functionals is called the **dual space** of X, and is denoted by X^*. The norm of $x^* \in X^*$ is given by $\|x^*\| := \sup_{\|x\|\le 1} |x^*(x)|$.

A **semi-norm** N on a vector space X is a function $N\colon X \to \mathbb{R}$ such that: $\forall\ x \in X$, $N(x) \ge 0$, $N(\alpha x) = |\alpha| N(x)$, and $N(x+y) \le N(x) + N(y)$. A collection $\mathcal{N}$ of seminorms is called **separating** if $N(x) = 0\ \forall\ N \in \mathcal{N}$ if and only if $x = 0$. For a seminorm N, define $S_N(a,\rho) := \{x \in X\colon N(x-a) < \rho\}$. $S_N(a,\rho)$ is called the **open** N **ball** of radius ρ centered at a. A set of the form $S = S_{N_1}(a,\rho_1) \cap \cdots \cap S_{N_j}(a,\rho_j)$ is called an **open** $\mathcal{N}$ **ball** centered at a. Given X and $\mathcal{N}$, let $X^{\mathcal{N}}$ denote X with the topology given by open $\mathcal{N}$ balls. One can check that $X^{\mathcal{N}}$ is Hausdorff if and only if $\mathcal{N}$ is separating, and $X^{\mathcal{N}}$ is a topological vector space when $\mathcal{N}$ is separating. Sometimes, $X^{\mathcal{N}}$ is called a **Fréchet space**. For example, consider the **Schwartz class**, $\mathcal{S}(\mathbb{R}) = \{f \in C^\infty\colon x^n f^{(k)} \in L_\infty, \forall n,k \ge 0\}$, equipped with the seminorms $\|f\|_{n,k} = \sup_{x\in\mathbb{R}} \left|x^n f^{(k)}(x)\right|$. Here $\mathcal{N}$ is separating since $\|f\|_{0,0} = 0$ implies $f = 0$.

A topological vector space is called **locally convex** if X has a neighborhood base at zero consisting of convex sets. That is, for any neighborhood U of 0, $\exists\ V$ convex such that $0 \in V \subseteq U$. It takes some work to show: if X is a topological vector space, and $\mathcal{N}$ is a family of seminorms, then $X^{\mathcal{N}} = X$ if and only if X is locally convex. As an example of the above we define the **weak topology** of X, denoted by $\sigma(X, X^*)$, as the topology on X induced by the seminorms $x \mapsto \|L(x)\|$, $L \in X^*$. In particular, we can form a basis of neighborhoods of x_0 using the seminorms:

$$U(x_0; \rho, x_1^*, \ldots, x_n^*) := \{x \in X\colon \left|x_j^*(x) - x_j^*(x_0)\right| < \varepsilon \text{ for } j = 1, \ldots, n\}$$

Note that $x_n \to x$ weakly (i.e. with respect to $\sigma(X, X^*)$) if and only if $\forall\, x^* \in X^*$, $x^*(x_n) \to x^*(x)$. A sequence $\{x_n\}_{n\geq 1}$ such that $x^*(x_n)$ converges for all $x^* \in X^*$ is called **weakly Cauchy**.

Analogously, we can define the **weak* topology** on X^*, denoted by $\sigma(X^*, X)$. Here a basis of neighborhoods of $x_0^* \in X^*$ is given by

$$U(x_0^*, \varepsilon, x_1, \ldots, x_n) := \{x^* \in X^* \colon\ |x^*(x_j) - x_0^*(x_j)| < \varepsilon \text{ for } j = 1, \ldots, n\}$$

For X a normed linear space, define the **canonical map** $\iota \colon X \to X^{**}$ by $\iota(x)(x^*) = x^*(x)$. One can check that ι is linear, and Thm. 10.5 shows that ι is an isometry:

$$\|\iota(x)\| = \sup_{\|x^*\|\leq 1} |\iota(x)(x^*)| = \sup_{\|x^*\|\leq 1} |x^*(x)| = \|x\|$$

In particular, $\iota \colon (X, \sigma(X, X^*)) \to (\iota(X), \sigma(X^{**}, X^*))$ is a homeomorphism, from X with the weak topology, to $\iota(X) \subseteq X^{**}$ with the weak* topology (of X^{**} acting on X^*). A Banach space is said to be **reflexive** if $\iota \colon X \to X^{**}$ is onto.

Let $K \subseteq X$ with X a topological vector space. A point $x \in K$ is called an **extreme point** if it is not contained in the interior of any **segment** $\{ty + (1-t)z \colon t \in [0,1], y, z \in K\} \subseteq K$. By interior we mean $\{ty + (1-t)z \colon t \in (0,1), y, z \in K\}$. If K is compact and convex, then a **face** $F \subseteq K$ is a subset such that: if $y \in K$ is in the interior of a segment in K, then the whole segment is in F. Let K be a nonempty compact, convex set. Denote the set of extreme points of K by $\mathcal{E}(K)$. Define $\widehat{K}$ as the intersection of all compact, convex sets containing $\mathcal{E}(K)$. By definition, $\widehat{K} \subseteq K$ and $\widehat{K} \neq \emptyset$. An **extreme face** F of K is a face of K where the only faces of F are $\emptyset$ and F.

Let μ be a probability measure. A subset $H \subseteq L_1(\mu)$ is called **uniformly integrable** if, for all $\varepsilon > 0$, $\exists\, \eta > 0$ with $\sup\left\{\int_A |f|\, d\mu \colon \mu(A) \leq \eta, f \in H\right\} \leq \varepsilon$. (Uniform integrability has a slightly different definition in the probability literature.)

An **associative algebra** $\mathcal{A}$ over a field F is a vector space over F with a bilinear, associative multiplication: $(ab)c = a(bc)$, $a(b+c) = ab + ac$, $(a+b)c = ac + bc$, and $a(\lambda c) = (\lambda a)c = \lambda(ac)$. A **Banach algebra** is a real or complex Banach space that is an associative algebra such that $\|ab\| \leq \|a\|\,\|b\|$. Examples include: $(C(X), \mathbb{C})$ for X a topological space, $B(V, V)$, and $L_1(\mathbb{R}^n)$ with convolution. Note that the latter has no identity. However, given a (complex) Banach algebra without identity, if we let $B := \{(a, \alpha) \colon a \in \mathcal{A}, \alpha \in \mathbb{C}\} = A \oplus \mathbb{C}$, and we define $(a, \alpha)(b, \beta) := (ab + \alpha b + \beta a, \alpha\beta)$ and $\|(a, \alpha)\| := \|a\| + |\alpha|$, then B has identity $e = (0, 1)$.

Let $\mathcal{A}$ be a Banach algebra with identity. Using power series (and that $\|a^n\| \leq \|a\|^n$), we see: if $\|a\| < 1$, then $1 - a$ is invertible (let $(1-a)^{-1} := \sum_{n\geq 0} a^n$, so $\|(1-a)^{-1}\| \leq \frac{1}{1-\|a\|}$). Using similar arguments, we see that: the **invertible elements** $\mathcal{U}$ form an open set, and the inverse map is continuous from $\mathcal{U}$ to itself. For $a \in \mathcal{U}$ and $\|x - a\| < \|a^{-1}\|^{-1}$, note that $\|a^{-1}x - 1\| = \|a^{-1}(x - a)\| \leq \|a^{-1}\|\,\|x - a\| < 1$, so $1 - (1 - a^{-1}x) = a^{-1}x$ has inverse b, so $1 = (ba^{-1})x$ and $a^{-1}xb = 1$, so $xb = a$, $x(ba^{-1}) = 1$, i.e. $(ba^{-1}) = x^{-1}$.

From now on, we assume that all Banach algebras $\mathcal{A}$ are complex and have an identity. Let $x \in \mathcal{A}$, and for convenience, λ denotes $\lambda \cdot 1 \in \mathcal{A}$. The **spectrum** of x is $\sigma(x) := \{\lambda \in \mathbb{C} \colon x - \lambda \text{ is not invertible}\}$. The **resolvent set** of x is $\rho(x) := \{\lambda \in \mathbb{C} \colon x - \lambda \text{ is invertible}\}$. The **resolvent** of x is the function $R(\lambda) = (x - \lambda)^{-1}$, defined for $\lambda \in \rho(x)$. The **spectral radius** is $r(x) := \sup\{|\lambda| : \lambda \in \sigma(x)\}$.

From our analysis of the invertible elements, we see that $r(a) \le \|a\|$. Thus, $\sigma(a)$ is bounded. In fact, it is compact, since $\rho(a)$ is open ($\lambda \in \mathbb{C} \mapsto a - \lambda \in \mathcal{A}$ is continuous, and $\rho(a) = \{\lambda \colon a - \lambda \in \mathcal{U}\}$), so $\sigma(a) = \rho(a)^c$ is closed. Also, $\sigma(a) \neq \emptyset$. To see this, define a **weakly analytic** function as a function ϕ from an open set $V \subseteq \mathbb{C}$ to a complex Banach space $\mathcal{A}$ such that $\xi \circ \phi$ is analytic for every $\xi \in \mathcal{A}^*$. Using power series, one can show: $R(\lambda) = (a-\lambda)^{-1}$ is weakly analytic on $\rho(a)$ and $\|R(\lambda)\| \to 0$ as $\lambda \to \infty$. So, if $\sigma(a)$ is empty, Liouville's Theorem (for bounded analytic functions) says that $\xi((a-\lambda)^{-1}) \equiv 0$, so $(a-\lambda)^{-1} \equiv 0 \ \forall \lambda$, which is a contradiction.

For the power series argument, let $\lambda_0 \in \rho(a)$. Write $a - \lambda = (1-\lambda_0)(a-(a-\lambda_0)^{-1}(\lambda-\lambda_0))$. Then $a-\lambda$ is invertible if $\|(a-\lambda_0)^{-1}(\lambda-\lambda_0)\| < 1$. Choose λ so the latter condition holds. Then $(a-\lambda)^{-1} = \sum_{n\ge 0}(a-\lambda_0)^{-n}(\lambda-\lambda_0)^n(a-\lambda_0)^{-1}$. So $\xi((a-\lambda)^{-1}) = \sum_{n\ge 0}\xi((a-\lambda_0)^{-n-1})(\lambda-\lambda_0)^n$, i.e. $\xi(R(\lambda))$ is analytic (in particular, for $\lambda \in \rho(a)$.) Now, $(a-\lambda)^{-1} = [\lambda(\lambda^{-1}a-1)]^{-1} = \lambda^{-1}(\lambda^{-1}a-1)^{-1}$ and $\|(\lambda^{-1}a-1)^{-1}\| \le (1-|\lambda|^{-1}\|a\|)^{-1} \to 1$ as $\lambda \to \infty$, so $\|R(\lambda)\| \to 0$ as $\lambda \to \infty$.

Below, $\mathcal{B}$ denotes a commutative Banach algebra with identity. A **character** α of $\mathcal{B}$ is a nonzero multiplicative linear function on $\mathcal{B}$, i.e. $\alpha(ab) = \alpha(a)\alpha(b)$. (A character is not assumed to be bounded a priori, but this can be proven. See Thm. 10.23(9).) The **spectrum** of $\mathcal{B}$ is the set $\widetilde{\mathcal{B}}$ of all characters of $\mathcal{B}$. For $a \in \mathcal{B}$, $\alpha \in \widetilde{\mathcal{B}}$ define $\hat{a}(\alpha) := \alpha(a)$. The map $a \mapsto \hat{a}$ from $\mathcal{B}$ into $C(\widetilde{\mathcal{B}})$ is called the **Gelfand map** or canonical map.

An **ideal** $I \subseteq \mathcal{B}$ is closed under addition (within I) and multiplication by elements of $\mathcal{B}$. An ideal is called **maximal** if $I \neq \mathcal{B}$ and I is not contained in any larger proper ideal. Define the **singular elements** $\mathcal{S}$ as the union of all maximal ideals in $\mathcal{B}$. Define the **radical** of $\mathcal{B}$ as the intersection of all maximal ideals in $\mathcal{B}$. Since $I \subseteq \overline{I} \subseteq \mathcal{S}$ is a proper ideal, we see that maximal ideals are closed, so the radical is closed. $\mathcal{B}$ is called **semisimple** if its radical is $\{0\}$.

An **involution** on a Banach algebra $\mathcal{B}$ is a map $\mathcal{B} \to \mathcal{B}$ written $a \mapsto a^*$ such that: $a^{**} = a$, $(a+b)^* = a^* + b^*$, $(\lambda a)^* = \overline{\lambda}a^*$, and $(ab)^* = b^*a^*$. Note that $1^* = 1$, since applying $*$ to $1 \cdot 1^* = 1^*$ gives $1^{**} \cdot 1^* = 1^{**}$, so since $1^{**} = 1$, we get $1 \cdot 1^* = 1$, so $1^* = 1$. We say a is **Hermitian** if $a = a^*$, **strongly positive** if $a = b^*b$ for some b, **positive** if $\sigma(a) \subseteq [0,\infty)$ and **real** if $\sigma(a) \subseteq \mathbb{R}$. An involution is **symmetric** if $1 + a^*a$ is invertible for all $a \in \mathcal{B}$.

A Banach $*$ algebra $\mathcal{B}$ is called $*$ **multiplicative** if $\|a^*a\| = \|a^*\|\,\|a\|$, $*$ **isometric** if $\|a^*\| = \|a\|$, and $*$ **quadratic** if $\|a^*a\| = \|a\|^2$. One can show that the first two together are equivalent to the third. The forward direction is clear. For the reverse, observe that $\|a\|^2 = \|a^*a\| \le \|a^*\|\,\|a\|$, so $\|a\| \le \|a^*\|$ (and also for $a = b^*$), so $\|a\| = \|a^*\|$. Also, $\|a^*a\| = \|a\|^2 = \|a^*\|\,\|a\|$. A B^* **algebra** is a quadratic $*$ algebra. (In modern terminology a C^* algebra is a multiplicative $*$ algebra. From the equivalences just shown, a B^* algebra is also a C^* algebra, and no one uses the term B^* algebra anymore).

Let $A, B \in B(H)$, i.e. let A, B be bounded linear operators on a Hilbert space H. Let $*$ denote the adjoint operation. Summarizing some properties of $B(H)$, we have: A^* is linear and bounded with $\|A^*\| = \|A\|$, $A^{**} = A$, $(\alpha A + \beta B)^* = \overline{\alpha}A^* + \overline{\beta}B^*$, $\|AB\| \le \|A\|\,\|B\|$, and $(AB)^* = B^*A^*$. Also note that $\|A^*A\| = \|A\|^2$. This follows since $\|A^*A\| \le \|A^*\|\,\|A\| = \|A\|^2$, and $\|Ax\|^2 = (A^*Ax, x) \le \|A^*A\|\,\|x\|^2$. From above, we saw that $B(X)$ is a Banach algebra whenever X is a Banach space. Thus, $B(H)$ is a B^* algebra.

In these notes, a C^***-algebra** on a Hilbert space H is a subalgebra $\mathcal{A}$ of $B(H)$ which is closed in norm and such that $A \in \mathcal{A}$ implies $A^* \in \mathcal{A}$. A subalgebra closed under taking adjoints is called a $*$ **subalgebra** of $\mathcal{B}(H)$. Note that $B(H)$ is a C^*-algebra. A **maximal abelian self-adjoint (m.a.s.a.) algebra** on H is a commutative algebra $\mathcal{A} \subseteq B(H)$ which is not contained in any larger commutative subalgebra, and such that $\mathcal{A}$ is a $*$-subalgebra. Let $S \subseteq B(H)$. Define $S' = \{A \in B(H)\colon AB = BA, \forall B \in S\}$. S' is then a subalgebra of $B(H)$ for any set S, and we call S' the **commutor algebra** of S.

Let (X, μ) be a measure space. Let $f \in L_\infty(\mu)$. Define $M_f \colon L_2(\mu) \to L_2(\mu)$ by $M_f(g) := fg$. Since $fg \in L_2$ for $g \in L_2$, M_f is everywhere defined, and $\|M_f g\|_2^2 \leq \|f\|_\infty^2 \|g\|_2^2$. Therefore, $\|M_f\| \leq \|f\|_\infty$. Note also that $M_{fg} = M_f M_g, M_{\alpha f + \beta g} = \alpha M_f + \beta M_g, M_f^* = M_{\overline{f}}$.

Unless otherwise specified, in this Section for a measure space (X, μ), we assume: every measurable set in X of positive measure contains a subset of finite strictly positive measure. (That is, μ has no infinite atoms). Under this assumption, $\|M_f\| = \|f\|_\infty$, which can be proven by considering the indicator function where $|f| = a$ for $\|f\|_\infty > a > 0$ (if such an a exists). The **multiplication algebra**, denoted by $\mathcal{M}(X, \mu)$, of (X, μ) is the algebra of operators on $L_2(X, \mu)$ consisting of all $M_f, f \in L_\infty$.

Let $D(w, \varepsilon) = \{z \in \mathbb{C}\colon |z - w| < \varepsilon\}$. If $f \in L_\infty(X, \mu)$, define the **essential range** of f as $\{w \in \mathbb{C}\colon \mu(f^{-1}(D(w, \varepsilon))) > 0 \text{ for all } \varepsilon > 0\}$. For (X, μ) with no infinite atoms, an exercise shows: $\sigma(M_f) =$ essential range of f. Let $\mathcal{A}$ be a subalgebra of $B(H)$. A vector $x \in H$ is called a **cyclic vector** for $\mathcal{A}$ if $\mathcal{A}x := \{Ax\colon A \in \mathcal{A}\}$ is dense in H. A **unitary operator** $U\colon H \to K$ between two Hilbert spaces is a linear surjective operator such that $\|Ux\| = \|x\|$ $\forall x \in H$. To emphasize the surjectivity, we write $U\colon H \twoheadrightarrow K$.

A bounded operator $A\colon H \to H$ is called: **normal** if $A^*A = AA^*$, **Hermitian** if $A = A^*$, **unitary** if A is onto and $\|Ax\| = \|x\|$ $\forall x \in H$, and **orthogonal** if H is real and A is unitary. Suppose H is a Hilbert space, $A\colon H \to H$ is linear, and $(Ax, x) = 0$ for all $x \in H$. Then (a) if H is complex, then $A = 0$ (b) if H is real and $A^* = A$ then $A = 0$. To prove this, we observe:

$$(A(x+y), x+y) - (A(x-y), (x-y)) = 2(Ax, y) + 2(Ay, x)$$

So $(Ax, y) + (Ay, x) = 0$. If H is real and $A = A^*$, this yields $(Ax, y) = 0$ for all x, y, so $A = 0$. For H complex, we can get the same conclusion by writing $(Ax, y) + (Ay, x) = 0$ and substituting $x \mapsto ix$ to get $i(Ax, y) - i(Ay, x) = 0$, etc.

Using the above we see: a linear function is **unitary** if and only if U is bounded and $UU^* = U^*U = id$. If the latter holds, then $\|Ux\|^2 = (U^*Ux, x) = \|x\|^2$, and $U(U^*x) = x$, so U is onto. In the reverse direction, if U is unitary, then $((U^*U - id)x, x) = \|Ux\|^2 - \|x\|^2 = 0$, so from the above we see that $U^*U - id = 0$. Using surjectivity of U (given x, let y such that $Uy = x$) we have $UU^*x = UU^*Uy = Uy = x$, so $UU^* = id$.

Let (X, μ) be a σ-finite measure space, and let $f \in L_\infty$. Using that $M_f^* = M_{\overline{f}}$, one can see that: (1) M_f is normal, (2) M_f is Hermitian if and only if f is real a.e. (i.e. $f = \overline{f}$ a.e.), and (3) M_f is unitary if and only if $|f| = 1$ a.e. (i.e. $f\overline{f} = 1$ a.e.). A sequence A_n of operators on a Banach space B **converges strongly** to a bounded operator A if $A_n x \to Ax$ for each $x \in B$. A_n **converges weakly** if $\langle A_n x, y\rangle \to \langle Ax, y\rangle$ $\forall x \in B, y \in B^*$. If B is a Hilbert space, weak convergence is therefore: $(A_n x, y) \to (Ax, y)$ $\forall x, y \in H$.

Let X be a set and let $\mathcal{S}$ be a σ-field in X. A projection valued measure on $\mathcal{S}$ is a function $E(\cdot)$ from $\mathcal{S}$ to projections on a Hilbert space H such that: (1) $E(\emptyset) = 0$, (2) $E(X) = id$,

(3) $E(A \cap B) = E(A)E(B)$, $A, B \in \mathcal{S}$ and (4) If $A_1, A_2, \dots$ is a disjoint sequence in $\mathcal{S}$ then

$$E(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} E(A_n) \quad \text{, a strongly convergent sum}$$

Let $(X, \mathcal{S})$ be a measurable space and $E(\cdot)$ a projection valued measure on $\mathcal{S}$ with values in $B(H)$. If $f = \sum_{j=1}^{n} a_j \chi_{B_j}$ is a simple complex valued measurable function on X, define the integral with respect to E as

$$\int f dE := \sum_{j=1}^{n} a_j E(B_j)$$

We now consider unbounded functions on Banach spaces. Let $T \colon X \to Y$ be a (possibly unbounded) function. If $\mathcal{D} \subseteq X$ is dense, and $T \colon D \to Y$, we say that T is **densely defined**. If T is a function from X to Y with domain $\mathcal{D}$, the **graph** of T is $G_T := \{(x, Tx) \colon x \in \mathcal{D}\}$. We say that T is **closed** if $G_T \subseteq X \times Y$ is closed.

Let H be a Hilbert space. Let $T \colon H \to H$ be linear and densely defined with domain $\mathcal{D}$. Define $\mathcal{D}_{T^*}$ as follows: $y \in \mathcal{D}_{T^*}$ if and only if the map $x \mapsto (Tx, y)$ is continuous from $\mathcal{D}$ to $\mathbb{C}$. For such a y, $\exists$ a unique $y^* \in H$ such that $(Tx, y) = (x, y^*)$ (by Riesz representation, Thm. 10.2). We define $T^* y := y^*$. Thus,

$$(Tx, y) = (x, T^* y) \quad \forall x \in \mathcal{D}_T, y \in \mathcal{D}_{T^*}$$

Write $A \subseteq B$ if $A = B$ on their common domains, and $\mathcal{D}_A \subseteq \mathcal{D}_B$. An important property is: if $A \subseteq B$, then $B^* \subseteq A^*$. (Since $B \supseteq A$, there are more x's to check that $x \mapsto (Bx, y)$ is continuous, so less y's will satisfy this condition, so $\mathcal{D}_{B^*} \subseteq \mathcal{D}_{A^*}$).

Let A be densely defined in H. A is **symmetric** if $A \subseteq A^*$ (i.e. $(Ax, y) = (x, Ay)$ $\forall x, y \in \mathcal{D}_A$). A is called **self-adjoint** if $A = A^*$ (i.e. $\mathcal{D}_A = \mathcal{D}_{A^*}$, and the operators agree on this domain.) A linear operator $A \colon H \to K$ between Banach spaces H, K is **compact** if the image of every bounded set has compact closure.

A **semigroup of operators** on a Banach space B is a function $s \mapsto T_s$ from $[0, \infty)$ to bounded operators on B such that: $T_0 = id$ and $T_{t+s} = T_t T_s$ for $s, t \geq 0$. A semigroup is called **strongly continuous** if for each $x \in B$, the function $t \mapsto T_t x$ is continuous from $[0, \infty)$ into B. Note that norm continuity of a semigroup (i.e. continuity of $\|T_t - T_s\|$) is stronger than strong continuity, since $\|T_t x - T_s x\| \leq \|T_t - T_s\| \|x\| \to 0$ as $t \to s$. A semigroup of operators is called a **contraction semigroup** if $\|T_t\| \leq 1$ for all $t \geq 0$. If $T_t \colon B \to B$ is a semigroup of linear operators, define $Af := \lim_{h \downarrow 0} \frac{T_h - id}{h} f$ with domain $\mathcal{D}_A := \{f \in B \colon Af \text{ exists}\}$. Then A is called the **infinitesimal generator** of the semigroup T_t. From the definition, we can see that A is a linear function on $\mathcal{D}_A$.

For $a \leq t \leq b$ let $u_t \colon [a, b] \to B$. We say u_t is **strongly continuous** at the point t if $\lim_{h \to 0} u_{t+h} = u_t$. If $\lim_{h \to 0} \frac{u_{t+h} - u_t}{h} = f$, then u_t is **strongly differentiable** at the point t and $f =: du_t/dt$. Let $a = t_0 < t_1 < \cdots < t_n = b$ and $\delta := \max_{1 \leq k \leq n} |t_k - t_{k-1}|$. If $\lim_{\delta \to 0} \sum_{k=0}^{n} u_{t_k}(t_k - t_{k-1})$ exists, then the function u_t is said to be **strongly integrable** on the interval $[a, b]$, and the limit is denoted by $\int_a^b u_t dt$.

If u_t is strongly continuous on $[a, b]$, then by mimicking the theory of the Riemann integral, u_t is strongly integrable on $[a, b]$. Let $T \colon B \to C$, $u_t \colon [a, b] \to B$. If u_t is strongly differentiable, then Tu_t is also strongly differentiable and $\frac{d(Tu_t)}{dt} = T\left(\frac{du_t}{dt}\right)$. If u_t is strongly integrable on $[a, b]$, then Tu_t is also strongly integrable on $[a, b]$ and $\int_a^b Tu_t dt = T(\int_a^b u_t dt)$.

If u_t is strongly integrable on the interval $[a, a+h]$ and strongly continuous from the right at a, then $\lim_{h\downarrow 0}\frac{1}{h}\int_a^{a+h} u_t dt = u_a$.

If du_t/dt is strongly continuous on $[a,b]$, then $\int_a^b \frac{du_t}{dt}dt = u_b - u_a$. To see this, let ξ be a linear functional. One of our properties above shows that $d(\xi u_t)/dt$ is Riemann integrable. So, using our other properties and the fundamental theorem of calculus for Riemann integrals,

$$\xi\left(\int_a^b \frac{du_t}{dt}dt\right) = \int_a^b \xi\left(\frac{du_t}{dt}\right)dt = \int_a^b \frac{d(\xi(u_t))}{dt} = \xi(u_b)-\xi(u_a)$$

Since the weak topology is Hausdorff, we conclude that $\int_a^b \frac{du_t}{dt}dt = u_b - u_a$, as desired. If u_t is strongly integrable on $[a,b]$ then u_{t-h} is strongly integrable on $[a+h, b+h]$ and $\int_{a+h}^{b+h} u_{t-h}dt = \int_a^b u_t dt$.

Let $A\colon B\to B$ be a bounded operator. Define the **exponential function** by $e^A := \lim_{N\to\infty}\sum_{n=0}^N \frac{1}{n!}A^n$. Since $\|A^n/n!\| \le \|A\|^n/n!$, the limit in the definition of e^A exists, and $\left\|e^A\right\| \le e^{\|A\|}$. Note that $e^{c\cdot id} = e^c\cdot id$. Also, if $AA' = A'A$, then $e^A e^{A'} = e^{A+A'}$ (using the same proof as in the case of the usual exponential function). We now show that $\|\frac{e^{tA}-id}{t} - A\| \to 0$ as $t\to 0$. Indeed, $\left\|e^{tA} - id - tA\right\| \le \sum_{n=2}^\infty \frac{\|A\|^n}{n!}t^n = e^{t\|A\|} - 1 - t\|A\|$. So, dividing by t and letting $t\to 0$ gives $\|\frac{e^{tA}-id}{t} - A\| \le \frac{e^{t\|A\|}-1}{t} - \|A\| \to 0$ as $t\to 0$. Now, if $AA' = A'A$, and if for $t\ge 0$ $\left\|e^{tA}\right\|\le 1, \left\|e^{tB}\right\| \le 1$, then for any $f\in B$, $\left\|e^{tA}f - e^{tB}f\right\| \le t\|Af - Bf\|$. To see this, note that

$$e^{tA}f - e^{tA'}f = (e^{\frac{tA}{n}})^n f - (e^{\frac{tA'}{n}})^n f = \left(\sum_{k=1}^n e^{\frac{k-1}{n}tA}e^{\frac{n-k}{n}tB}\right)(e^{\frac{tA}{n}}f - e^{\frac{tA'}{n}}f)$$

So, $\|e^{tA}f - e^{tA'}f\| \le n\|e^{tA/n}f - e^{tA'/n}f\|$. But $\frac{e^{tA/n}-id}{t/n} \to A$ and $\frac{e^{tA'/n}-id}{t/n}\to A'$ as $n\to\infty$, so $n(e^{tA/n} - e^{tA'/n}) \to t(A-B)$ as $n\to\infty$. So, letting $n\to\infty$ gives $\|e^{tA}f - e^{tA'}f\| \le t\|A-B\|$, as desired.

Theorem 10.1. ***(Hilbert space projections)*** *Let H be a Hilbert space, $W\subseteq H$ a closed convex set, $u\in H$, $M\subseteq H$ a closed subspace. Define $M^\perp := \{h\in H\colon (h,m) = 0, \forall m\in M\}$.*

(a) *$\exists\ v\in W$ with $\|u-v\| = \inf_{w\in W}\|u-w\|$.*
(b) *Every $u\in H$ can be uniquely written as $u = m+v$, $m\in M$, $v\in M^\perp$. (We therefore write $H = M\oplus M^\perp$)*
(c) *$(M^\perp)^\perp = M$*

Proof of (a): Let $a := \inf_{w\in W}\|u-w\|$. Let $\{w_n\}\subseteq W$ be a minimizing sequence. The parallelogram law says

$$\|2u - (w_n+w_m)\|^2 + \|w_n - w_m\|^2 = 2(\|u-w_m\|^2 + \|u-w_n\|^2) \to 4a^2 \quad (*)$$

as $m,n\to\infty$. But $\frac{1}{2}(w_n+w_m)\in M$, so $4\left\|u - \frac{1}{2}(w_n+w_m)\right\|^2 \ge 4a^2$, by definition of a. Then from the left side of $(*)$, $\|w_n - w_m\|^2\to 0$, so $\{w_n\}$ is Cauchy, i.e. $v := \lim w_n$ exists, and continuity of the norm gives $\|u-v\| = a$.

(b): First observe that $M^\perp$ is closed and $M\cap M^\perp = 0$ by definition of $M^\perp$. Uniqueness follows since $M\cap M^\perp = 0$. To get existence, use part (a) to find $v\in M$ with $\|u-v\| =$

$\inf_{m\in M}\|u-m\|$. Let $m \in M$ with $\|m\| = 1$. Then $v + (u-v,m)m \in M$, and by definition of v,

$$\|u-v\|^2 \leq \|u-v-(u-v,m)m\|^2 = \|u-v\|^2 - |(u-v,m)|^2$$

Thus $(u-v,m) = 0$, i.e. $u-v \in M^\perp$, so $u = v + (u-v)$.

(c): By definition, $M \subseteq M^{\perp\perp}$. For $u \in M^{\perp\perp}$, apply part (b) to get $u = m + m^\perp$, so $0 = m^\perp + (m-u)$, $m^\perp \in M^\perp, m-u \in M^{\perp\perp}$. Apply part (b) again to $M^\perp$, so $H = M^\perp \oplus M^{\perp\perp}$. By uniqueness of this decomposition for $0 \in H$, we conclude $m^\perp = 0$, $m-u = 0$, so $u = m \in M$, i.e. $M^{\perp\perp} = M$. □

Theorem 10.2. ***(Riesz Representation Theorem, Hilbert space version)*** *Let ℓ be a continuous linear functional on a Hilbert space H. Then $\exists$ unique $v \in H$ with $\ell(u) = (u,v)$. Also, $\|\ell\| = \|v\|$.*

Proof. Uniqueness is clear. For existence, if $\ell = 0$ take $v = 0$. Otherwise let $M = \{u\colon \ell(u) = 0\}$. Observe that M is a closed subspace and $M \neq H$. So we can let $w \neq 0$, $w \in M^\perp$, via Thm. 10.1(b). Then $\ell(w) \neq 0$. Let $v = (\overline{\ell(w)}/\|w\|^2)w$. Then $\ell(u - (\ell(u)/\ell(w))w) = 0$, so $u - (\ell(u)/\ell(w))w \in M$, and $v \in M^\perp$ so

$$\langle u, v\rangle = \left\langle u - \left(u - \frac{\ell(u)}{\ell(w)}w\right), v\right\rangle = \left\langle \frac{\ell(u)}{\ell(w)}w, \frac{\overline{\ell(w)}}{\|w\|^2}w\right\rangle = \ell(u)$$

Finally, Cauchy-Schwarz shows $|\ell(u)| = |(u,v)| \leq \|u\|\|v\|$ and $|(v,v)| = |\ell(v)| \leq \|\ell\|\|v\|$. □

Theorem 10.3. ***(Hahn-Banach, Abstract Form)*** *Let X be a real vector space, and let $p\colon X \to \mathbb{R}$ be a function such that $p(x+x') \leq p(x) + p(x')$ and $p(tx) = tp(x)$ $(x, x' \in X, t \geq 0, t \in \mathbb{R})$. Let f be a linear functional on a subspace Y of X with $f(y) \leq p(y)$ for all $y \in Y$. Then $\exists$ a linear functional F on X with $F(y) = f(y)$ $\forall$ $y \in Y$, and such that $F(x) \leq p(x)$ $\forall$ $x \in X$.*

(Hahn-Banach, Functional Form) *Let $Y \subseteq X$ be a vector subspace of the normed linear space X. Let $y^* \in Y^*$. Then $\exists$ $x^* \in X^*$ with $\|x^*\| = \|y^*\|$ and $x^*(y) = y^*(y)$, $\forall$ $y \in Y$.*

Proof of Abstract Form: Trick: use Zorn's Lemma, and then extend by one dimension. For any two linear functionals $(f_1, \mathrm{Dom}(f_1))$, $(f_2, \mathrm{Dom}(f_2))$ that are dominated by p, write $f_1 \leq f_2$ if $f_1 = f_2$ on $\mathrm{Dom}(f_1) \cap \mathrm{Dom}(f_2)$, and $\mathrm{Dom}(f_1) \subseteq \mathrm{Dom}(f_2)$. Thus, $f_1 \leq f_2$ if and only if f_2 extends f_1. Any chain of such extensions has an upper bound F. (For $x \in \cup_\alpha \mathrm{Dom}(f_\alpha)$, $x \in \mathrm{Dom}(f_{\alpha_0})$ for some α_0, so define $F(x) := f_{\alpha_0}(x)$.) By construction, F is then linear, since $x, y \in \mathrm{Dom}(F)$ implies $x, y \in \mathrm{Dom}(f_{\alpha_1})$ for some α_1. By Zorn's Lemma, we can let (f_0, Y_0) be a maximal extension. It remains to prove $Y_0 = X$, by contradiction.

Assume $y_1 \in X$, $y_1 \notin Y_0$. Let $Y_1 = \mathrm{span}\{y_1, Y_0\}$. A simple argument shows: every $x \in Y_1$ can be written uniquely as $x = y + cy_1$, $y \in Y_0, c \in \mathbb{R}$. Define f_1 on Y_1 by $f_1(y + cy_1) := f_0(y) + ck$ for some $k \in \mathbb{R}$, chosen below. It remains to show that $f_1 \leq p$ on Y_1. Let $y, y' \in Y_0$ and observe:

$$f_0(y') - f_0(y) = f_0(y'-y) \leq p(y'-y) \leq p(y'+y_1) + p(-y_1-y)$$

That is,

$$-p(-y_1-y) - f_0(y) \leq p(y'+y_1) - f_0(y') \qquad (*)$$

Let k be any number between: the sup of the left side of $(*)$ over $y \in Y_0$, and the inf of the right side of $(*)$ over $y' \in Y_0$. We need three cases. If $c = 0$, then $f_1(x) = f_0(x) \leq p(x)$. If

$c > 0$, then for $x = y + cy_1$, the choice of k for $y' = c^{-1}y$ gives

$$f_1(x) = f_0(y) + ck \leq f_0(y) + c(p(c^{-1}y) + y_1 - f_0(c^{-1}y)) = p(y + cy_1) = p(x)$$

If $c < 0$, then for $x = y + cy_1$, the choice of k for $y' = c^{-1}y$ gives

$$f_1(x) = f_0(y) + ck \leq f_0(y) + c(-p(y_1 - c^{-1}y) - f_0(c^{-1}y)) = p(y + cy_1) = p(x)$$

So, $f_1(x) \leq p(x)$, contradicting the maximality of f_0.

Proof of Functional Form: If the scalars are real, apply part (a) to $p(x) = \|y^*\|\,\|x\|$, $f = y^*$. If the scalars are complex, extend $\Re(y^*)$ to $F\colon X \to \mathbb{R}$ via part (a), then define $x^*(x) := F(x) - iF(ix)$. One can check that x^* is complex linear. Also, for $y \in Y$

$$(\Re\, y^*)(iy) + i(\Im\, y^*)(iy) = y^*(iy) = iy^*(y) = -(\Im\, y^*)(y) + i(\Re\, y^*)(y)$$

so $(\Re\, y^*)(iy) = -(\Im\, y^*)(y)$, implying that

$$x^*(y) = (\Re\, y^*)(y) - i(\Re\, y^*)(iy) = (\Re\, y^*)(y) + i(\Im\, y^*)(y) = y^*(y)$$

so that x^* extends y^*. If $x^*(x) = re^{i\theta}$, then $\Im(x^*(e^{-i\theta}x)) = 0 = -F(ie^{-i\theta}x)$, so

$$|x^*(x)| = x^*(e^{-i\theta}x) = F(e^{-i\theta}x) \leq \|y^*\|\,\|e^{-i\theta}x\| = \|y^*\|\,\|x\|$$

□

Theorem 10.4. ***(Hahn-Banach, Geometric form, a)*** *Let Y be a closed subspace of a normed linear space X. Let $x_0 \in X$, $x_0 \notin Y$. Then $\exists$ $x^* \in X^*$ such that $x^*(Y) = 0$ and $x^*(x_0) = 1$, with $\|x^*\| = d(x_0, Y)^{-1}$.*

(Hahn-Banach, Geometric form, b) *Let X be a real normed linear space, $C \subseteq X$ a convex open set containing 0. If $x_0 \in X$, $x_0 \notin C$, then $\exists$ $\phi\colon X \to \mathbb{R}$ continuous with $\phi(x_0) = 1$ and $\phi(C) < 1$*

(Hahn-Banach, Geometric form, c) *Let X be a locally convex linear space (that is, 0 has a neighborhood basis of convex sets), $A, B \subseteq X$ disjoint, convex, with A open. Then $\exists$ a continuous linear functional f on X and $\alpha \in \mathbb{R}$ such that $\Re(f(A)) < \alpha$ and $\Re(f(B)) \geq \alpha$.*

(Hahn-Banach, Geometric form, d) *Let X be a locally convex space, $A, B \subseteq X$ disjoint closed convex sets, A compact. Then $\exists$ a continuous linear functional f on X and $\alpha \in \mathbb{R}$ such that $\Re(f(A)) < \alpha$ and $\Re(f(B)) > \alpha$.*

Proof of (a): Let $d = d(x_0, Y)$, $Z = \operatorname{span}\{x_0, Y\}$. As above, every $x \in Z$ can be written uniquely as $x = y + cx_0$. Define $z^*(x) := c$. Then z^* is linear and for $c \neq 0$, and $\|x\| = |c|\,\|c^{-1}y + x_0\| \geq |c|\,d = d\,|z^*(x)|$. Thus, $\|z^*\| \leq d^{-1}$. Letting $\|x_0 - y_n\| \to d$, $\{y_n\} \in Y$ shows $1 = z^*(x_0 - y_n) \leq \|z^*\|\,\|x_0 - y_n\| \to d\,\|z^*\|$, so $\|z^*\| \geq d^{-1}$. Finally, apply Thm. 10.3(b) to extend z^*.

Proof of (b): (Sketch) On the one-dimensional space $\operatorname{span}\{x_0\}$, define $\phi(tx_0) = t$. Observe that ϕ is dominated by the norm induced by C ($\|y\|_C := \inf\{r > 0\colon y \in rC\}$). Apply Thm. 10.3(a).

Proof of (c): (Sketch) Assume scalars are real. Let $a_0 \in A, b_0 \in B$, $x_0 := b_0 - a_0$, $C := A - B + x_0$. Then C is convex, open, and it contains zero. Also, $x_0 \notin C$. Apply part (b) to C and x_0. Observe $\phi(b_0) = \phi(a_0) + 1$ and $\phi(a) < \phi(b) + \phi(a_0) - \phi(b_0) + 1$, $a \in A, b \in B$, so $\phi(a) < \phi(b)$. Let $\alpha = \inf_{b \in B} \phi(b)$, so $\phi(a) \leq \alpha \leq \phi(b)$. Since A is open, $\phi(a) = \alpha$ cannot occur, so $\phi(a) < \alpha$.

Proof of (d): We claim that $C := B - A$ is closed. Given the claim, note that $0 \notin C$, so C^c contains an open neighborhood U of 0. So apply part (c) to U and $B - A$ to get some real nonconstant continuous linear functional with $f(B-A) \geq c$ and $f(U) \leq c$. Let $x \in X$ such that $f(x) = 1$. For $\alpha > 0$ small, $\alpha x \in U$. Since $f(\alpha x) = \alpha$, $f(U) \supseteq [0, \varepsilon)$ for some small ε. Therefore, $f(B) - f(A) = f(B-A) \geq \varepsilon$, as desired.

We now prove that $C = B - A$ is closed. Let $d \in \overline{B-A}$, U a neighborhood of d, $A_U := \{a \in A \colon a \in B - U\}$. Note that $A_U \neq \emptyset$ since $d \in \overline{B-A}$. (Let $d_n \to d$, $d_n \in B - A$. Without loss of generality, $d_n \in U$. Then $d_n = b_n - a_n$, so $a_n = b_n - d_n$, $d_n \in U$. Since A is compact, after taking a subsequence we may assume that $a_n \to a$, $a \in A$, so $b_n \to b$ as well. Since B is closed, $b \in B$. Now $d = \lim(b_n - a_n) = b - a$, $b \in B, a \in A$, so $a = b - d \in A_U$, as desired.)

Observe $U \subseteq V$ implies $A_U \subseteq A_V$. So, any finite subset of the sets $\{A_U\}$ has the nonempty intersection. So, the compact sets $\{\overline{A_U}\}$ have a common element a_0. (If not, then $\cap_U A_U$ is empty. Write $A_U =: A \smallsetminus B_U$. Since $\cap_U A_U = \cap_U (A \smallsetminus B_U) = A \smallsetminus (\cup_U B_U)$, $\cup_U B_U = A$. But B_U is open and A is compact, so $A = \cup_{i=1}^n B_{U_i}$, implying $\cap_{i=1}^n A_{U_i} = \emptyset$, a contradiction.)

Let N be a neighborhood of 0. Then $U = N + d$ is a neighborhood of d, so $a_0 \in B - U = B - N - d$, so $(N + a_0) \cap (B - N - d) \neq \emptyset$. Shifting an N from the right to the left, we get $(N + N + a_0) \cap (B - d) \neq \emptyset$. Let M be a neighborhood of 0. By continuity of addition, there exists an open neighborhood N of 0 such that $N + N \subseteq M$. So, any neighborhood of a_0 intersects $B - d$. Since B is closed, $B - d$ is closed. Therefore, by the limit point definition of closedness, $a_0 \in B - d$, so $d \in B - a_0 \subseteq B - A$, as desired. □

Theorem 10.5. ***(Hahn-Banach Corollaries)*** *Let X be a normed linear space. If $x_0 \in X$, $x_0 \neq 0$, then $\exists\ x^* \in X^*$ with $\|x^*\| = 1$ and $x^*(x_0) = \|x_0\|$. Thus, setting $x_0 = x - x'$, we see that the weak topology is Hausdorff. Also, $\|x_0\| = \sup_{\|x^*\| \leq 1} |x^*(x_0)|$.*

Proof. First, use Thm. 10.4(a) with $Y = 0$. The formula for $\|x_0\|$ follows, since $\sup_{\|x^*\| \leq 1} |x^*(x_0)| \leq \sup_{\|x^*\| \leq 1} \|x^*\| \|x_0\| = \|x_0\|$. □

Theorem 10.6. ***(Uniform Boundedness Principle/ Banach-Steinhaus)*** *Let $\{L_\alpha\}$ be a set of bounded linear operators from a Banach space X into a normed linear space Y. Assume the L_α are pointwise bounded, i.e. $\|L_\alpha(x)\| \leq C_x$ for all α. Then $\exists\ C$ with $\|L_\alpha\| \leq C$ for all α (C independent of x).*

Proof. Let $F_n := \{x \in X \colon \|L_\alpha(x)\| \leq n, \forall \alpha\}$. Since $F_n = \cap_\alpha (\|\cdot\| \circ L_\alpha)^{-1}[0, n]$, F_n is closed. By assumption, $\cup_n F_n = X$. Since X is a complete metric space, Baire's Category Theorem shows that some F_N contains a nonempty open ball $B = B(b, 2r) \subseteq X$. Then $\|L_\alpha(x)\| \leq N$ for all α and for all $x \in B$. Dilating and translating B gives our result. Specifically, for $\|x\| \leq 1$ we have $rx + b \in B$, so

$$\|L_\alpha(x)\| = r^{-1} \|L_\alpha(rx)\| \leq r^{-1}(\|L_\alpha(rx + b)\| + \|L_\alpha(b)\|) \leq r^{-1}(N + C_b)$$

i.e. $\|L_\alpha\| \leq r^{-1}(N + C_b)$. □

Theorem 10.7. ***(Interior Mapping (Open Mapping) Principle)***

(a) *Let $L \colon X \to Y$ be a continuous linear operator between Banach spaces, which is onto. Then L is open (i.e. its inverse is continuous, if it exists).*

(b) *Suppose $\overline{L(B_X)}$ contains an open ball. ($B_X := \{x \colon \|x\| \leq 1\} = B(0,1)$). Then $\exists r > 0$ with $L(B_X) \supseteq r \cdot B_Y$, and (by scaling) $L(X) = Y$.*

Proof of (a): Trick: Baire, show L is open at zero by approximations, then translate. Let $B(0, n) = \{x \colon \|x\| < n\}$. Write $Y = L(X) = L(\cup_{n \geq 1} \overline{B(0, n)}) = \cup_{n \geq 1} L(\overline{B(0, n)})$. Thus

$Y = \cup_{n\geq 1}\overline{L(B(0,n))}$. Baire's Category Theorem says some $\overline{L(B(0,n))}$ contains an open ball. By scaling (and using continuity) $\overline{L(B(0,n))} = (2n)\overline{L(B(0,1/2))}$, so $\overline{L(B(0,1/2))}$ also contains an open ball V. A continuity argument shows that $V - V \subseteq \overline{L(B(0,1))}$. (Let $v, v' \in \overline{L(B(0,1/2))}$, let $v_n, v_n' \in L(B(0,1/2))$ with $v_n \to v, v_n' \to v'$. Then $v_n - v_n' \in L(B(0,1))$, $v_n - v_n' \to v - v' \in \overline{L(B(0,1))}$.) Since $V - V$ is open, it is also a neighborhood containing zero. Thus, $\overline{L(B(0,1))} \supseteq B(0,s)$ for some $s > 0$. Using continuity again,

$$\overline{L(B(0,t))} \supseteq B(0,st) \quad \forall t > 0 \qquad (*)$$

We now show that $L(\overline{B(0,c)}) \supseteq B(0,sc/2)$ for all $c > 0$. Let $y \in B(0,sc/2)$. Using $(*)$ with $t = c/2$ gives $x_1 \in \overline{B(0,c/2)}$ with $\|y - L(x_1)\| < 2^{-2}sc$. Suppose $x_1, \ldots, x_{n-1}$ satisfy $\|y - L(x_1 + \cdots + x_{n-1})\| < 2^{-n}sc$. Then $y - L(x_1 + \cdots + x_{n-1}) \in B(0, 2^{-n}sc)$, so $(*)$ with $t = 2^{-n}c$ gives $x_n \in \overline{B(0,2^{-n}c)}$ with $\|y - L(x_1 + \cdots + x_n)\| < 2^{-(n+1)}sc$. Since the norms of the x_n decrease exponentially, $x = \sum_{n=1}^{\infty} x_n$ exists and $\|x\| \leq c$. By definition of the x_i and since L is continuous we get $y = L(x)$, $x \in \overline{B(0,c)}$, as desired.

So, $L(B(0,2c)) \supseteq B(0,sc/2)$. By translating this result, we see: for U open, each $w \in L(U)$ has a neighborhood contained in $L(U)$, as desired.

Proof of (b): Argue as in part (a). □

Theorem 10.8. ***(Closed graph theorem)*** *Let $L\colon X \to Y$ be a linear function between Banach spaces. Suppose the graph $G := \{(x, L(x))\colon x \in X\}$ of L is closed in $X \times Y$. (If $x_n \to x$ and $f(x_n) \to y$, then $x \in Dom(L)$ and $f(x) = y$). Then L is bounded.*

Proof. Trick: Use a projection. Define the Banach space $X \oplus Y$ via the norm $\|(x,y)\| := \|x\|_X + \|y\|_Y$. By assumption, $G \subseteq X \oplus Y$ is closed. In particular, G is a Banach space. Define $P\colon G \to X$ by $P((x, L(x))) = x$, and observe that P is linear, continuous and bijective. Therefore, $P^{-1}(x) = (x, L(x))$ is continuous, from the Interior Mapping Principle (Thm. 10.7(a)). Finally, note that the projection $\Pi\colon X \oplus Y \to Y$ defined by $\Pi(x,y) := y$ is bounded, so $\Pi|_G$ is bounded, so $(\Pi|_G) \circ P^{-1} = L$ is bounded. □

Theorem 10.9. ***(Choquet)*** *Let X be a metrizable, compact, convex subset of a locally convex space E. Let $x_0 \in X$. Then $\exists$ a probability measure $\mu = \mu_{x_0}$ on X that is supported on the extreme points $\mathcal{E}(X)$ of X, and such that $x_0 = \int_{\mathcal{E}(X)} x d\mu(x)$ weakly. (That is, for any continuous linear functional F on X, $F(x_0) = \int_{\mathcal{E}(X)} F(x) d\mu(x)$).*

Proof. Strategy: consider $X^* \subseteq C(X)$, construct a strictly convex function, apply Hahn-Banach. By Thm. 10.10, $C(X)$ is separable. Let $A \subseteq C(X)$ be the set of continuous affine functions. Let $\{h_n\}_{n=1}^{\infty} \subseteq A$ be dense on the unit sphere of A. Observe that $f = \sum_{n\geq 1} 2^{-n} h_n^2$ converges uniformly and is strictly convex and nonnegative. Let $B = \operatorname{span}\{A, \overline{f}\}$. For $g \in C(X)$, define $\overline{g}(x) := \inf\{h(x)\colon h \in A, h \geq g\}$ (this is the upper concave envelope of the graph of g). Since $\overline{g + u} \leq \overline{g} + \overline{u}$, $p(g) := \overline{g}(x_0)$ is a subadditive function on $C(X)$, which is also homogeneous. Define $\ell\colon B \to \mathbb{R}$ by $\ell(h + rf) = h(x_0) + r\overline{f}(x_0)$. On B, note that $\ell \leq p$. (For $r > 0$, $h + r\overline{f} = \overline{h + rf}$, and for $r < 0$, $h + rf$ is concave, so $\overline{h + rf} = h + rf \geq h + r\overline{f}$).

Since $\ell \leq p$ on B, Hahn-Banach (Thm. 10.3(a)) gives $m \in C(X)^*$ which extends ℓ and remains dominated by p. Since $m \leq p$, m is nonpositive on nonpositive functions, and Riesz's Representation Theorem shows $m(f) = \int f d\mu$ for μ a regular Borel (probability) measure (using $1 \in A$, so $m(1) = 1$). Now, $f \leq \overline{f}$ so $\mu(f) \leq \mu(\overline{f})$. Conversely, if $h \in A$

and $h \geq f$, then $h \geq \overline{f}$, so $h(x_0) = m(h) = \mu(h) \geq \mu(\overline{f})$. So, taking the infimum over $h \in A, h \geq f$, we get $\overline{f}(x_0) \geq \mu(\overline{f})$. But $\mu(f) = m(f) = \overline{f}(x_0)$, so $\mu(f) \geq \mu(\overline{f})$. Thus, $\mu(f) = \mu(\overline{f})$, i.e. μ vanishes outside $E := \{x \colon f(x) = \overline{f}(x)\}$. Finally, E is supported on the extreme points of X, since if $x = (1/2)(y + z)$, then strict convexity of f implies $f(x) < (1/2)(f(y) + f(z)) \leq (1/2)(\overline{f}(y) + \overline{f}(z)) \leq \overline{f}(z)$. (And actually, E is equal to the extreme points). □

Theorem 10.10. ***(Separability Conditions)*** *Let X be a locally compact separable metric space, with μ a Borel measure on X. For $1 \leq p < \infty$ we have*

(a) $(C_c(X), \|\cdot\|_{sup})$ *is a separable normed linear space*
(b) $L_p(X, \mu)$ *is separable*

Theorem 10.11. ***(Characterization of Locally Convex Spaces)*** *Let X be a topological linear space. Let $\mathcal{N}$ be the family of continuous semi-norms on X. Let $X^{\mathcal{N}}$ be the Fréchet space formed by X and $\mathcal{N}$. Then $X^{\mathcal{N}} = X$ iff X is locally convex*

Theorem 10.12. ***(Alaoglu Theorem/ Banach-Alaoglu)*** *Let X be a normed linear space. Then the unit ball $B_{X^*} = \{x^* \colon \|x^*\| \leq 1\}$ of X^* is weak* compact.*

Proof. Let A be the set of scalar valued functions ξ on X with $\|\xi(x)\| \leq \|x\|$ for all $x \in X$. Equivalently, $A = \prod_{x \in X}\{\lambda \in \{\text{scalars}\} \colon |\lambda| \leq \|x\|\}$. Then A with the product topology is compact by Tychonoff's Theorem. By the definition of the product topology, a basic open neighborhood of some ξ_0 is $\{\xi \colon |\xi(x_j) - \xi_0(x_j)| < \varepsilon, j = 1, \ldots, n\}$. Now for fixed $x \in X$, the projection map $\xi \mapsto \xi(x)$ is continuous, from A (with the product topology) to scalars. (Given ξ in the inverse image of a small open interval, ξ is contained in an open set in A). Consider the natural embedding $B_{X^*} \subseteq A$. By the definition of the weak* topology, the topology induced by $B_{X^*} \subseteq A$ is exactly the weak* topology.

Putting everything together, let $x, y \in X$, α, β scalars, and observe: $\xi(\alpha x + \beta y) - \alpha\xi(x) - \beta\xi(y)$ is a continuous function of ξ, from A to scalars (since it is a composition of continuous functions). Therefore, $\{\xi \colon \xi(\alpha x + \beta y) - \alpha\xi(x) - \beta\xi(y) = 0\}$ is closed in A (being an inverse image of zero). Therefore,

$$B_{X^*} = \cap_{x,y,\alpha,\beta}\{\xi \colon \xi(\alpha x + \beta y) - \alpha\xi(x) - \beta\xi(y) = 0\}$$

is closed in the compact set A. □

Remark 10.13. If X is separable then $(B_{X^*}, \sigma(X^*, X))$ is metrizable. Using the homeomorphism $\iota \colon (X, \sigma(X, X^*)) \to (X^{**}, \sigma(X^{**}, X^*))$, we see: if X^* is separable then $(B_X, \sigma(X, X^*))$ is metrizable. To prove the first assertion, let $\{x_n\}_{n \geq 1}$ be dense in B_X. Then, define $\rho(x_1^*, x_2^*) = \sum_{n \geq 1} 2^{-n}\, |x_1^*(x_n) - x_2^*(x_n)|$. Essentially by definition (using open balls), ρ is $\sigma(X^*, X)$ continuous. That is, a point x in an open ρ ball U has $x \in V \subseteq U$, V a basic open set in the weak* topology. And the reverse inclusion of open sets holds by scaling appropriately. So, for X separable, B_{X^*} is sequentially compact.

Theorem 10.14. ***(Properties of Faces)***

(a) $\cap\{faces\} = face$
(b) *A face of a face is a face.*
(c) *For $A \colon H \to \tilde{H}$ continuous and linear, if K is compact and convex, then $A(K)$ is too.*

(d) *If $\widetilde{F}$ is a face of $\widetilde{K} := A(K)$, then $F := K \cap A^{-1}(\widetilde{F})$ is a face of K.*

(e) *If $\widetilde{F} \subsetneqq \widetilde{K} := A(K)$, then $F := K \cap A^{-1}(\widetilde{F}) \neq K$.*

(f) *Let X be a topological vector space such that X^* separates points of X (e.g. X is normed linear, using Thm. 10.5, or X is locally convex, using Thm. 10.11). Then a nonempty extreme face is an extreme point.*

Proof of (f): (Properties (a),(b),(c) and (e) are routine. For (d), apply A to a segment containing a point of F.). View X as a vector space over $\mathbb{R}$. Let $\xi \in X^*$ be real valued. Let $F \subseteq K$ be an extreme face. By definition of a face, F is compact, convex. Since $\xi(F)$ is compact, convex from (c), $\xi(F) = [a, b]$. Now, $\{a\}$ is a face of $\xi(F)$, so $F \cap \xi^{-1}(a)$ is a nonempty face of F. Since F is extreme, $F = F \cap \xi^{-1}(a)$, i.e. $\xi(F) = \{a\}$. If F had two distinct points, we could choose ξ which separates them, by assumption on X^*. Therefore, F has only one point. □

Theorem 10.15. ***(Krein-Milman)*** *Let X be a locally convex topological vector space. If K is compact and convex, then $\widehat{K} = K$.*

Proof. We will use Thm. 10.14 multiple times, without further note. Recall that $\widehat{K}$ is the intersection of all compact, convex sets containing the extreme points of K. So, by definition, $\widehat{K} \subseteq K$. We will see below that $\widehat{K} \neq \emptyset$. For ξ a real continuous linear functional, $\xi(\widehat{K}) \subseteq \xi(K) = [a, b]$. Now, $K \cap \xi^{-1}(a)$ is a face of K. Let T be the set of non-empty faces of $K \cap \xi^{-1}(a)$. For $t_1, t_2 \in T$, write $t_1 \leq t_2$ if $t_2 \subseteq t_1$ (i.e. we reverse the inclusion). Every chain has a nonempty upper bound, since the intersection of nested compact sets is nonempty. (Recall: we take the definition of compactness, take the contrapositive, and then take complements and apply de Morgan's law. Finally, use the finite intersection property.) By Zorn's Lemma, $\exists$ a nonempty extreme face in $K \cap \xi^{-1}(a)$, which is therefore an extreme point. Thus, $(K \cap \xi^{-1}(a)) \cap \widehat{K} \neq \emptyset$, so $a \in \xi(\widehat{K})$, so the endpoints of $\xi(K)$ are in $\xi(\widehat{K})$, so (by convexity of $\widehat{K}$), $\xi(K) \subseteq \xi(\widehat{K})$. Thus, $\xi(K) = \xi(\widehat{K})$ for all real linear functionals ξ, so $\widehat{K} = K$ from Hahn-Banach in Geometric Form (Thm. 10.4(d)). (If $\widehat{K} \subsetneqq K$, let $x \in K \setminus \widehat{K}$, so Hahn-Banach gives ξ with $\xi(\widehat{K}) < \alpha, \xi(x) > \alpha$.) □

Remark 10.16. Given X, Y, we consider whether or not $Y = X^*$. If so, then $B_Y = B_{X^*}$ is compact and convex, so this theorem says $\widehat{B_Y} = B_Y$. Thus, analyzing the extreme points of B_Y (existence, nonexistence, finiteness, etc.) can tell us whether not not $Y = X^*$ for any X. For example, the unit ball of $C([0, 1], \mathbb{R})$ only has extreme points ± 1, and that of $L_1([0, 1], \mathbb{R})$ has no extreme points.

Theorem 10.17. ***(Weak Topology Equivalence)***

(a) *The $\sigma(X, X^*)$ topology on X is the weakest topology such that all norm-continuous linear functionals are still continuous. Similarly, the $\sigma(X^*, X)$ topology on X^* is the weakest topology which makes all functionals in $\iota(X) \subseteq X^{**}$ continuous.*

(b) *The space of all continuous linear functionals on $(X, \sigma(X, X^*))$ equals X^*.*

(c) *Let $\phi_0, \ldots, \phi_n$ be linear forms on a linear space X (i.e. ignore any topology). Then the following are equivalent: (i) $\phi_0 \in span\{\phi_j\}_{j=1}^n$, and (ii) $\ker \phi_0 \supseteq \cap_{j=1}^n \ker \phi_j$*

(d) *The space of all continuous linear functionals on $(X^*, \sigma(X^*, X))$ equals X.*

Proof of (a): Apply definitions.

(b): The forward direction follows since the weak topology is weaker than the strong topology. The reverse is part (a).

(c): (i) easily implies (ii). For the reverse, define $\pi \colon X \to \mathbb{K}^n$ by $\pi(x) = \{\phi_j(x)\}_{j=1}^n$ ($\mathbb{K}$ = scalars). Then $\ker \pi = \cap_{j=1}^n \ker \phi_j$, so (ii) implies that ϕ_0 induces a linear form $\Lambda \colon \mathbb{K}^n \to \mathbb{K}$, i.e. $\phi_0 = \Lambda\pi$. (For $z = \{\phi_j(x)\}_{j=1}^n$, z is in the linear space $\pi(X)$, so define $\Lambda(z) = \phi_0(x)$. By (ii), Λ is well-defined. Λ is linear on $\pi(X)$, so extend Λ linearly to $\mathbb{K}^n$.) From linear algebra, $\phi_0(x) = \sum_{j=1}^n \alpha_j \phi_j(x)$.

(d): Let ϕ be a linear functional on X^* continuous in $\sigma(X^*, X)$. Then $\{x^* \in X^* \colon |\phi(x^*)| < 1\} \supseteq \{x^* \in X^* \colon |x_j(x^*)| < \varepsilon, j = 1, \ldots, n\}$ for some $\varepsilon > 0$, $x_1, \ldots, x_j \in X^*$. In particular, viewing $x_j \in X^{**}$, $\ker \phi \supseteq \cap_{i=1}^j \ker(x_j)$. Finally, apply part (c). □

Theorem 10.18. ***(Goldstine)*** *The closed unit ball of X is $\sigma(X^{**}, X^*)$-dense in the closed unit ball of X^{**}. (Here we identify B_X with $\iota(B_X) \subseteq X^{**}$)*

Proof. Trick: use X^{***}, Hahn-Banach, contradiction. Let $V = \overline{B_X}^{\sigma(X^{**},X^*)}$. Assume for the sake of contradiction $\exists\ x^{**} \in X^{**}$ with $\|x^{**}\| \leq 1$, $x^{**} \notin V$. Alaoglu (Thm. 10.12) says V is $\sigma(X^{**}, X^*)$-compact. So, Hahn-Banach, geometric form (Thm. 10.4(d)) gives ϕ a linear functional on $(X^{**}, \sigma(X^{**}, X^*))$ with $\phi(x^{**}) > \sup\{\phi(v) \colon v \in V\}$. Theorem 10.17(d) says $\phi(x^{**}) = x^{**}(x_0^*)$, for some $x_0^* \in X^*$. But then

$$\|x_0^*\| = \sup_{\substack{x \in X \\ \|x\| \leq 1}} |x_0^*(x)| \leq \sup_{v \in V} |v(x_0^*)| = \sup_{v \in V} |\phi(v)| < \phi(x^{**}) = x^{**}(x_0^*) \leq \|x_0^*\|$$

which is a contradiction. □

Theorem 10.19. ***(Reflexive equivalences)*** *Let X be a Banach space. The following are equivalent:*

(a) *X is reflexive ($\iota(X) = X^{**}$)*
(b) *X^* is reflexive*
(c) *B_X is $\sigma(X, X^*)$ compact*
(d) *Every subspace of X is reflexive*
(e) *Every quotient space of X is reflexive*

Proof: (a) implies (c): Using ι, we know that $(B_X, \sigma(X, X^*))$ is homeomorphic to $(B_{X^{**}}, \sigma(X^{**}, X^*))$ so apply Alaoglu (Thm. 10.12).

(c) implies (a): $\iota(B_X)$ is compact, so apply Goldstine (Thm. 10.18).

(d) implies (a): True by definition.

(a) implies (d): Let $Y \subseteq X$ be a norm-closed subspace. By, say, Hahn-Banach (Thm. 10.5), Y is $\sigma(X, X^*)$ closed. By (c), $B_Y \subseteq B_X$ is $\sigma(X, X^*)$ compact. By restriction of X^*, B_Y is $\sigma(Y, Y^*)$ compact, so Y is reflexive, since (c) implies (a).

(a) implies (b): By reflexivity, $(X^*, \sigma(X^*, X)) = (X^*, \sigma(X^*, X^{**}))$, so Alaoglu (Thm. 10.12) shows B_{X^*} is $\sigma(X^*, X^{**})$ compact, so use: (c) implies (a).

(b) implies (a): X^* reflexive implies X^{**} is reflexive (since (a) implies (b)) so X is reflexive, using $\iota(X) \subseteq X^{**}$ and that (a) implies (d).

(e)$\Longleftrightarrow$(a): $(X/Y)^* \subseteq X^*$ from Hahn-Banach, so use: (a)$\Longleftrightarrow$(b)$\Longleftrightarrow$(d). □

Remark 10.20. Let (S, μ) be a σ-finite measure space with infinitely many disjoint sets of positive measure. Then L_1, L_∞ are not reflexive. Since $(L_1)^* = L_\infty$, the previous theorem says, it suffices to show that L_1 is not reflexive. So it suffices to find $\phi \in (L_\infty)^*$ with $\phi \notin \iota(L_1)$.

Let $\{E_n\}_{n=1}^{\infty}$ be disjoint sets of positive measure. Let $Y \subseteq L_\infty$ be the set of functions that are constant on each E_n, with $y \in Y$ satisfying: $\lim_{n\to\infty} y(E_n)$ exists. Let $y^*(y) = \lim_{n\to\infty} y(E_n)$. Then y^* has norm 1, and Hahn-Banach (Thm. 10.3) gives an extension $x^* \in (L_\infty)^*$. For the sake of contradiction, assume $\exists\ g \in L_1$ such that $x^*(f) = \int_S fgd\mu$. In particular, $x^*(f) = \int_S fgd\mu$ for all $f \in Y$. Then $x^*(1_{E_n}) = y^*(1_{E_n}) = 0$, but

$$1 = y^*(1_{\cup_{n=1}^\infty E_n}) = x^*(1_{\cup_{n=1}^\infty E_n}) = \int_S g(1_{\cup_{n=1}^\infty E_n}) = \sum_{n=1}^{\infty}\int_{E_n} g = \sum_{n=1}^{\infty} x^*(1_{E_n}) = 0$$

a contradiction. So, no such g exists, i.e. $x^* \notin \iota(L_1)$.

Theorem 10.21. ***(Principle of Local Reflexivity)*** *Let X be a Banach space and let $E \subseteq X^{**}$ and $F \subseteq X^*$ be finite dimensional subspaces. Given $\varepsilon > 0$ there exists an operator $T\colon E \to X$ such that*

(1) $\|T\|\,\|T^{-1}|_{T(E)}\| \leq 1+\varepsilon$
(2) $T|_{E\cap X} = id$
(3) $f(Te) = e(f)$ *for all* $f \in F$ *and* $e \in E$

Proof. We begin with the following Lemma:

Let $\{A_j\}_{j=1}^N$ be bounded, norm-open convex subsets of X and let $\tilde{A}_j$ be the norm interior of the $\sigma(X^{**}, X^*)$-closure of A_j in X^{**}.

(a) If $\cap_{j=1}^N \tilde{A}_j \neq \emptyset$ then $\cap_{j=1}^N A_j \neq \emptyset$
(b) If we have a map $T\colon X \to Y$ with Y a finite dimensional Banach space then $T^{**}(\cap_{j=1}^N \tilde{A}_j) = T(\cap_{j=1}^N A_j)$.

Proof of Lemma, (a): We prove the contrapositive statement. Let $X_N = \oplus_{i=1}^N (X)$ and let $A = \{\{x_j\}_{j=1}^N \in X_N \colon x_j \in A_j, j = 1,\dots,N\}$. Then $A \subseteq X_N$ is bounded, norm-open and convex. If $\cap_{j=1}^N A_j = \emptyset$ then $A \cap V = \emptyset$ for $V = \{\{x_j\}_{j=1}^N \in X_N \colon x_j = x_1 \text{ for } j = 1,\dots,N\}$. Let $\tilde{A} = \{\{x_j^{**}\}_{j=1}^N \in X_N^{**} \colon x_j^{**} \in \tilde{A}_j, j = 1,\dots,N\}$ and let $V^{**} = \{\{x_j^{**}\}_{j=1}^N \in X_N^{**} \colon x_j^{**} = x_1^{**} \text{ for } j = 1,\dots,N\}$.

If $A \cap V = \emptyset$ then since V is a closed subspace, Hahn-Banach, Thm. 10.4(a) says there exists $\phi = (\phi_1,\dots,\phi_N) \in X_N^*$ with $\phi|_V = 0$ and $\phi(a) > 0$ for all $a \in A$. By Goldstine, Thm. 10.18, A is $\sigma(X^{**}, X^*)$ dense in $\tilde{A}$, so $\phi(a^{**}) \geq 0$ for all $a^{**} \in \tilde{A}$. But $\tilde{A}$ is open so $\phi(\tilde{A})$ is open, so $\phi(a^{**}) > 0$ for all $a^{**} \in \tilde{A}$. But B_V is $\sigma(X^{**}, X^*)$ dense in $B_{V^{**}}$, so $\phi(V^{**}) = 0$, so $\tilde{A} \cap V^{**} = \emptyset$, i.e. $\cap_{j=1}^N \tilde{A}_j = \emptyset$.

Proof of Lemma, (b): Note that $T^{**}(\cap_{j=1}^N \tilde{A}_j)$ and $T(\cap_{j=1}^N A_j)$ are open, convex sets and since $\iota(A_j) \subseteq \tilde{A}_j$, we have $T(\cap_{j=1}^N A_j) \subseteq T^{**}(\cap_{j=1}^N \tilde{A}_j)$. Suppose this containment is strict. Using Hahn-Banach, Thm. 10.4(d), there exists $p \in T^{**}(\cap_{j=1}^N \tilde{A}_j)$ and there exists a functional ϕ on Y^{**} and $\alpha > \beta$ such that $\phi(p) > \alpha > \beta > \phi(T(\cap_{j=1}^N A_j))$.

Let $z^{***} \in X^{***}$ with $z^{***} = T^{***}(\phi)$. Let $A_j^+ = A_j \cap \{x \in X \colon z^{***}(x) > \alpha\}$, $\tilde{A}_j^+ = \tilde{A}_j \cap \{x^{**} \in X^{**} \colon z^{***}(x^{**}) > \alpha\}$. By definition of ϕ, z^{***}, T^*, we have $\cap_{j=1}^N A_j^+ = \emptyset$. However, $\cap_{j=1}^N \tilde{A}_j^+ \neq \emptyset$, contradicting (a) and proving (b). To verify the nonempty intersection, write $p = T^{**}(z)$, so that $z^{***}(z) = T^{***}(\phi)(z) = \phi(T^{**}(z)) = \phi(p) > \alpha$, i.e. $z \in \cap_{j=1}^N \tilde{A}_j^+$.

Proof of Theorem: Let $\dim(E) = n$ and $\dim(E \cap X) = n-k$. Let $\{x_j^{**}, e_j^*\}_{j=1}^n$ be a biorthogonal system in $E \times E^*$ such that $\operatorname{span}\{x_j^{**}\}_{j=k+1}^n = E \cap X$ and $\|x_j^{**}\| = 1$. From biorthogonality, the identity $id\colon E \to X^{**}$ can be written as $id(e) = \sum_{j=1}^n e_j^*(e)x_j^{**}$. We will find $x_1, \dots, x_k \in X$ such that $T\colon E \to X$ defined by $T(e) = \sum_{j=1}^k e_j^*(e)x_j + \sum_{j=k+1}^n e_j^*(e)x_j^{**}$ satisfies the conclusion of the theorem. Note that property (2) is already satisfied by this T. Let $Z = \oplus_{i=1}^k(X)$ and let $\delta > 0$. Fix the following three sets: $\{f_j\}_{j=1}^M$ a basis of F, $\{x_j^*\}_{j=1}^R \subseteq B_{X^*}$ a set such that for every $e \in E$, $\|e\| \leq (1+\delta)\sup\{|x_j^*(e)|\colon j = 1, \dots, R\}$, and $\{e_j\}_{j=1}^N$ a δ-net in B_E.

Write $e_j = \sum_{r=1}^n \lambda_r^j x_r^{**}$. For $j = 1, \dots, N$, define $C_j \subseteq Z$ by

$$C_j := \{\{x_s\}_{s=1}^k\colon \left\|\sum_{s=1}^k \lambda_s^j x_s + \sum_{s=k+1}^n \lambda_s^j x_s^{**}\right\| < (1+\delta)\|e_j\|$$
$$\text{and } \|x_s\| < 1+\delta,\ s = 1, \dots, k\}$$

Note that the C_j are norm-open, bounded and convex. Let $\tilde{C}_j$ denote the norm interior of the $\sigma(X^{**}, X^*)$-closure of C_j in X^{**}. Since $\{x_s^{**}\}_{s=1}^k \in \cap_{j=1}^N \tilde{C}_j \subseteq Z^{**}$, part (a) of the Lemma gives $\{z_s\}_{s=1}^k \in \cap_{j=1}^N C_j \neq \emptyset$.

Let $\mathbb{K} = \{\text{scalars}\}$ and define $S\colon Z \to \mathbb{K}^{M\cdot k} \oplus \mathbb{K}^{R\cdot k}$ by

$$S(\{x_s\}_{s=1}^k) := \{f_j(x_s), x_k^*(x_s)\}_{j=1,\dots,M,k=1,\dots,R,s=1,\dots,k}$$

From part (b) of the Lemma, there exists $\{z_s\}_{s=1}^k \in \cap_{j=1}^N C_j$ with $S(\{z_s\}_{s=1}^k) = S^{**}(\{z_s^{**}\}_{s=1}^k)$. By the last equality, from our formula for id, by choice of $\{f_j\}$ and by definition of T, property (3) follows. It remains to show that property (1) holds. Using again the equality $S(\{z_s\}_{s=1}^k) = S^{**}(\{z_s^{**}\}_{s=1}^k)$, we see that $x_j^*(Te) = x_j^*(e)$. So, by the defining properties of $\{x_j^*\}$,

$$\|Te\| \geq \sup_{j=1,\dots,R} \left|x_j^*(Te)\right| = \sup_{j=1,\dots,R} \left|x_j^*(e)\right| \geq (1+\delta)^{-1}\|e\| \qquad (*)$$

Given $e \in B_E$, fix e_j with $\|e - e_j\| \leq \delta$. Recall the formula $e_j = \sum_{r=1}^n \lambda_r^j x_r^{**}$. Since $\lambda_r^j = e_r^*(e_j)$ and since $\{z_s\}_{s=1}^k \subseteq C_j$, the definitions of T and C_j show that $\|Te_j\| \leq (1+\delta)\|e_j\|$. Therefore $\|Te\| \leq \|Te_j\| + \|T(e - e_j)\| \leq (1+\delta)\|e_j\| + \delta\|T\| \leq (\|e\| + \delta)(1+\delta) + \delta\|T\|$, so

$$\|Te\| \leq \|e\| + 2\delta + \delta^2 + \delta\|T\| \qquad (**)$$

We crudely estimate $\|T\|$ by $\|T\| \leq (1+\delta)\sum_{j=1}^n \left\|e_j^*\right\| \leq 2\sum_{j=1}^n \left\|e_j^*\right\|$ for $\delta < 1$. So, for δ sufficiently small, $(*)$ and $(**)$ give condition (1), proving the theorem. □

Theorem 10.22. ***(Spectral Mapping theorem)*** *Let $a \in \mathcal{A}$, $\mathcal{A}$ a complex Banach algebra with identity. Recall: $\sigma(a) = \{\lambda \in \mathbb{C}\colon x - \lambda \cdot 1$ is not invertible$\}$, and $\sigma(a)$ is compact and nonempty. If p is a polynomial, then $p(\sigma(a)) = \sigma(p(a))$.*

Proof. See Thm. 10.26(4). □

Theorem 10.23. ***(Facts about ideals, etc.)*** *An **ideal** $I \subseteq \mathcal{B}$ is closed under addition (within I) and multiplication by elements of $\mathcal{B}$. An ideal is called **maximal** if $I \neq \mathcal{B}$ and I is not contained in any larger proper ideal. Define the **singular elements** $\mathcal{S}$ as the union of all maximal ideals in $\mathcal{B}$.*

1. *If $\{0\}$ is the only proper ideal in $\mathcal{B}$, then $\mathcal{B}$ is a field.*
2. *If I is a maximal ideal in $\mathcal{B}$, then $\mathcal{B}/I$ is a field.*
3. *For $a \in \mathcal{B}$, a is invertible if and only if a does not belong to any maximal ideal.*

(4) *If $\alpha \in \widetilde{\mathcal{B}}$, then $\alpha(1) = 1$. (Use $\alpha(1) = \alpha(1)\alpha(1)$.)*
(5) *If I is a proper ideal in $\mathcal{B}$, then $\overline{I}$ is a proper ideal. (Let $a_n \to a$, $a_n, b \in I$, $a \in \overline{I}$. Then $ba = \lim ba_n \in \overline{I}$. Also, $I \subseteq \mathcal{S}$, $1 \notin \mathcal{S}$.)*
(6) *If I is a maximal ideal then $I = \overline{I}$. (Follows from (5)).*
(7) *Let B be a Banach space, and let K be a closed subspace of B. Then B/K with $\|x + K\| := \inf\{\|y\| : y \in x + K\}$ is a Banach space. Moreover, if B is a Banach algebra with identity and K is a closed proper two sided ideal in B, then B/K is a Banach algebra, with the norm as before.*
(8) *For ξ a linear function on a Banach space B, ξ is continuous if and only if $\ker \xi$ is closed.*
(9) *Any character α (i.e. $\alpha \in \widetilde{\mathcal{B}}$) is continuous.*
(10) *(Gelfand-Mazur) The only complex Banach algebra with unit which is a division algebra is $\mathbb{C}$.*
(11) *There is a one to one correspondence between characters and maximal ideals given by $\alpha \mapsto \ker \alpha$*

Proof of (7): The triangle inequality is mostly routine. Without loss of generality, suppose $\{x_n + K\}_{n\geq 1}$ is Cauchy with $\|x_n - x_m + K\| < 2^{-\min(n,m)}$. By definition of the norm, we may take $\{k_n\}_{n\geq 1} \subseteq K$ with $\|x_n - x_{n+1} - k_{n+1}\| < 2^{-n+1}$. Let $z_n = x_n + \sum_{j=1}^{n} k_n$. Then $\{z_n\}_{n\geq 1}$ is Cauchy, so $z_n \to x$ for some $x \in B$. We claim that $x_n + K \to x + K$ in B/K. This follows since

$$\|x_n - x + K\| \leq \|x_n - z_n + K\| + \|z_n - x + K\| = \|z_n - x + K\| \leq \|z_n - x\|$$

The second assertion is mostly routine.

Proof of (8): If ξ is continuous, $\xi^{-1}(0) = \ker \xi$ is closed. Conversely, if $K := \ker \xi$ is closed, then from (7), B/K is a Banach space. Define $\zeta \colon B/K \to \{\text{scalars}\}$ by $\zeta(x + K) = \xi(x)$ (this is well defined and linear with trivial kernel). For $y, z \in B/K, y, z \neq 0$ we therefore have $\zeta(\zeta(z)y - z\zeta(y)) = 0$, so $z = \frac{\zeta(z)}{\zeta(y)}y$, i.e. B/K is one-dimensional. So, ζ is clearly continuous. Also, $\Pi \colon B \to B/K$ defined by $\Pi(x) = x + K$ is also easily continuous, so $\xi = \zeta \circ \Pi$ is continuous.

Proof of (9): For α a character, $I := \{a \in \mathcal{B} \colon \alpha(a) = 0\} = \ker \alpha$ is an ideal that is proper, since $\alpha(1) = 1$. For any $a \in \mathcal{B}$, write $a = (a - \alpha(a)1) + \alpha(a)1$. This shows that $\dim_{\mathbb{C}}(\mathcal{B}/I) = 1$, so I is maximal, hence closed (by (6)), so α is continuous by (8).

Proof of (10): Let $a \in \mathcal{A}$, $\lambda \in \sigma(a)$. Then $a - \lambda 1$ is not invertible, so $a - \lambda 1 = 0$, so $a = \lambda 1$, i.e. $\mathcal{A}$ is scalar multiples of 1.

Proof of (11): From the proof of (9), we see that $\ker \alpha$ is a maximal ideal. Now, let I be a maximal ideal. I is closed from (6), so $\mathcal{B}/I$ is a field from (2) and a complex Banach algebra from (7). From (10), this means $\mathcal{B}/I$ is isomorphic to $\mathbb{C}$. Let $\beta \colon \mathcal{B} \to \mathcal{B}/I \approx \mathbb{C}$ be the natural (projection) homomorphism. By definition, β is a character, and $I = \ker \beta$, so any maximal ideal is the kernel of some character. This shows surjectivity. For injectivity, if $\ker \alpha = \ker \beta = I$, then $1 \notin I$, and from the proof of (9), we may write any $a \in \mathcal{B}$ as $a = c + \lambda 1$, $c \in I, \lambda \in \mathbb{C}$. Then $\alpha(a) = \lambda = \beta(a)$. □

Theorem 10.24. ***(Characterization of Characters)*** *Let $\mathcal{B}$ be a commutative Banach algebra with identity. Then the set of characters $\widetilde{\mathcal{B}}$ is a closed subset of the unit ball of $\mathcal{B}^*$ in the weak* topology. In particular, $\widetilde{\mathcal{B}}$ is a compact Hausdorff space in this topology.*

Proof. Let $\alpha \in \widetilde{\mathcal{B}}$. We first show that $\|\alpha\| \le 1$. Let $a \in \mathcal{B}$, $\|a\| \le 1$. Then $\|a^n\| \le 1$, so $\{a^n\}$ is a bounded set. Since $|\alpha(a^n)| = |\alpha(a)|^n$ we must have $|\alpha(a)| \le 1$, via Thm. 10.23(9). Therefore, $\widetilde{\mathcal{B}}$ is contained in the unit ball of $\mathcal{B}^*$. Now, using the same reasoning as in Alaoglu's Theorem (Thm. 10.12), $\{\xi \in \mathcal{B}^* \colon \xi(ab) = \xi(a)\xi(b)\}$ (for $a, b \in \mathcal{B}$ fixed) is closed in the weak* topology. Therefore, $\cap_{a,b\in\mathcal{B}}\{\xi \in \mathcal{B}^* \colon \xi(ab) = \xi(a)\xi(b)\}$ is weak* closed. Simiarly, $\{\xi \in \mathcal{B}^* \colon \xi(1) = 1\}$ is weak* closed, so $\widetilde{\mathcal{B}}$ is weak* closed (recall that we defined characters to be nonzero). □

Theorem 10.25. ***(Gelfand)*** *The canonical map is a homomorphism from $\mathcal{B}$ into $C(\widetilde{\mathcal{B}})$ with norm at most one. (Recall the definition: $a \mapsto \hat{a}$, where $\hat{a}(\alpha) := \alpha(a)$.)*

Proof. $\widehat{ab}(\alpha) = \alpha(ab) = \alpha(a)\alpha(b) = \widehat{a}(\alpha)\widehat{b}(\beta)$. So, the Gelfand map is a homomorphism. For all $\alpha \in \widetilde{\mathcal{B}}$ we have $|\widehat{a}(\alpha)| = |\alpha(a)| \le \|a\|$ from Thm. 10.24, so $\|\hat{a}\|_\infty \le \|a\|$. □

Theorem 10.26. ***(Facts about the Gelfand map, etc.)***

(1) *The kernel of the canonical map is the radical of $\mathcal{B}$ (i.e. the intersection of maximal ideals). So, the canonical map is injective if and only if the radical is trivial. (Surjectivity will be dealt with below.)*

(2) $\widehat{1}(\alpha) = \alpha(1) = 1$ *for all $\alpha \in \widetilde{\mathcal{B}}$ (Use Thm. 10.23*(4)*)*

(3) *$\lambda \in \sigma(a)$ if and only if $\lambda \in$ range of $\widehat{a}$, i.e. $\sigma(a) = \mathcal{R}(\widehat{a})$*

(4) *The Spectral mapping theorem (Thm. 10.22) follows from* (3).

(5) *$r(a) = \|\widehat{a}\|_\infty \le \|a\|$, from (3) and Thm. 10.25. Therefore, $r(a+b) \le r(a) + r(b)$ and $r(ab) \le r(a)r(b)$.*

(6) *The following are equivalent: $a \in$ radical, $\widehat{a} = 0$, $\|\widehat{a}\|_\infty = 0$, and $r(a) = 0$. (Apply* (1) *and* (5)*).*

(7) *$r(a^n) = r(a)^n$, and $r(a) = \lim_{n\to\infty} \|a^n\|^{1/n}$*

(8) *$\|\widehat{a}\|_\infty = \|a\|$ for all $a \in \mathcal{B}$ if and only if $\|a^2\| = \|a\|^2$ for all $a \in \mathcal{B}$.*

(9) *Let $\mathcal{B}$ be a symmetric Banach algebra ($1 + a^*a$ is invertible). If a is Hermitian, then a is real. If a is strongly positive, then a is positive.*

(10) *Let $\mathcal{B}$ be a commutative $*$ algebra with unit. Then the following are equivalent: (i) $\mathcal{B}$ is symmetric, (ii) Hermitian implies real, (iii) $\widehat{(a^*)}(\alpha) = \overline{\widehat{a}(\alpha)}$, and (iv) Every maximal ideal is closed under $*$.*

Proof of (1): If $\hat{a} = 0$ then $\alpha(a) = 0$ for all α. So $a \in \ker\alpha\ \forall\alpha$, so a is in every maximal ideal by Thm. 10.23(11). We can reverse this argument to see that the radical is in the kernel of the canonical map.

Proof of (3): a is invertible, if and only if a is not in a maximal ideal, if and only if $\widehat{a}(\alpha) \ne 0$ for each α (from (1)). So, by negation, $\lambda \in \sigma(a)$, if and only if $a - \lambda 1$ is in some maximal ideal, if and only if there exists some α with $\alpha(a - \lambda 1) = 0$, if and only if $\widehat{a}(\alpha) - \lambda\widehat{1}(\alpha) = 0$, if and only if $\widehat{a}(\alpha) = \lambda$, from (2).

Proof of (4): Using (3) and Thm. 10.25,

$$\sigma(p((a))) = \mathcal{R}(\widehat{p(a)}) = \mathcal{R}(p(\widehat{a})) = p(\sigma(a))$$

Proof of (7): From the remarks at the beginning of this Section, $\sigma(a)$ is compact, so $\exists$ $\lambda \in \sigma(a)$ with $|\lambda| = r(a)$. So, $\lambda^n \in \sigma(a^n)$ by (4) (Thm. 10.22), so $r(a^n) \ge |\lambda^n| = r(a)^n$. Conversely, $\exists\ \lambda_0 \in \sigma(a^n)$ with $r(a^n) = |\lambda_0|$. By (4), $\exists\ \lambda \in \sigma(a)$ with $\lambda^n = \lambda_0$, so $r(a)^n \ge |\lambda|^n = |\lambda_0| = r(a^n)$.

To prove the second assertion, we use power series. Let ξ be a linear functional. For λ small enough, $(1-\lambda a)^{-1} = \sum_{n\geq 0} a^n\lambda^n$, and $\xi((1-\lambda a)^{-1}) = \sum_{n\geq 0}\xi(a^n)\lambda^n$. From the discussion near the beginning of this Section, $\xi((1-a\lambda)^{-1})$ is analytic for $(1/\lambda)\notin\sigma(a)$, so the infinite sum converges for $|\lambda| < 1/r(a)$. For fixed ξ and λ as above, $\{|\lambda^n|\,|\xi(a^n)| : n = 0,1,2,\ldots\}$ is therefore a bounded set. So, the Uniform boundedness principle (Thm. 10.6) says $\{\lambda^n a^n\}$ is a bounded set, i.e. $\|\lambda^n a^n\| \leq K$, $K > 0$. So, $\limsup \|a^n\|^{1/n} \leq 1/|\lambda|$ for $r(a) < 1/|\lambda|$, so $\limsup \|a^n\|^{1/n} \leq r(a)$. But $r(a)^n = r(a^n) \leq \|a^n\|$ from (5), so $r(a) \leq \liminf \|a^n\|^{1/n}$.

Proof of (8): From (5), $\|\widehat{a}\|_\infty = \|a\|$ if and only if $r(a) = \|a\|$. Now, if $\|a^2\| = \|a\|^2$ for all a, then $\|a\| = \left\|a^{2^n}\right\|^{1/2^n}$, so $\|a\| = \lim_n \left\|a^{2^n}\right\|^{1/2^n} = r(a)$ from (7). If $r(a) = \|a\|$ for all a, then $\|a^2\| = r(a^2) = r(a)^2 = \|a\|^2$, from (7).

Proof of (9): Let $a = a^*$. By scaling, it suffices to show $a - i$ is invertible. But $(a-i)(a+i)(1+a^*a)^{-1} = 1$ and $(1+a^*a)(a+i)(a-i) = 1$, so $a-i$ is invertible. For the second assertion, let $a = b^*b$. Then $a = a^*$, so by the first assertion, a is real. Finally, for $\alpha < 0$ we write $b^*b - \alpha = -\alpha\left(\left(\frac{b}{\sqrt{-\alpha}}\right)^*\left(\frac{b}{\sqrt{-\alpha}}\right)+1\right)$, which is invertible, so $\sigma(a)\subseteq[0,\infty)$.

Proof of (10): (i) implies (ii) follows from (9). To show (ii) implies (iii), let $f = a + a^*$, $g = i(a - a^*)$. Then f, g are Hermitian, so $\sigma(f),\sigma(g)\subseteq\mathbb{R}$. Using the Gelfand map and (3), $\alpha(f),\alpha(g)\in\mathbb{R}$ for a character α, i.e. $\alpha(f) = \overline{\alpha(f)}, \alpha(g) = \overline{\alpha(g)}$. So $\alpha(a)+\alpha(a^*) = \overline{\alpha(a)} + \overline{\alpha(a^*)}$ and $i(\alpha(a)-\alpha(a^*)) = -i(\overline{\alpha(a)} - \overline{\alpha(a^*)})$, i.e. $-\alpha(a)+\alpha(a^*) = \overline{\alpha(a)} - \overline{\alpha(a^*)}$. Adding the first and third equality, $\overline{\alpha(a)} = \alpha(a^*)$.

To show (iii) implies (iv), let I be a maximal ideal. By Thm. 10.23(11), $I = \ker\alpha$ for some character α. For $a\in I$, $0 = \alpha(a) = \overline{\alpha(a)} = \alpha(a^*)$, so $a^*\in I$. To show (iv) implies (i), first observe that (iv) implies (iii). For α a character, let $b = c - \alpha(c)1$, so $\alpha(b) = 0$. Then $b, b^*\in\ker\alpha$, i.e. $\alpha(b^*) = \alpha(c^*) - \alpha((\alpha(c)1)^*) = \alpha(c^*) - \overline{\alpha(c)} = 0$, proving (iii). Now, $\alpha(c^*c) = \alpha(c^*)\alpha(c) = |\alpha(c)|^2$, so $\alpha(1+c^*c) = 1+|\alpha(c)|^2 \neq 0$, so $1+c^*c$ is not in any maximal ideal by Thm. 10.23(11), i.e. it is invertible. □

Theorem 10.27. ***(Gelfand Map for a*** B^* ***algebra)*** *If $\mathcal{B}$ is a commutative B^* algebra with identity, then the canonical map (Gelfand map) is an isometric isomorphism onto $C(\widetilde{\mathcal{B}})$. So, using Thm. 10.24, $\mathcal{B}$ is isometrically isomorphic to the algebra of complex valued functions on a compact Hausdorff space.*

Proof. Claim 1: If $\mathcal{B}$ is a commutative $*$-multiplicative Banach algebra with identity, then $\|a\| = r(a)$ for all $a\in\mathcal{B}$. (From the remarks at the beginning of this Section, a B^* algebra is $*$-multiplicative. By Thm. 10.26(3), $r(a) = \|a\|$ if and only if $\|\widehat{a}\| = \|a\|$, so the Gelfand map is an isometry with complete image.)

Claim 2: A commutative B^* algebra with identity is symmetric and semisimple. (In particular, via Thm. 10.26(1), the Gelfand map is injective).

Claim 3: If $\mathcal{B}$ is commutative, and symmetric (with unit), the image of $\mathcal{B}$ under the canonical map is dense in $C(\widetilde{\mathcal{B}})$. (Combining this with the first claim gives surjectivity, proving the theorem).

Proof of Claim 1: Let b be Hermitian. Then $\|b^2\| = \|b^*b\| = \|b^*\|\,\|b\| = \|b\|^2$, $\left\|b^{2^n}\right\| = \|b\|^{2^n}$, so $r(b) = \|b\|$ from Thm. 10.26(7). For a arbitrary, a^*a is Hermitian, so $\|a^*\|\,\|a\| = \|a^*a\| = r(a^*a) \leq r(a^*)r(a) \leq \|a^*\|\,r(a)$, using Thm. 10.26(5) twice. So, $\|a\| \leq r(a)$, and Thm. 10.26(5) says $r(a) \leq \|a\|$, so $\|a\| = r(a)$.

Proof of Claim 2: Semisimplicity (radical $= 0$) follows from Thm. 10.26(6) and Claim 1. Now, using Thm. 10.26(10), it suffices to show: if $a = a^*$, then a is real. By scaling, if suffices to show $a - i$ is invertible, i.e. $1 + ia$ is invertible, i.e. $1 \notin \sigma(-ia)$, i.e. $\lambda + 1 \notin \sigma(\lambda - ia)$ (for some $\lambda \in \mathbb{R}$). If $\lambda + 1 \in \sigma(\lambda - ia)$ for all $\lambda \in \mathbb{R}$, then since $\sup_{\gamma \in \sigma(b)} |\gamma| =: r(b) \leq \|b\|$ from Thm. 10.26(5), we have

$$(\lambda + 1)^2 \leq \|\lambda - ia\|^2 = \|(\lambda + ia)(\lambda - ia)\| = \left\|\lambda^2 + a^2\right\| \leq \lambda^2 + \left\|a^2\right\|$$

using $*$-multiplicativity and the triangle inequality. But then $2\lambda + 1 \leq \|a^2\|$, which is a contradiction for λ large. Therefore, $\lambda + 1 \notin \sigma(\lambda - ia)$ for some $\lambda \in \mathbb{R}$.

Proof of Claim 3: Let $\alpha_1 \neq \alpha_2 \in \widetilde{\mathcal{B}}$, $\beta_1, \beta_2 \in \mathbb{C}$. Let $a \in \mathcal{B}$ such that $\alpha_1(a) \neq \alpha_2(a)$. Choose (using basic linear algebra) λ, μ such that $\lambda\alpha_1(a) + \mu = \beta_1$, $\lambda\alpha_2(a) + \mu = \beta_2$. Let $b := \lambda a + \mu$. Then $\widehat{b}(\alpha_1) = \beta_1$, $\widehat{b}(\alpha_2) = \beta_2$. The claim then follows by Stone-Weierstrass Theorem. Note that $\widetilde{\mathcal{B}}$ is a compact Hausdorff space from Thm. 10.24, the image of the Gelfand map clearly contains constants, and Thm. 10.26(10)(iii) shows that this image is closed under conjugation. □

Theorem 10.28. ***(Facts about maximal abelian self adjoint (m.a.s.a.) algebras, etc.)*** *(Recall $\mathcal{A}'$ is a set of commutators, and $\mathcal{M}$ denotes the multiplication algebra on L_2.)*

(1) *$\mathcal{A} \subseteq B(H)$ is maximal abelian if and only if $\mathcal{A} = \mathcal{A}'$. If $\mathcal{A} \subseteq B(H)$ is m.a.s.a. then $\mathcal{A} = \mathcal{A}'$.*
(2) *A m.a.s.a algebra $\mathcal{A}$ is a C^* algebra (and a B^* algebra).*
(3) *Let (X, μ) be a σ-finite measure space. Then $\mathcal{M}(X, \mu)$ is a m.a.s.a. algebra.*
(4) *Let $\mathcal{A}$ be any $*$ subalgebra of $B(H)$. Suppose K is a closed subspace of H and P is the projection on K. Then K is invariant under $\mathcal{A}$ if and only if $P \in \mathcal{A}'$*
(5) *If H is separable and $\mathcal{A}$ is m.a.s.a. on H then $\mathcal{A}$ has a cyclic vector. ($\exists$ z with $\mathcal{A}z$ dense in H.)*

Proof of (1): Let $\mathcal{A}$ be maximal abelian and let $B \in \mathcal{A}'$. We first note that $id, 0 \in \mathcal{A}$, since the set of operators of the form A or $A + id$, with $A \in \mathcal{A}$ or $A = 0$, contains $\mathcal{A}$ and is abelian. Also, the set of operators of the form $A_0 + A_1 B + \cdots + A_n B^n$ is an abelian algebra containing $\mathcal{A}$, so $B \in \mathcal{A}$, i.e. $\mathcal{A}' \subseteq \mathcal{A}$, so $\mathcal{A} = \mathcal{A}'$. Conversely, if $C \supseteq A$ is an abelian algebra, then $C \subseteq \mathcal{A}' = \mathcal{A}$, so $C = \mathcal{A}$, i.e. $\mathcal{A}$ is maximal abelian.

Let $\mathcal{A}$ be m.a.s.a., and let $B \in \mathcal{A}'$. We first note that $id, 0 \in \mathcal{A}$, since the set of operators of the form A or $A + id$, with $A \in \mathcal{A}$ or $A = 0$, contains $\mathcal{A}$ and is abelian and self adjoint. Now, for $A \in \mathcal{A}, A^* \in \mathcal{A}$, so $AB = BA, A^*B = BA^*$, so $B^*A^* = A^*B^*, B^*A = AB^*$, so $B^* \in \mathcal{A}'$. Write $A = (A + A^*)/2 + i(-i(A - A^*)/2) =: X + iY$. Then $X, Y \in \mathcal{A}', X^* = X, Y^* = Y$. The set of operators of the form $A_0 + A_1 X + \cdots + A_n X^n$, $A_j \in \mathcal{A}$ is a commutative self adjoint algebra containing $\mathcal{A}$, so it is $\mathcal{A}$. This set of operators also contains X, since $1 \in \mathcal{A}$. Therefore, $X \in \mathcal{A}$. Similarly, $Y \in \mathcal{A}$, so $B \in \mathcal{A}$, i.e. $\mathcal{A}' \subseteq \mathcal{A}$. By commutativity of $\mathcal{A}$, $\mathcal{A} \subseteq \mathcal{A}'$. Therefore $\mathcal{A} = \mathcal{A}'$.

Proof of (2): Let $A_n \in \mathcal{A}$, $A_n \to A$. For any $B \in \mathcal{A}$ we have $AB - BA = \lim(A_n B - BA_n) = 0$, so $A \in \mathcal{A}' = \mathcal{A}$, from (1). Also, by definition of m.a.s.a., $A \in \mathcal{A}$ implies $A^* \in \mathcal{A}$. Finally, $\|A^*A\| = \|A\|^2$, from the remarks at the beginning of this Section.

Proof of (3): Assume $\mu(X) < \infty$. Let $T \in (\mathcal{M}(X, \mu))'$. Let $g = T(1)$. If $f \in L_\infty$ then $TM_f 1 = M_f T1$, so $T(f) = fg$, i.e. $Tf = M_g f$ for $f \in L_\infty$. By approximating appropriate level sets with indicator functions (as in the proof that $\|M_f\| = \|f\|_\infty$), we see that $\|g\|_\infty \le \|T\|$. Since M_g and T are both bounded on L_2, the equation $T|_{L_\infty} = M_g|_{L_\infty}$ extends by continuity to L_2. Thus, $T \in \mathcal{M}(X,\mu)$ and $\mathcal{M}(X,\mu)$ is maximal abelian from (1). Since $M_g^* = M_{\overline{g}}$, $\mathcal{M}(X,\mu)$ is self-adjoint. The general σ-finite case follows the usual "piecing together" argument. Write $X = \cup_{j=1}^\infty X_j$, $\mu(X_j) < \infty$, X_j disjoint. If $T \in (\mathcal{M}(X,\mu))'$, T commutes with $M_{1_{X_j}}$, so T leaves the subspace $\{f \in L_2(X)\colon f|_{X_j^c} = 0\}$ ($= L_2(X_j)$) invariant, etc.

Proof of (4): If $P \in \mathcal{A}', x \in K$ then $Ax = APx = PAx \in K$. Conversely, if $AK \subseteq K$ with $x \in H$, then $APx \in K$, so $APx = PAPx$. Since $A^* \in \mathcal{A}$, we similarly have $A^*P = PA^*P$. So, $PA = P^*A = (A^*P)^* = (PA^*P)^* = PAP = AP$, so $P \in \mathcal{A}'$. (A projection satisfies $P = P^*$ by Thm. 10.1(b)).

Proof of (5): Let $x \in H$, and let $\overline{\mathcal{A}x}$ be the smallest closed subspace containing $\mathcal{A}x$. Note that $id \in \mathcal{A}$ so $x \in \mathcal{A}x$. Since $\mathcal{A}x$ is invariant under $\mathcal{A}$, so is $\overline{\mathcal{A}x}$ (via a limiting argument). Suppose $y \perp \mathcal{A}x$. Since $(Ay, Bx) = (y, A^*Bx) = 0$, $\mathcal{A}y \perp \mathcal{A}x$. Let $E = \{x_\alpha\}$ be an orthonormal set such that $\mathcal{A}x_\alpha \perp \mathcal{A}x_\beta$ for $\alpha \neq \beta$. These sets exists (consider e.g. singletons), so Zorn's lemma gives a maximal such set, E. Note that $H = \text{closed span}_\alpha\{\mathcal{A}x_\alpha\}$ (for if not, this would contradict maximality of E). Since H is separable, $E = \{x_\alpha\}_{\alpha\in\mathbb{N}}$ is countable. Set $z = \sum_{n\ge1} 2^{-n}x_n$. This is our desired cyclic vector. To see this, let P_n project onto $\overline{\mathcal{A}x_n}$. From (4), $P_n \in \mathcal{A}'$, so $P_n \in \mathcal{A} = \mathcal{A}'$ by (1), so $\mathcal{A}z \supseteq \mathcal{A}P_nz = \mathcal{A}2^{-n}x_n = \mathcal{A}x_n$, so $\overline{\mathcal{A}z} \supseteq \text{closed span}_n\{\mathcal{A}x_n\} = H$. □

Theorem 10.29. ***(Unitary Diagonalization of m.a.s.a algebra)*** *Let $\mathcal{A}$ be a m.a.s.a. algebra on a separable Hilbert space H. Then $\exists$ a finite measure space (X,μ) and a unitary operator $U\colon H \twoheadrightarrow L^2(X,\mu)$ such that $U\mathcal{A}U^{-1} = \mathcal{M}(X,\mu)$*

Proof. Idea: Take "Rayleigh quotient" of (unit) cyclic vector (separating vector), use inverse Gelfand map. Using the Gelfand map again, define $U_0Az = \widehat{A}$.

Let z be a unit cyclic vector for $\mathcal{A}$ (using Thm. 10.28(5)). Then z is also a **separating vector** for $\mathcal{A}$ (i.e. if $A \in \mathcal{A}$ and $Az = 0$ then $A = 0$) since if $Az = 0$ then $\forall\ B \in \mathcal{A}$, $ABz = BAz = 0$, so $A\mathcal{A}z = 0$, but $\mathcal{A}z$ is dense, so $A = 0$. From Thm. 10.28(2), $\mathcal{A}$ is a B^* algebra (with identity). Let $X = \text{spectrum}(\mathcal{A}) = \widetilde{\mathcal{A}}$. From Thm. 10.27, the Gelfand map $A \mapsto \widehat{A}$ is an isometric isomorphism $\mathcal{A} \twoheadrightarrow C(X)$.

Define $\Lambda\colon C(X) \to \mathbb{C}$ by

$$\Lambda(\widehat{A}) := (Az, z)$$

Λ is a bounded linear functional, as it is the composition of bounded linear functionals. In fact, $|\Lambda(\widehat{A})| \le \|A\| = \|\widehat{A}\|_\infty$ (using that the Gelfand map is an isometry). Λ is positive since

$$\Lambda(\overline{\widehat{A}}\widehat{A}) = \Lambda(\widehat{A^*}\widehat{A}) = \Lambda(\widehat{A^*A}) = (A^*Az, z) = \|Az\|^2 \ge 0$$

(using Claim 2 of Thm. 10.27, Thm.10.26(10)(iii), and Thm. 10.25). So, by Riesz's Representation Theorem and recalling that X is compact, $\exists$ a unique regular Borel measure μ on X such that $\Lambda(\widehat{A}) = \int \widehat{A}d\mu$. Using Thm. 10.26(2), note that $\mu(X) = \int 1 d\mu = \Lambda(1) = \|z\|^2 = 1$, so μ is a probability measure. Define $U_0\colon \mathcal{A}z \to L_2(X,\mu)$ by

$$U_0Az := \widehat{A}$$

U_0 is well defined since $Az = 0$ implies $A = 0$. U_0 is also linear and densely defined by Thm. 10.28(5). Moreover,

$$\|U_0Az\|^2 = \int \overline{\widehat{A}}\widehat{A}d\mu = \Lambda(\widehat{A^*A}) = (Az, Az) = \|Az\|^2$$

So, U_0 is an isometry, which extends by continuity to $U \colon H \to L_2(X, \mu)$, with $\|Ux\| = \|x\|$ $\forall\ x \in H$. Since U is an isometry and H is a Hilbert space, the range of U is a complete (hence closed) subspace of $L_2(\mu)$ (which contains $C(X)$ by definition of the Gelfand map), so the range of U is all of of $L_2(\mu)$, by the proof of Thm. 10.10(b). Thus, U is unitary.

Now, if $A, B \in \mathcal{A}$, $UAU^{-1}\widehat{B} = UABz = \widehat{AB} = M_{\widehat{A}}B$, so $UAU^{-1} = M_{\widehat{A}}$ on a dense set $\{\widehat{B} \colon B \in \mathcal{A}\} \subseteq L_2(X, \mu)$, so $UAU^{-1} = M_{\widehat{A}}$ on all of $L_2(X, \mu)$. Let $\mathcal{N} := U\mathcal{A}U^{-1}$, and let $\mathcal{M} := \mathcal{M}(X, \mu)$. We just showed $\mathcal{N} \subseteq \mathcal{M}$. Now, if $T \in \mathcal{M}$, then $T \in \mathcal{N}'$, so $U^{-1}TU \in \mathcal{A}'$. But $\mathcal{A}' = \mathcal{A}$ (Thm. 10.28(1)), so $U^{-1}TU \in \mathcal{A}$, i.e. $T \in \mathcal{N}$, therefore $\mathcal{M} = \mathcal{N}$. □

Theorem 10.30. ***(Spectral theorem, Multiplication Operator Form)*** *Let $\{A_\alpha\}$ be a family of bounded normal operators on a complex separable Hilbert space H. Assume the family is commuting: $A_\alpha A_\beta = A_\beta A_\alpha$ $\forall \alpha, \beta$, and $A_\alpha A_\beta^* = A_\beta^* A_\alpha$ $\forall \alpha, \beta$. Then $\exists$ a finite measure space (X, μ) and a unitary operator $U \colon H \to L_2(X, \mu)$ and for each α $\exists$ a function $f_\alpha \in L_\infty$ such that $UA_\alpha U^{-1} = M_{f_\alpha}$*

Proof. Let $\mathcal{A}_0$ be the algebra generated by the $\{A_\alpha, A_\alpha^*\}$. Note that $\mathcal{A}_0$ is a commutative $*$ algebra. Using Zorn's lemma, $\exists$ a largest such $*$ algebra $\mathcal{A}$. We then claim that $\mathcal{A} = \mathcal{A}'$. If $B \in \mathcal{A}'$, then $B^* \in \mathcal{A}'$, so $C := B + B^* \in \mathcal{A}'$. Then the algebra generated by $\mathcal{A}$ and C is commutative and self-adjoint, so $C \in \mathcal{A}$. Similarly, $i(B - B^*) \in \mathcal{A}$, so $B \in \mathcal{A}$, so $\mathcal{A}' = \mathcal{A}$, so $\mathcal{A}$ is m.a.s.a., so we may apply Thm. 10.29. □

Theorem 10.31. ***(Spectral theorem, Projection Valued Measure Form)*** *Let A be a bounded normal operator on a separable Hilbert space H. There exists a unique projection valued Borel measure E on $\mathbb{C}$ with compact support such that*

$$A = \int_{\mathbb{C}} zdE$$

Moreover, if D is any bounded operator on H, then D commutes with A and A^* if and only if D commutes with $E(B)$ for all Borel sets B.

Proof idea: From the previous theorem, $UAU^{-1} = M_f$. Now, $G(B) := M_{1_{f^{-1}(B)}}$ is a projection valued measure. Define $E(B) := U^{-1}G(B)U$.

Theorem 10.32. ***(Spectral theorem for a bounded Hermitian operator)*** *If A is a bounded Hermitian operator on a separable Hilbert space H, then $\exists$ a unique projection-valued Borel measure $E(\cdot)$ on the line with compact support such that*

$$A = \int_{-\infty}^{\infty} \lambda dE(\lambda)$$

And for all real Borel sets B, $E(B) \subseteq \sigma(A)$

Theorem 10.33. ***(Spectral theorem for a unitary operator)*** *Let U be a unitary operator on a separable Hilbert space. Then $\exists$ a unique projection-valued Borel measure $E(\cdot)$*

on $[0, 2\pi)$ such that

$$U = \int_0^{2\pi} e^{i\theta} dE(\theta)$$

And $E(B) \subseteq \sigma(U)$ for all Borel sets B.

Theorem 10.34. ***(von Neumann's criteria for self adjointness)*** *Let T be a symmetric operator on a Hilbert space H. Then the following three statements are equivalent:*

(a) *T is self-adjoint*
(b) *T is closed and* $\ker(T^* + i) = \ker(T^* - i) = \{0\}$
(c) *Range($T + i$)=Range($T - i$)= H*

Theorem 10.35. ***(Spectral Theorem, Unbounded case, Multiplication operator form)*** *Let T be a self-adjoint operator on a separable Hilbert space H. Then $\exists$ a finite measure space (X, μ), a unitary operator $U\colon H \to L_2(X, \mu)$ and a real valued measurable function f on X such that*

$$UTU^{-1} = M_f$$

Proof. (Sketch) From Thm. 10.34, $T + i$ is bijective from $\mathcal{D}_T$ to H, so its inverse $(T + i)^{-1}$ exists. One can also check that $\|(T + i)\phi\|^2 = ((T + i)\phi, (T + i)\phi) = \|T\phi\|^2 + \|\phi\|^2 \geq \|\phi\|^2$, so $(T + i)^{-1}$ is bounded, and we can apply the bounded spectral theorem and write $U(T + i)^{-1}U = M_g$. Since $(T + i)^{-1}$ is bijective, so is M_g. In particular, g can be zero only on a set of measure zero. Therefore $f := (1/g) - i$ is well-defined. This is our desired f. □

Theorem 10.36. ***(Spectral Theorem, Unbounded case, Projection valued measure form)*** *Let T be a self-adjoint operator on a separable complex Hilbert space H. Then $\exists$ a projection valued measure $E(\cdot)$ on the Borel sets of the line such that*

$$T = \int_{-\infty}^{\infty} \lambda dE(\lambda)$$

Theorem 10.37. ***(Facts about Compact Operators, etc.)*** *Let $A\colon H \to K$ be a linear map between two Banach spaces.*

(1) *If A is compact then A is bounded.*
(2) *Let A be compact and let $B(0, n) = \{x\colon \|x\| \leq n\}$. Then $\overline{A(B(0, n))}$ is compact and therefore separable. Since the range of A is $\cup_{n\geq 1}\overline{A(B(0, n))}$, the range of A is separable.*
(3) *Let $\{A_n\}$ be a sequence of compact operators such that $\|A_n - A\| \to 0$. Then A is compact.*
(4) *Define $\alpha_n(A) := \inf\{\|A - A_n\| : A_n\colon H \to K, rank(A_n) < n\}$, where $n = 1, 2, \ldots$. For $n, m \geq 1$ and for all operators A, B:*

$$\alpha_{n+m-1}(A + B) \leq \alpha_n(A) + \alpha_m(B)$$
$$\alpha_{n+m-1}(A \circ B) \leq \alpha_n(A) \cdot \alpha_m(B)$$

(5) *For $0 < p < \infty$ define $A^p(X, Y)$ as the set of all operators $A\colon X \to Y$ with $a_p(A) := (\sum_{n=1}^{\infty} \alpha_n(A)^p)^{\frac{1}{p}} < \infty$. The quantity $a_p(A)$ is a quasi-norm that makes $A^p(X, Y)$ a quasi-Banach operator ideal.*
(6) *If $A \in A^p(X, Y)$ and $B \in A^q(Y, Z)$ then $BA \in A^s(X, Z)$ with $1/s = 1/p + 1/q$.*

(7) *If $A\colon X \to Y$ is compact then $\alpha_n(A) = \alpha_n(A^*)$. In particular, A^* is also compact, by (3).*

Proof of (3): Let $\{x_n\}$ with $\|x_n\| \leq 1$. A diagonalization argument gives a subsequence $\{y_n\} \subseteq \{x_n\}$ such that for all j, $A_j y_n$ converges as $n \to \infty$. Then

$$\begin{aligned} \|Ay_k - Ay_\ell\| &\leq \|Ay_k - A_i y_k\| + \|A_i y_k - A_i y_\ell\| + \|A_i y_\ell + Ay_\ell\| \\ &\leq 2\|A - A_i\| + \|A_i y_k - A_i y_\ell\| \end{aligned}$$

So $\limsup_{k,\ell\to\infty} \|Ay_k - Ay_\ell\| \leq 2\|A - A_i\|$ which can be made arbitrarily small. Therefore, $A(B(0,1))$ is a compact set, as desired.

Proof of (4): Given A_n, B_m of rank less than n, m respectively, consider $(A+B)-(A_n+B_m)$. The first inequality follows. For the second inequality, note that

$$\|AB - (A_nB + AB_m - A_nB_m)\| = \|(A - A_n)(B - B_n)\| \leq \|A - A_n\| \, \|B - B_m\|$$

So, using that $\operatorname{rank}(A_nB + AB_m - A_nB_m) \leq \operatorname{rank}(A_n(B - B_m)) + \operatorname{rank}(AB_m) < n + m - 1$, we get

$$\alpha_{n+m-1}(AB) \leq \inf_{A_n, B_m} \|A - A_n\| \, \|B - B_m\| = \alpha_n(A)\alpha_n(B)$$

Proof of (5): Using that the α_n's are decreasing in n, (4), and the L_p (quasi)-triangle inequality,

$$\begin{aligned} a_p(A+B) = \left(\sum_{n=1}^{\infty} \alpha_n(A+B)^p\right)^{\frac{1}{p}} &\leq \left(2\sum_{n=1}^{\infty} \alpha_{2n-1}(A+B)^p\right)^{\frac{1}{p}} \\ &\leq 2^{\frac{1}{p}} \left(\sum_{n=1}^{\infty} (\alpha_n(A) + \alpha_n(B))^p\right)^{\frac{1}{p}} \leq C_p(a_p(A) + a_p(B)) \end{aligned}$$

Since also $a_p(\lambda A) = |\lambda| \, a_p(A)$, we see that $a_p(A)$ is a quasi-norm and $A^p(X,Y)$ is a linear subspace of $L(X,Y)$. Since $\alpha_n(A) = 0$ if $\operatorname{rank}(A) < n$, $A^p(X,Y)$ contains finite rank operators.

Since $\alpha_1(A) = \|A\|$ and the sequence $\{\alpha_n(A)\}$ is decreasing in n, (4) shows $\alpha_n(CBA) \leq \alpha_n(C)\alpha_n(B)\alpha_n(A) \leq \|C\| \, \alpha_n(B) \, \|A\|$, so $a_p(CBA) \leq \|C\| \, a_p(B) \, \|A\|$. In particular, if $B \in A^p(X,Y)$, $A \in A^p(Z,X)$ and $C \in A^p(Y,V)$, then $CBA \in A^p(Z,V)$. For any rank-one operator $A\colon X \to Y$ of the form $A(x) = x^*(x) \cdot y$, $a_p(A) = \alpha_1(A) = \|A\| = \|x^*\| \, \|y\|$.

It remains to show that $(A^p(X,Y), a_p(\cdot))$ is complete. This procedure is mostly routine. Starting with a Cauchy sequence, take a rapidly convergent subsequence, use $\alpha_1(\cdot) = \|\cdot\|$ to get a limit, use the (quasi)-triangle inequality for ℓ_p, and so on.

Proof of (6): From (4) and Hölder's inequality,

$$\begin{aligned} a_s(BA) = \left(\sum_{n=1}^{\infty} \alpha_n(BA)^s\right)^{\frac{1}{s}} &\leq 2^{\frac{1}{s}} \left(\sum_{n=1}^{\infty} \alpha_{2n-1}(BA)^s\right)^{\frac{1}{s}} \\ &\leq 2^{\frac{1}{s}} \left(\sum_{n=1}^{\infty} \alpha_n(A)^s \cdot \alpha_n(B)^s\right)^{\frac{1}{s}} \leq 2^{\frac{1}{s}} a_p(A) a_q(B) \end{aligned}$$

Proof of (7): If A_n has $\operatorname{rank}(A_n) < n$ with $\|A - A_n\| < \alpha_n(A) + \varepsilon$ then A_n^* has $\operatorname{rank}(A_n^*) < n$ (if S is a finite independent spanning set for $A_n(X)$, then $A_n^*(y^*)$ is determined by evaluating y^* on S, so apply Thm. 10.17(c)). So, $\alpha_n(A^*) \le \|A^* - A_n^*\| = \|A - A_n\| < \alpha_n(A) + \varepsilon$, so $\alpha_n(A^*) \le \alpha_n(A)$. It remains to show that $\alpha_n(A^*) \ge \alpha_n(A)$. It suffices to show $\alpha_n(A^{**}) \ge \alpha_n(A)$, since combining this with our first inequality gives $\alpha_n(A^*) \ge \alpha_n(A^{**}) \ge \alpha_n(A)$.

Fix n, $\varepsilon > 0$, and V a finite ε-net in $T(B_X)$. Since T is compact, Goldstine, Thm. 10.18, and Hahn-Banach, Thm. 10.4(d) show that $\iota(T(B_X))$ is norm-dense in $T^{**}(B_{X^{**}})$, so $\iota(V)$ is also an ε-net in $T^{**}(B_{X^{**}})$. Let $T_n \colon X^{**} \to Y^{**}$ with $\operatorname{rank}(T_n) < n$ and $\|T^{**} - T_n\| \le \alpha_n(T^{**}) + \varepsilon$. Let $E = \operatorname{span}\{\iota(V) \cup T_n(X^{**})\}$. From the principle of local reflexivity, Thm. 10.21, there exists $\phi \colon E \to Y$ with $\|\phi\| \le 1 + \varepsilon$ and $\phi|_{E \cap \iota(Y)} = id$, so $\phi|_{\iota(V)} = id$. For $x \in X$ with $\|x\| \le 1$ fix $v \in V$ with $\|Tx - v\| \le \varepsilon$. Then

$$\begin{aligned}
\|Tx - \phi T_n \iota x\| &\le \varepsilon + \|v - \phi T_n \iota x\| \le \varepsilon + (1+\varepsilon)\, \|\iota(v) - T_n \iota x\| \\
&\le \varepsilon + (1+\varepsilon)(\|\iota(v) - \iota(Tx)\| + \|\iota(Tx) - T_n \iota x\|) \\
&\le \varepsilon + (1+\varepsilon)(\varepsilon + \|T^{**}\iota x - T_n \iota x\|) \\
&\le \varepsilon + (1+\varepsilon)(\varepsilon + \varepsilon + \alpha_n(T^{**}))
\end{aligned}$$

using that ι is a linear isometry, that $\iota(Tx) = T^{**}\iota(x)$, and the definition of T_n. Since $\phi T_n \iota$ has rank less than n and ε is arbitrary, we conclude that $\alpha_n(T) \le \alpha_n(T^{**})$. □

Theorem 10.38. ***(Riesz Theory for Compact Operators)*** *Let $C \colon H \to H$ be a compact operator on a Banach space H. Let $B = id - C$ and $B^* = id - C^*$.*

1. *$\ker(B)$ is finite dimensional.*
2. *If B is injective then B is surjective.*
3. *$dim \ker(B) = dim \ker(B^*)$.*
4. *Every nonzero point λ of the spectrum of C is an eigenvalue of finite multiplicity. (That is $dim \ker(\lambda - C)$ is finite.) Moreover, the multiplicity of λ for C is the same as for C^*. The only possible cluster point of the spectrum of C is zero.*
5. *If C has an infinite number of eigenvalues then 0 is a cluster point of eigenvalues. So, the eigenvalues can be arranged in a sequence converging to zero. (Just apply (4).)*
6. *If C is a compact normal operator on a separable complex Hilbert space H, then there is a finite or infinite sequence P_n of mutually orthogonal finite dimensional projections such that*

$$C = \sum_{n=1}^{\infty} \lambda_n P_n \qquad (\text{or } C = \sum_{n=1}^{k} \lambda_n P_n)$$

where $\{\lambda_n\}$ are the nonzero eigenvalues of C and the series converges in the operator norm. Moreover, H has an orthonormal basis consisting of eigenvectors of C.

Proof of (1): We begin with a few claims.

Claim 1: A finite dimensional subspace $F \subseteq H$ is closed in H.

Claim 2: Let $H_0 \subseteq H$ be a closed proper subspace. For any $\varepsilon > 0$, $\exists\, x_0 \in H$ with $\|x_0\| = 1$ and $\|x - x_0\| \ge 1 - \varepsilon$ for all $x \in H_0$.

Claim 3: Any locally compact Banach space is finite dimensional.

To prove Claim 1, recall that any two norms on a finite dimensional linear space are equivalent, so the norm on F is equivalent to any Euclidean norm on F. Therefore F is complete in its own norm. Let $\{x_n\}_{n=1}^{\infty} \subseteq F$ be such that $x_n \to x \in H$. Then $\{x_n\}$ is Cauchy in F, but F is complete, so $x_n \to y \in F$. By uniqueness of limits in H, $x = y$. So $x \in F$ and F is closed in H.

To prove Claim 2, let $\varepsilon < 1$, $z_0 \notin H_0$, $d = \inf_{x \in H_0} \|x - z_0\|$. For any $\delta > 0$, $\exists\ z \in H_0$ with $\|z - z_0\| \leq d + \delta$. Let $\delta := \varepsilon d/(1-\varepsilon)$, so that z is determined from the previous sentence, and let $x_0 := (z - z_0)/\|z - z_0\|$. Then $\|x_0\| = 1$ and if $x \in H_0$, then $(\|z - z_0\|)x - z \in H_0$, so

$$\|x - x_0\| = \frac{\|(\|z - z_0\|)x - z + z_0\|}{\|z - z_0\|} \geq \frac{d}{\|z - z_0\|} \geq \frac{d}{d + \delta} = 1 - \varepsilon$$

To prove Claim 3, we instead prove its contrapositive. Assume H is infinite dimensional. We construct a sequence $x_1, x_2, \ldots \subseteq H$ with $\|x_n\| = 1$, $\|x_i - x_j\| \geq 1/2$, $i \neq j$. Let x_1 with $\|x_1\| = 1$. Given $x_1, \ldots, x_n$, let $H_0 = \operatorname{span}\{x_1, \ldots, x_n\}$. By Claim 1, H_0 is closed. By Claim 2, $\exists\ x_{n+1}$ such that $\|x_i - x_{n+1}\| \geq 1/2$ for $i = 1, \ldots, n$. The sequence $\{x_n\}$ has no Cauchy subsequence, so the closed unit ball of H is not compact. By scaling, the closed ball of radius $r > 0$ is also not compact.

We can now finally prove (1). Let $H_0 = \ker(B)$. Then $x \in H_0$ if and only if $Cx = x$, i.e. $C|_{H_0} = id$. But C maps the unit ball of H_0 into a totally bounded set, i.e. the unit ball of H_0 is locally compact. Therefore H_0 is finite dimensional by Claim 3.

Proof of (2): We first claim that $B = id - C$ satisfies $\|Bx\| \geq m\|x\|$ for all $x \in H$. If not, then there exists a sequence $\{y_n\}$ with $\|y_n\| = 1$ and $By_n \to 0$. Since C is compact, we may take a subsequence $\{y_{n_j}\}$ so that Cy_{n_j} converges. Since $By_{n_j} = (id - C)y_{n_j}$ converges also, $(B + C)y_{n_j} = y_{n_j}$ converges. Let $y = \lim y_{n_j}$. Since $\|y_n\| = 1$, $\|y\| = 1$, but $By = \lim By_n = 0$ so B is not injective. Since we have arrived at a contradiction, the claim follows.

Let $M \subseteq H$ be a closed subspace. Then $B(M)$ is closed. To see this, suppose $Bx_n \to y$. Then $\{x_n\}$ is bounded since $\|Bx_n\| \geq m\|x_n\|$. We may take a subsequence so that Cx_{n_j} converges, so $(B + C)x_{n_j} = x_{n_j}$ converges. Let $x = \lim x_{n_j}$, so that $Bx = \lim Bx_{n_j} = y$, i.e. $y \in B(M)$ as desired.

Suppose for the sake of contradiction that $\operatorname{Range}(B) \neq H$. Let $H_0 = H, H_1 = BH_0, H_2 = BH_1$, and so on. Then H_{m+1} is a closed and proper subspace of H_m. (If $E, F \subseteq H$ are closed with $F = BH, E = BF = B^2H$ and $f \in F$ then $Bf =: e \notin F$ implies $B^2 f = Be$ with $B^2 f \in E$ and $Be \notin E$ by injectivity, a contradiction. So $E \subseteq F$.) By Claim 2 of (1), $\exists$ $x_n \in H_n$ with $\|x_n\| = 1$ and $d(x_n, H_{n+1}) \geq 1/2$. If $n > m$,

$$Cx_m - Cx_n = x_m - Bx_m - x_n + Bx_n =: x_m - x, \text{ with } x \in H_{m+1}$$

So $\|Cx_m - Cx_n\| \geq d(x_m, H_{m+1}) \geq 1/2$. So the sequence $\{Cx_n\}$ contains no Cauchy subsequence, contradicting compactness of C. Therefore $\operatorname{Range}(B) = H$.

Proof of (3): Using (1) and Thm. 10.37(7), let $x_1, \ldots, x_n$ be a basis for $\ker(B)$ and let $\eta_1, \ldots, \eta_\nu$ be a basis for $\ker(B^*)$. By Hahn-Banach (Thm. 10.4(a)) $\exists\ \xi_j \in H^*$ with $\xi_j(x_i) = \delta_{ij}$, $i, j = 1, \ldots, n$. Let $K = \ker B^* \subseteq H^*$ and let $R\colon H \to H^{**}$ be such that $R(x) = \iota(x)|_K$. We claim that $R\colon H \to K^*$ is surjective. If not, $\exists\ u \in K$, $u \neq 0$ such that, for all $x \in H$, $u(x) = R(x)(u) = 0$. But then $u = 0$, a contradiction. Therefore $R\colon H \to K^*$ is surjective. So using Hahn-Banach again and the surjectivity of R, $\exists\ y_1, \ldots, y_\nu$ in H such that $\eta_j(y_i) = \delta_{ij}$, $i, j = 1, \ldots, \nu$.

Suppose for the sake of contradiction that $n < \nu$ and define

$$C'x := Cx + \sum_{j=1}^{n} \xi_j(x) y_j$$

Since C' is the sum of two compact operators, C' is compact. Let $B' = id - C'$. We claim that B' is injective. To see this, suppose $B'x_0 = 0$. Since $B - B' = (id - C) - (id - C') = C' - C$, we see that $Bx_0 = \sum_{j=1}^{n} \xi_j(x_0) y_j$. Using $\eta_i \in \ker B^*$, the previous sentence, and the choice of the y_j,

$$0 = B^* \eta_i(x_0) = \eta_i(Bx_0) = \sum_{j=1}^{n} \xi_j(x_0) \eta_i(y_j) = \xi_i(x_0)$$

Therefore $Bx_0 = 0$, so $x_0 = \sum_{j=1}^{n} \alpha_j x_j$ by definition of the x_j. But then $0 = \xi_i(x_0) = \alpha_i$ by definition of the ξ_i, so $x_0 = 0$. We conclude that $\ker B' = 0$.

By (2), B' is also surjective. So $\exists\, x \in H$ with $y_{n+1} = B'x$. Since $B - B' = C' - C$,

$$\begin{aligned} 1 = \eta_{n+1}(y_{n+1}) = \eta_{n+1}(B'x) &= \eta_{n+1}(Bx) - \eta_{n+1}\left(\sum_{j=1}^{n} \xi_j(x) y_j \right) \\ &= B^*(\eta_{n+1})(x) - \sum_{j=1}^{n} \xi_j(x) \eta_{n+1}(y_j) = 0 - 0 = 0 \end{aligned}$$

Since $\eta_{n+1} \in \ker B^*$ and using the definition of η_{n+1}. Since we have achieved a contradiction, we conclude that $n \geq \nu$.

Since $n \geq \nu$, $\dim \ker B \geq \dim \ker B^*$. Since C^* is compact by Thm. 10.37(7), we conclude that $\dim \ker B^* \geq \dim \ker B^{**}$. Since $B^{**} \iota x = \iota(Bx)$, $\dim \ker B^{**} \geq \dim \ker B$. Combining these inequalities concludes the proof.

Proof of (4): Suppose $\lambda \neq 0$ is in $\sigma(C)$. Then $id - \lambda^{-1} C$ is not invertible. Since $\lambda^{-1} C$ is compact, $1 - \lambda^{-1} C$ may not be invertible due to either a lack of injectivity (in which case λ is an eigenvalue, of finite multiplicity by (1)), or due to a lack of surjectivity (in which case $1 - \lambda^{-1} C$ is also not injective by (2), so the first case applies). If $1 - \lambda^{-1} C$ is bijective, then it is also invertible by the Open Mapping Theorem, Thm. 10.7 (or by the proof of (2)). By (3), $\dim \ker(\lambda - C) = \dim \ker(\lambda - C^*)$ for $\lambda \neq 0$, using $\ker(\lambda - C) = \ker(1 - \lambda^{-1} C)$.

We now prove by contradiction that zero is the only possible cluster point of $\sigma(C)$. Let $\lambda_n \in \sigma(C)$ with $\lambda_n \to \lambda \neq 0$. Without loss of generality, $\lambda_n \neq \lambda_m$ for $n \neq m$ and $\exists\, \varepsilon > 0$ with $|\lambda_n| \geq \varepsilon$ for all n. By the part of the theorem already proven, $\exists\, x_n \neq 0$ with $Cx_n = \lambda_n x_n$. We claim that $\{x_n\}$ is linearly independent. If not, let N be the smallest integer such that a linear relation holds: $x_N = \sum_{j=1}^{N-1} \alpha_j x_j$ with some $\alpha_j \neq 0$. Then $\lambda_N x_N = C x_N = \sum_{j=1}^{N-1} \alpha_j \lambda_j x_j$, i.e. $\lambda_N \sum_{j=1}^{N-1} \alpha_j x_j = \sum_{j=1}^{N-1} \alpha_j \lambda_j x_j$, i.e. $\sum_{j=1}^{N-1} \alpha_j (\lambda_N - \lambda_j) x_j = 0$. By minimality of N and the distinctness of the λ_j, $\alpha_j = 0$ for $j = 1, \ldots, n-1$, a contradiction. Therefore, $\{x_n\}$ is linearly independent.

Let $H_n := \operatorname{span}\{x_1, \ldots, x_n\}$. Then H_n is a strictly increasing sequence of subspaces. By Claim 2 of (1), $\exists\, y_n \in H_n$ with $\|y_n\| = 1$ and $\|y_n - x\| \geq 1/2$ for all $x \in H_{n-1}$. Let $y \in H_n$

so that $y = \sum_{j=1}^{n} \alpha_j x_j$ and $Cy - \lambda_n y = \sum_{j=1}^{n} \alpha_j(\lambda_j - \lambda_n)x_j \in H_{n-1}$. Let $n > m$. Then

$$\begin{aligned} \|Cy_n - Cy_m\| &= \|(Cy_n - \lambda_n y_n) + \lambda_n y_n - \lambda_m y_m + (\lambda_m y_m - Cy_m)\| \\ &=: \|\lambda_n y_n - z\|, \text{ with } z \in H_{n-1} \\ &\geq \frac{|\lambda_n|}{2} \geq \frac{\varepsilon}{2} \end{aligned}$$

using the choice of y_n and the definition of ε. Therefore, $\{Cy_m\}$ contains no Cauchy subsequence, a contradiction. We conclude that $\lambda = 0$, as desired.

Proof of (6), sketch: From the Spectral Theorem, Thm. 10.31, we may write

$$C = \int_{\sigma(C)} \lambda dE(\lambda)$$

Let $\lambda_1, \lambda_2, \ldots$ be the nonzero eigenvalues of C. Let $P_n := E(\{\lambda_n\})$. □

Theorem 10.39. *(**Facts about Semigroups of Operators, etc.**) Let $T_t\colon H \to H$ be a strongly continuous contraction semigroup. Define H_0 as the set of $f \in H$ such that $\lim_{t\downarrow 0} T_t f = f$.*

1. *H_0 is closed*
2. *T_t leaves the subspace H_0 invariant. For $f \in H_0$, $T_t f$ is strongly continuous on $t \geq 0$ ($\lim_{s\to 0} T_{t+s} f = T_t f$). For $f \in \mathcal{D}_A$, $Af \in H_0$.*
3. *$H_0 = \overline{\mathcal{D}_A}$*
4. *If $f \in \mathcal{D}_A$, then $T_t f$ is strongly differentiable, and $\frac{dT_t f}{dt} = AT_t f = T_t Af$. Also, $T_t f - f = \int_0^t T_s Af ds$*
5. *The operator A is closed. (If $f_n \in \mathcal{D}_A$, if $f_n \to f$ and if $Af_n \to g$ then $f \in \mathcal{D}_A$ and $Af = g$).*
6. *Let $g \in H_0$, $\lambda > 0$. The equation $\lambda f - Af = g$ has exactly one solution $f \in \mathcal{D}_A$. Moreover, f satisfies $f = R_\lambda(g) := \int_0^\infty e^{-\lambda t} T_t g dt$. The operator R_λ is linear and $\|R_\lambda g\| \leq \frac{1}{\lambda}\|g\|$. (So, $(\lambda - A)\colon \mathcal{D}_A \to H_0$ is bijective, and $(\lambda - A)^{-1} = R_\lambda$. R_λ is called the **resolvent** of A. Note also that $R_\lambda(H_0) = \mathcal{D}_A$. Moreover, the proof shows that $R_\lambda(H) \subseteq H_0$.)*
7. *Let U be a linear operator that extends A. Suppose $\mathcal{D}$ satisfies: (i) $\mathcal{D}_A \subseteq \mathcal{D} \subseteq H_0$, (ii) $\mathcal{D} \subseteq \mathcal{D}_U$ and $U(\mathcal{D}) \subseteq H_0$, (iii) If $Uf = f$ and $f \in \mathcal{D}$ then $f = 0$. Given such $\mathcal{D}$, we conclude that $\mathcal{D} = \mathcal{D}_A$.*

Proof of (1): Let $f_n \in H_0$ with $f_n \to f$. Then

$$\|T_t f - f\| \leq \|T_t f - T_t f_n\| + \|T_t f_n - f_n\| + \|f_n - f\| \leq 2\|f_n - f\| + \|T_t f_n - f_n\|$$

Now choose n so that $\|f_n - f\|$ is small, and then choose t so that $\|T_t f_n - f_n\|$ is small.

Proof of (2): Let $f \in H_0$. If $h \downarrow 0$, then $\|T_{t+h} f - T_t f\| = \|T_t(T_h f - f)\| \leq \|T_h f - f\| \to 0$. If $h \uparrow 0$, $\|T_{t+h} - T_t f\| = \|T_{t+h}(f - T_{-h} f)\| \leq \|f - T_{-h} f\| \to 0$. Therefore, $\lim_{h\to 0} T_{t+h} f = T_t f$. The first two assertions follow. Now, let $f \in H$. If $\lim_{t\downarrow 0} T_t f$ is not equal to f, or this limit does not exist, then Af is undefined. Therefore, $\mathcal{D}_A \subseteq H_0$. Let $f \in \mathcal{D}_A$ so that $f \in H_0$. Then the third assertion follows by the definition $A(f) := \lim_{h\downarrow 0} \frac{T_h f - f}{h}$ and from (1).

Proof of (3): As shown in (2), $\mathcal{D}_A \subseteq H_0$. It is therefore sufficient to show that any $f \in H_0$ can be approximated by elements of $\mathcal{D}_A$. If $f \in \mathcal{D}_A$, then $T_t f$ is strongly continuous for

$t \geq 0$ by (2), so $T_t f$ is strongly integrable on any finite interval in $[0, \infty)$. Let $g_a := \int_0^a T_t f dt$. Observe

$$T_h g_a = \int_0^a T_h T_t f dt = \int_0^a T_{h+t} f dt = \int_h^{h+a} T_t f dt = g_a + \int_a^{a+h} T_t f dt - \int_0^h T_t f dt$$

Therefore

$$\lim_{h \downarrow 0} \frac{T_h g_a - g_a}{h} = \lim_{h \downarrow 0} \left(\frac{1}{h} \int_a^{a+h} T_t f dt - \frac{1}{h} \int_0^h T_t f dt \right) = T_a f - f$$

So, $g_a \in \mathcal{D}_a$. Finally, by definition of g_a, $\lim_{h \downarrow 0} g_h / h = f$, as desired.

Proof of (4): For $t, h \geq 0$,

$$T_t \left(\frac{T_h f - f}{h} \right) = \frac{T_t T_h f - T_t f}{h} = \frac{T_{t+h} f - T_t f}{h} = \frac{T_h T_t f - T_t f}{h} \qquad (*)$$

If $f \in \mathcal{D}_A$, then $\lim_{h \downarrow 0} \frac{T_h f - f}{h} = Af$, so $\lim_{h \downarrow 0} T_t \left(\frac{T_h f - f}{h} \right) = T_t A f$. So, using $(*)$, we conclude that $\lim_{h \downarrow 0} \frac{T_{t+h} f - T_t f}{h} =: \frac{d^+ T_t f}{dt}$, with $T_t f \in \mathcal{D}_A$, and with $\frac{d^+ T_t f}{dt} = A T_t f = T_t A f$.

We now need to show that $\lim_{h \downarrow 0} \frac{T_t f - T_{t-h} f}{h} = \frac{d^+ T_t f}{dt}$. For $t > h > 0$, we have

$$\begin{aligned} \left\| \frac{T_t f - T_{t-h} f}{h} - T_t A f \right\| &\leq \left\| T_{t-h} \left(\frac{T_h f - f}{h} - Af \right) \right\| + \| T_{t-h} (Af - T_h A f) \| \\ &\leq \left\| \frac{T_h f - f}{h} - Af \right\| + \| Af - T_h A f \| \end{aligned}$$

By (2), $Af \in H_0$, so, $\lim_{h \downarrow 0} \frac{T_t f - T_{t-h} f}{h} = T_t A f = \frac{d^+ T_t f}{dt}$, as desired. Finally, $\int_0^t T_s A f ds = \int_0^t \frac{d T_s f}{ds} ds = T_t f - f$, as desired.

Proof of (5): From (4), $T_t f_n - f_n = \int_0^t T_s A f_n ds$. Letting $n \to \infty$, (noting that $\| \int_0^t (T_s A f_n - T_s g) \| \leq \int_0^t \| T_s \| \, \| A f_n - g \|$), we get $T_t f - f = \int_0^t T_s g ds$. So $\lim_{t \to 0} \frac{T_t f - f}{t} = \lim_{t \downarrow 0} \frac{1}{t} \int_0^t T_s g ds = g$. So $f \in \mathcal{D}_A$ and $Af = g$, as desired.

Proof of (6): First note that $\int_0^\infty e^{-\lambda t} T_t g$ is well-defined since the integrand is strongly continuous by (2), with norm bounded by $e^{-\lambda t} \| g \|$. Linearity of R_λ follows by its definition. Note that $\left\| \int_0^\infty e^{-\lambda t} T_t g dt \right\| \leq \int_0^\infty e^{-\lambda t} \| g \| dt = \frac{1}{\lambda} \| g \|$. Also,

$$\begin{aligned} T_h f &= \int_0^\infty e^{-\lambda t} T_{t+h} g dt = \int_h^\infty e^{-\lambda (t-h)} T_t g dt \\ &= e^{\lambda h} \int_h^\infty e^{-\lambda t} T_t g dt = e^{\lambda h} \left(f - \int_0^h e^{-\lambda t} T_t g dt \right) \end{aligned}$$

by definition of f. So, $\frac{T_h f - f}{h} = \frac{e^{\lambda h} - 1}{h} f - \frac{1}{h} e^{\lambda h} \int_0^h e^{-\lambda t} T_t g dt$. Letting $h \downarrow 0$, the right side of this equality goes to $\lambda f - g$. Therefore, $f \in \mathcal{D}_A$, with $Af = \lambda f - g$, as desired.

We now treat uniqueness. Let g such that the equation $\lambda f - Af = g$ has two solutions. By subtracting these solutions, $\exists \ \phi \in \mathcal{D}_A$ with $\lambda \phi - A \phi = 0$. Since $d(T_t \phi)/dt = A T_t \phi = T_t A \phi = \lambda T_t \phi$ by (4), we conclude that $d(e^{-\lambda t} T_t \phi)/dt = e^{-\lambda t} \lambda T_t \phi + T_t \phi (-\lambda) e^{-\lambda t} = 0$. Integrating this in s, we get $\int_0^t \frac{d(e^{-\lambda s} T_s \phi)}{ds} ds = e^{-\lambda t} T_t \phi - \phi = 0$. So, $\| \phi \| \leq e^{-\lambda t} \| \phi \|$ for all $t > 0$, i.e. $\phi = 0$.

Proof of (7): Let $f \in \mathcal{D}$. Then $f - Uf =: g \in H_0$. By (6), $\exists \ \tilde{f} \in \mathcal{D}_A$ with $\tilde{f} - A \tilde{f} = g$. Since $A \subseteq U$, we conclude that $\tilde{f} - U \tilde{f} = g$ as well. Therefore, $h := f - \tilde{f} \in \mathcal{D}$ satisfies $Uh = h$, so $h = 0$, i.e. $f = \tilde{f} \in \mathcal{D}_A$. □

Theorem 10.40. *(**Characterization of Semigroups**) Let $T_t \colon H \to H$ be a strongly continuous semigroup of bounded linear operators (so $H = H_0$). Then its infinitesimal generator A is a closed densely defined linear function. T_t is uniquely determined by A in the sense that distinct semigroups have distinct infinitesimal generators. Moreover, if $f \in \mathcal{D}_A$ then $u(t) := T_t f$ solves the differential equation*

$$\frac{du}{dt}(t) = Au(t), \qquad t \geq 0, u(0) = f$$

Proof. Closedness of A follows from Thm. 10.39(5), and the density of $\mathcal{D}_A$ in H_0 follows from Thm. 10.39(3). We now prove two claims

Claim 1: Let $\phi(t)$ be a Borel function such that for all $\lambda > 0$, $\int_0^\infty e^{-\lambda t}\phi(t)dt = 0$. Then $\phi(t) = 0$ for almost all t.

Claim 2: Let $u_t \colon [0, \infty) \to H$ be a strongly continuous function such that for all $\lambda > 0$, $\int_0^\infty e^{-\lambda t} u_t dt = 0$. Then $u_t = 0$ for all t.

Proof of Claim 1: Letting $\lambda = n = 1, 2, \ldots$, and substituting $z = e^{-t}$ gives $\int_0^1 z^{n-1}\psi(z)dz = 0$ with $\psi(z) := \phi(-\log(z))$. (In particular, $\psi \in L_1$.) Applying Weierstrass's Approximation Theorem, $\int_0^1 f(z)\psi(z)dz = 0$ for all continuous functions f. Arguing as in Thm. 10.10, $\int_0^1 f(z)\psi(z)dz = 0$ for all bounded measurable functions f. In particular, letting $m > 0$ and $f(z) = \psi(z)1_{|\psi| \leq m}$ gives $\int_0^1 \psi(z)^2 1_{|\psi| \leq m}(z)dz = 0$. Letting $m \to \infty$ by the Monotome Convergence Theorem, $\int_0^1 \psi(z)^2 dz = 0$, i.e. $\psi = 0$, i.e. $\phi = 0$ as desired.

Proof of Claim 2: Let ξ be a linear functional. Then $\int_0^\infty e^{-\lambda t}\xi(u_t)dt = 0$, so $\xi(u_t) = 0$ for almost all t by Claim 1. Since $\xi(u_t)$ is continuous, $\xi(u_t) = 0$ for all t. By Hausdorfness of the weak topology, $u_t = 0$.

We now prove our uniqueness statement. Suppose two semigroups T_t, T'_t have the same infinitesimal generator A. We will show that $T_t f = T'_t f$ for all $f \in H_0$ (with H_0 identical for both T_t and T'_t). First, recall from Thm. 10.39(3) that H_0 is the strong closure of the domain of the infinitesimal generator. Therefore, H_0 is in fact the same for both T_t and T'_t. By Thm. 10.39(6), R_λ is identical for T_t, T'_t. As a result, $\forall\ f \in H_0\ \forall\ \lambda > 0$, $\int_0^\infty e^{-\lambda t} T_t f dt = \int_0^\infty e^{-\lambda t} T'_t f dt$. By Claim 2, $T_t f - T'_t f = 0\ \forall\ f \in H_0\ \forall\ \lambda > 0$, giving our desired uniqueness.

Let $f \in \mathcal{D}_A$. We now show that $u_t := T_t f$ is the unique solution of the equation $du_t/dt = Au_t$ such that (i) u_t is strongly differentiable and its derivative is strongly continuous for $t \geq 0$, (ii) $\|u_t\|$ is bounded, and (iii) $u_0 = \lim_{t \downarrow 0} u_t = f$.

Thm. 10.39(4) shows that u_t satisfies $du_t/dt = Au_t$ and (i) holds (using also $Af \in H_0$ by Thm. 10.39(2)). Then (ii) follows from (i) since $\|T_t f\| \leq \|f\|$. Also, using Thm. 10.39(3), $f \in \mathcal{D}_A \subseteq H_0$, so (iii) holds. We now prove uniqueness. Suppose u_t is a solution of du_t/dt satisfying (i),(ii) and such that $u_0 = \lim_{t \downarrow 0} u_t = 0$. We will show that $u_t = 0$. Define $v_t := e^{-\lambda t} u_t$. By definition of u_t, $dv_t/dt = -\lambda e^{-\lambda t} u_t + e^{-\lambda t} du_t/dt = -\lambda v_t + Av_t$. So, by Thm. 10.39(6), $v_t = -R_\lambda(dv_t/dt)$.

Now, dv_t/dt is strongly continuous, and R_λ is bounded by Thm. 10.39(6), so $\int_0^t v_s ds = -R_\lambda \int_0^t (dv_s/ds)ds = -R_\lambda(v_t - v_0) = -R_\lambda v_t$. By (ii), $\|v_t\| \to 0$ as $t \to \infty$, so $\|R_\lambda v_t\| \leq \|R_\lambda\| \, \|v_t\| \to 0$ as $t \to \infty$. So, for all $\lambda > 0$, $\int_0^\infty v_s ds = \int_0^\infty e^{-\lambda s} u_s ds = 0$. Applying Claim 2 shows $u_t = 0$ for all $t > 0$, as desired. □

11. Appendix: Isoperimetric Inequalities

Theorem 11.1. ***(Prékopa-Leindler Inequality)*** *Let $f, g, m\colon \mathbb{R}^n \to [0, \infty)$ be integrable. Suppose $\forall x, y \in \mathbb{R}^n$ $m(\lambda x + (1-\lambda)y) \geq f(x)^\lambda g(y)^{1-\lambda}$. Then*

$$\int_{\mathbb{R}^n} m(x)dx \geq \left(\int_{\mathbb{R}^n} f(x)dx\right)^\lambda \left(\int_{\mathbb{R}^n} g(x)dx\right)^{1-\lambda}$$

Proof. We induct on n. Let μ_n denote Lebesgue measure. Note that $\mu_1(A+B) \geq \mu_1(A) + \mu_1(B)$ for $A, B \subseteq \mathbb{R}$. (To prove this, by approximation we may assume compactness, and then translation with $\max(A) = 0 = \min(B)$ shows $A + B \supseteq A \cup B$, etc.) By scaling, we may assume $\|f\|_\infty = \|g\|_\infty = 1$. Let $A := \{x\colon f(x) \geq t\}$, $B := \{x\colon g(x) \geq t\}$, $C := \{x\colon m(x) \geq t\}$. Let $x \in A, y \in B$. Then $m(\lambda x + (1-\lambda)y) \geq f(x)^\lambda g(y)^{1-\lambda} \geq t$. Thus, $\lambda A + (1-\lambda)B \subseteq C$. Applying our claim, we have $\mu_1(C) \geq \mu_1(\lambda A + (1-\lambda)B) \geq \lambda\mu_1(A) + (1-\lambda)\mu_1(B)$. So, applying the definitions of A, B, C, integrating with respect to t, and applying AMGM gives

$$\int_{\mathbb{R}} m(x)dx \geq \lambda \int_{\mathbb{R}} f(x)dx + (1-\lambda)\int_R g(x)dx \geq (\int_{\mathbb{R}} f(x)dx)^\lambda (\int_{\mathbb{R}} g(x)dx)^{1-\lambda}$$

Now, let $(t, s) \in \mathbb{R} \times \mathbb{R}^{n-1}$. We can check that $f_{t_0}(s) = f(t_0, s)$, $g_{t_1}(s) = g(t_1, s)$, $m_t(s) = m(t, s)$ satisfy the hypotheses of the theorem, for $t = \lambda t_0 + (1-\lambda)t_1$. So, we can apply the inductive hypothesis (for the case $n-1$). But the result is the $n = 1$ hypotheses, viewed as functions of the t's. Thus applying the $n = 1$ case of the theorem gives the desired result. □

Theorem 11.2. ***(Brunn-Minkowski Inequality)*** *Let $A, B \subseteq \mathbb{R}^n$ measurable, nonempty.*

(1) $\forall$ $\lambda \in [0,1]$ $\mathrm{vol}(\lambda A + (1-\lambda)B) \geq \mathrm{vol}(A)^\lambda \mathrm{vol}(B)^{1-\lambda}$
(2) $\mathrm{vol}(A+B)^{1/n} \geq \mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}$

Proof. To prove (1), let $m = 1_{(\lambda A + (1-\lambda)B)}$, $f = 1_A$, $g = 1_B$, and apply Prékopa-Leindler (Thm. 11.1). To prove (2), let $\tilde{A} = \frac{A}{\mathrm{vol}(A)^{1/n}}$, $\tilde{B} = \frac{B}{\mathrm{vol}(B)^{1/n}}$, $\lambda = \frac{\mathrm{vol}(A)^{1/n}}{\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}}$, and apply (1). □

Theorem 11.3. ***(The Isoperimetric Inequality)*** *Let $A \subseteq \mathbb{R}^n$ as above. Let $B = rB_2^n$ be the Euclidean ball, scaled so that $\mathrm{vol}(A) = \mathrm{vol}(B) = r^n \mathrm{vol}(B_2^n)$. Then $\mathrm{vol}_{n-1}(\partial B) \leq \mathrm{vol}_{n-1}(\partial A)$.*

Proof. We show that $\mathrm{vol}(A_\varepsilon) \geq \mathrm{vol}(B_\varepsilon)$. Observe that $r = \left(\frac{\mathrm{vol}(A)}{\mathrm{vol}(B_2^n)}\right)^{\frac{1}{n}}$. So,

$$\begin{aligned} \mathrm{vol}(A_\varepsilon)^{\frac{1}{n}} &= \mathrm{vol}(A + \varepsilon B_2^n)^{\frac{1}{n}} && \text{, definition of } A_\varepsilon \\ &\geq \mathrm{vol}(A)^{\frac{1}{n}} + \mathrm{vol}(\varepsilon B_2^n)^{\frac{1}{n}} && \text{, Brunn-Minkowski, Thm. 11.2} \\ &= r\,\mathrm{vol}(B_2^n)^{\frac{1}{n}} + \varepsilon\,\mathrm{vol}(B_2^n)^{\frac{1}{n}} && \text{, definition of } r \\ &= \mathrm{vol}((r+\varepsilon)B_2^n)^{\frac{1}{n}} \\ &= \mathrm{vol}(B_\varepsilon)^{\frac{1}{n}} \end{aligned}$$

as desired. □

Theorem 11.4 (**Spherical Isoperimetric Inequality**)**.** *Among all domains of fixed volume on the sphere, one with minimal boundary volume is the geodesic ball.*

Proof due to Figiel-Lindenstrauss-Milman-1977. Idea: if we start with an optimal set that is not a geodesic ball, we can apply a finite number of symmetrizations to it so that its interior is squished into a smaller region. This gives a contradiction, so we had a ball at the beginning. The technical device of outer radius allows the argument to proceed rigorously.

Let $S^{n-1} \subseteq \mathbb{R}^n$ be the unit sphere centered at the origin, and let $A \subseteq S^{n-1}$ be closed. Given two antipodal points $a, b \in S^{n-1}$, let $\gamma \subseteq S^{n-1}$ be a geodesic joining a and b. We define the symmetrization $\sigma_\gamma(A)$ as follows. For each $y \in \gamma$, let Π_y be the plane containing y, such that Π_y is perpendicular to the line in $\mathbb{R}^n$ connecting a and b. Note that $\Pi_y \cap S^{n-1}$ is a dilation and translation of S^{n-2}, so let $\mu_{n-2,y}$ be the normalized Haar measure on $\Pi_y \cap S^{n-1}$. We let $\sigma_\gamma(A) \cap \Pi_y$ be a geodesic ball in S^{n-2} with center y, such that $\mu_{n-2,y}(\sigma_\gamma(A) \cap \Pi_y) = \mu_{n-2,y}(A \cap \Pi_y)$. From Fubini's Theorem, we have $\mu_{n-1}(B) = \mu_{n-1}(A)$, where μ_{n-1} is normalized Haar measure on S^{n-1}.

We say $\sigma_\gamma(A)$ is the **symmetrization** of A with respect to γ. Let $r(A) := \min\{r > 0 \colon \exists\, x \in S^{n-1}, A \subseteq \overline{B(x,r)}\}$ be the (outer) **radius** of A. Here $B(x,r) := \{y \in S^{n-1} \colon d(x,y) < r\}$ is the open ball of radius r centered at y on S^{n-1}, and d is the usual metric on S^{n-1}, so that $d(x,y) = \cos^{-1}(\langle x, y\rangle)$ for all $x, y \in S^{n-1}$. The minimum in the definition of $r(A)$ exists by closedness of A and $\overline{B(x,r)}$. We claim that $\sigma_\gamma(A)$ is closed.

To show this, we use the Hausdorff distance on closed sets in S^{n-1}, $\delta(A,B) := \min\{r > 0 \colon A_r \supseteq B, B_r \supseteq A\}$. (Here $A_r := \{x \in S^{n-1} \colon d(x,A) < r\}$.) Let $B := \sigma_\gamma(A)$. Recall that the set of closed sets is a complete metric space with respect to the metric δ. Note that the function $y \mapsto \mu_{n-2,y}(A) = \mu_{n-2,y}(B)$ is upper semicontinuous in $y \in \gamma$, i.e. $\mu_{n-2,y_0}(A) \geq \limsup_{y \to y_0} \mu_{n-2,y}(A)$ when $y, y_0 \in \gamma$. This follows by the definition of the product topology and by the closedness of A. Now, writing S^{n-1} as $S^{n-2} \times [-1,1]/\sim$ where $(x,1) \sim (x',1)$ and $(x,-1) \sim (x',-1)$ $\forall\, x, x' \in S^{n-2}$, we can treat $A \subseteq S^{n-1}$ as a closed set in the product topology of $S^{n-2} \times [-1,1]$. Given $(x,y) \in B^c \subseteq S^{n-1} \times [-1,1]$, we wish to find a box $F \times G \subseteq S^{n-2} \times [-1,1]$ with F, G open, so that $(x,y) \in F \times G$ and $F \times G$ is disjoint from B. Since $B \cap \Pi_y$ is a geodesic ball (which is not all of S^{n-2}), we can find $F \times G$ as required, by the upper semicontinuity of $y \mapsto \mu_{n-2,y}(B)$. (Specifically, our inability to find such a box $F \times G$ would violate this upper semicontinuity.)

Below we also use that $\mu_{n-1}(\cdot)$ is upper semi-continuous with respect to δ, that is if $A^{(1)}, A^{(2)}, \ldots \subseteq S^{n-1}$ satisfy $\lim_{k\to\infty} \delta(A^{(k)}, A) = 0$, then $\mu_{n-1}(A) \geq \limsup_{k\to\infty} \mu_{n-1}(A^{(k)})$. To see this, let $x_k \in A^{(k)}$ for any $k \geq 1$. Since $d(x_k, A) \leq \delta(A^{(k)}, A) \to 0$ as $k \to \infty$, any limit point of the set $\{x_k\}_{k=1}^\infty$ must be contained in A. Therefore, for any fixed $\varepsilon > 0$, there exists $K > 0$ such that $k \geq K$ implies $A^{(k)} \subseteq A_\varepsilon$. Let $\lambda > \mu_{n-1}(A)$. Since μ_{n-1} is a Borel measure, there exists an open set U such that $A \subseteq U$ and $\mu_{n-1}(U) < \lambda$. Since A is compact, $d(A, U^c) > 0$, and there exists $\varepsilon > 0$ such that $A_\varepsilon \subseteq U$. Combining these observations, $\limsup_{k\to\infty} \mu_{n-1}(A^{(k)}) \leq \mu_{n-1}(A_\varepsilon) \leq \mu_{n-1}(U) < \lambda$. Therefore, $\limsup_{k\to\infty} \mu_{n-1}(A^{(k)}) \leq \mu_{n-1}(A)$, as desired.

We are now ready to proceed by inducting on n. For the case $n = 1$, the theorem is clear. We require the following claims, which are proven by induction.

Claim 1: Let $A \subseteq S^{n-1}$ be closed, and define

$$\begin{aligned} M(A) := \{C \subseteq S^{n-1} \colon & C \text{ is closed}, \\ & \mu_{n-1}(C) = \mu_{n-1}(A),\ \mu_{n-1}(C_\varepsilon) \leq \mu_{n-1}(A_\varepsilon)\ \forall\, \varepsilon > 0\} \end{aligned}$$

Then there is a $B \in M(A)$ with minimal radius, i.e. $\min\{r(C) \colon C \subseteq M(A)\}$ exists.

Claim 2: Let $A \subseteq S^{n-1}$ be closed. Then for every half circle γ, $\sigma_\gamma(A) \in M(A)$.

Claim 3: Let $B \subseteq S^{n-1}$ be a closed set that is not a geodesic ball. There exists a finite family of half circles $\{\gamma_i\}_{i=1}^n \subseteq S^{n-1}$ so that $r(\sigma_{\gamma_n}(\sigma_{\gamma_{n-1}}(\cdots\sigma_{\gamma_1}(B)\cdots))) < r(B)$.

We prove the theorem assuming these claims. By definition of $M(A)$, $B \in M(A)$ and $C \in M(B)$ implies $C \in M(A)$. So, using Claim 2, $B \in M(A)$ and $\sigma_{\gamma_1}(B) \in M(B)$ implies $\sigma_{\gamma_1}(B) \in M(A)$, $\sigma_{\gamma_2}(\sigma_{\gamma_1}(B)) \in M(A)$, etc. Using Claim 3, we therefore see that an element of minimal (outer) radius in $M(A)$ must be a geodesic ball. Claim 1 says that this minimal element must exist, so $M(A)$ must contain a geodesic ball. The theorem is therefore proven. We now prove the claims.

Proof of Claim 1: $B \mapsto r(B)$ is continuous (with respect to the Hausdorff metric for $B \subseteq S^{n-1}$), so it suffices to show that $M(A)$ is a closed subset in the space of closed subsets of S^{n-1} (since the latter space is compact with respect to δ). Let $B^{(1)}, B^{(2)}, \ldots \in M(A)$ with $\lim_{k\to\infty} \delta(B^{(k)}, B) = 0$ for some $B \subseteq S^{n-1}$, and let $\varepsilon \geq 0$. We will show $B \in M(A)$. For any fixed $\eta > 0$, there exists $K > 0$ such that, if $k \geq K$, then $B \subseteq B^{(k)}_\eta$, so $B_\varepsilon \subseteq B^{(k)}_{\varepsilon+\eta}$. So, for all $k \geq K$, $\mu_{n-1}(B_\varepsilon) \leq \mu_{n-1}(B^{(k)}_{\varepsilon+\eta}) \leq \mu_{n-1}(A_{\varepsilon+\eta})$, since $B^{(1)}, B^{(2)}, \ldots \in M(A)$. Therefore,

$$\mu_{n-1}(B_\varepsilon) \leq \inf_{\eta>0} \mu_{n-1}(A_{\varepsilon+\eta}) = \mu_{n-1}(\cap_{\eta>0} A_{\varepsilon+\eta}) = \mu_{n-1}(A_\varepsilon)$$

So, letting $\varepsilon = 0$, we get $\mu_{n-1}(B) \leq \mu_{n-1}(A)$. Moreover, $\mu_{n-1}(B) \geq \limsup_{k\to\infty} \mu_{n-1}(B^{(k)}) = \mu_{n-1}(A)$, using the upper semicontinuity of $\mu_{n-1}(\cdot)$ mentioned above, and the definition of $B^{(1)}, B^{(2)}, \ldots \in M(A)$. So $B \in M(A)$, as desired.

Proof of Claim 2: Let $A \subseteq S^{n-1}$ be closed and let γ be a half circle on S^{n-1} joining $z \in S^{n-1}$ with $-z$. Let u be the midpoint of γ. As usual, identify $S^{n-2,u} := S^{n-1} \cap \Pi_u$ with S^{n-2}. For any $y \in \gamma, y \neq \pm x$, define a map $\tau_y \colon S^{n-2,y} \to S^{n-2,u}$ by letting $\tau_y(x) := \gamma \cap S^{n-2,u}$ for any $x \in S^{n-2,y}$. (Note that this intersection is a single point). By applying polar coordinates, we see that there exists a function f such that, if $y_1, y_2 \in \gamma$ and if $x_1 \in S^{n-2,y_1}, x_2 \in S^{n-2,y_2}$, we have

$$d(x_1, x_2) = f(y_1, y_2, d(\tau_{y_1}(x_1), \tau_{y_2}(x_2))).$$

Moreover, for y_1, y_2 fixed, f is monotonically increasing with respect to its third argument, $d(\tau_{y_1}(x_1), \tau_{y_2}(x_2)) \leq \pi$.

For every $y_1, y_2 \in \gamma, \varepsilon > 0$ (with $d(y_1, y_2) < \varepsilon$) there is an $\eta(y_1, y_2, \varepsilon)$ so that, for every $C \subseteq S^{n-2,y_1}$, we have

$$C_\varepsilon \cap S^{n-2,y_2} = \tau_{y_2}^{-1}((\tau_{y_1} C)_{\eta(y_1,y_2,\varepsilon)}) \qquad (*)$$

To see this, it suffices to consider the case that $C = \{x_1\}$. Then

$$\begin{aligned} C_\varepsilon \cap S^{n-2,y_2} &= \{x_2 \in S^{n-2,y_2} \colon d(x_1, x_2) < \varepsilon\} \\ &= \{x_2 \in S^{n-2,y_2} \colon f(y_1, y_2, d(\tau_{y_1}(x_1), \tau_{y_2}(x_2))) < \varepsilon\} \\ &= \{x_2 \in S^{n-2,y_2} \colon d(\tau_{y_1}(x_1), \tau_{y_2}(x_2)) < \eta\} \end{aligned}$$

Here η is determined by the existence and monotonicity of f. (If $d(y_1, y_2) \geq \varepsilon$, then $C_\varepsilon \cap S^{n-2,y_2} = \emptyset$.) Note that the subscript ε on the left of $(*)$ denotes an ε neighborhood in S^{n-1}, whereas the subscript η on the right of $(*)$ denotes an η neighborhood in S^{n-2}. Let $A^y := A \cap S^{n-2,y}$. By fixing $y_2 = y$ and varying $y_1 = z$ in $(*)$, we have

$$\tau_y((A_\varepsilon)^y) = \cup_{\{z\in\gamma\colon d(z,y)<\varepsilon\}} (\tau_z(A^z))_{\eta(z,y,\varepsilon)} \qquad (**)$$

Substituting $B := \sigma_\gamma(A)$ gives

$$\tau_y((B_\varepsilon)^y) = \cup_{\{z\in\gamma\colon d(z,y)<\varepsilon\}}(\tau_z(B^z))_{\eta(z,y,\varepsilon)} \qquad (\dagger)$$

By definition of B, $\tau_z(B^z)$ is a geodesic ball in $S^{n-2,u}$ $\forall$ $z \in \gamma$, and $\mu_{n-2,u}(\tau_z(B^z)) = \mu_{n-2,u}(\tau_z(A^z))$. So, the induction hypothesis (i.e. the full theorem) says

$$\mu_{n-2,u}((\tau_z(B^z))_{\eta(z,y,\varepsilon)}) \leq \mu_{n-2,u}((\tau_z(A^z))_{\eta(z,y,\varepsilon)}) \qquad (\ddagger)$$

for admissible y, z, ε. Since the sets on the right side of $(\dagger)$ are all $(n-2)$-dimensional geodesic balls with the same center, we have

$$\begin{aligned}\mu_{n-2,u}(\tau_y(B_\varepsilon)^y) &= \sup_{z\in\gamma\colon d(z,y)\leq\varepsilon} \mu_{n-2,u}((\tau_z(B^z))_{\eta(z,y,\varepsilon)}) \\ &\leq \sup_{z\in\gamma\colon d(z,y)\leq\varepsilon} \mu_{n-2,u}((\tau_z(A^z))_{\eta(z,y,\varepsilon)}) \quad \text{, from } (\ddagger) \\ &\leq \mu_{n-2,u}(\tau_y(A_\varepsilon)^y) \quad \text{, from } (**)\end{aligned}$$

Re-writing this inequality, we see that for every $y \in \gamma, y \neq \pm x$ we have

$$\mu_{n-2,y}((B_\varepsilon)^y) \leq \mu_{n-2,y}((A_\varepsilon)^y)$$

So by Fubini's Theorem, we can integrate this inequality to get $\mu_{n-1}(B_\varepsilon) \leq \mu_{n-1}(A_\varepsilon)$, so that $B \in M(A)$ as desired.

Proof of Claim 3: Let $B \subseteq S^{n-1}$ be closed, and suppose B is not a geodesic ball. Let $r = r(B)$ as above, and let $u \in S^{n-1}$ be such that $B \subseteq \overline{B(u,r)}$. Let γ be a half circle with midpoint u, so that we will symmetrize with respect to γ, leaving $\overline{B(u,r)}$ fixed. Since B is not a geodesic ball, $E := B^c \cap \partial B(u,r) \neq \emptyset$.

We need two observations. First, any symmetrization σ_γ does not decrease the set E. That is, $E \subseteq (\sigma_\gamma(B))^c \cap \partial B(u,r)$. Second, we can find symmetrizations that increase E. To see the second claim, let $G \subseteq \partial B(u,r)$ be a relatively open set. Given any $x \in \partial B(u,r) \smallsetminus G$, there exists a relatively open set $G_x \subseteq \partial B(u,r)$ and γ_x such that $x \in G_x$, and $G_x \cap \sigma_{\gamma_x}(B) = \emptyset$. To construct γ_x, consider the straight line ℓ (in $\mathbb{R}^n$) between x and some point $y \in B^c \cap \partial B(u,r)$ (which exists since B is not a ball). Let P reflect $\partial B(u,r)$ across a hyperplane perpendicular to ℓ and intersecting ℓ at its midpoint. Then, let G_x be a small ball (in $\partial B(u,r)$) around x disjoint from G, such that $PG_x \subseteq B^c \cap \partial B(u,r)$ (which is possible since B is closed). Observe that G_x does what we claimed above. Also note that G_x, γ_x depend on x and G, but not on B.

Now, apply the above observations to B and $G := B^c \cap \partial B(u,r)$ to produce γ_1, G_{x_1}. Then, apply these same observations to $\sigma_{\gamma_1}(B)$ and $G := \sigma_{\gamma_1}(B)^c \cap \partial B(u,r)$ to produce γ_2 and G_{x_2}, and so on. By compactness of S^{n-1} (using a cover by $\{G_{x_i}\}_{i\geq 1}$), after a finite number of symmetrizations we have $\sigma_{\gamma_n}(\cdots\sigma_{\gamma_1}(B)\cdots)$ disjoint from $\partial B(u,r)$. Therefore, $r(\sigma_{\gamma_n}(\cdots\sigma_{\gamma_1}(B)\cdots)) < r(B)$. □

As an application, we prove the following concentration of measure result. Note that the exponential dependence on n implies that almost all of a high dimensional sphere is close to any given set of Haar measure 1/2. Put another way, a high dimensional sphere has a "large waist."

Theorem 11.5 (**Concentration of measure on the sphere**). *Let μ be the normalized Haar measure on S^{n+1}. Let $A \subseteq S^{n+1}$, let $\varepsilon > 0$, and define $A_\varepsilon := \{x \in S^{n+1}\colon \exists y \in S^{n+1}$ with $d_{S^{n+1}}(x,y) \leq \varepsilon\}$. If $\mu(A) \geq 1/2$ then $\mu(A_\varepsilon) \geq 1 - \sqrt{\frac{\pi}{8}}e^{-\varepsilon^2 n/2}$.*

Proof. By Theorem. 11.4, it suffices to prove this claim for geodesic balls, i.e. it suffices to analyze the quantity

$$\mu(B(\pi/2+\varepsilon)) = \frac{\int_{-\pi/2}^{\varepsilon}\cos^n(t)dt}{\int_{-\pi/2}^{\pi/2}\cos^n(t)dt}.$$

For any $n \geq 1$, let $I_n := \int_0^{\pi/2}\cos^n(t)dt$. Changing variables and using $\cos(t) \leq e^{-t^2/2}$, valid for any $0 \leq t \leq \pi/2$ (which follows since $f(t) := \log\cos t$ satisfies $f''(t) = -1/\cos^2(t) \leq -1$ for all $0 \leq t \leq \pi/2$),

$$\begin{aligned} 1-\mu(B(\pi/2+\varepsilon)) &= \int_{\varepsilon}^{\pi/2}\cos^n(t)\frac{dt}{2I_n} = \frac{1}{\sqrt{n}}\int_{\varepsilon\sqrt{n}}^{(\pi/2)\sqrt{n}}\cos^n(t/\sqrt{n})\frac{dt}{2I_n} \\ &\leq \frac{1}{\sqrt{n}}\int_{\varepsilon\sqrt{n}}^{(\pi/2)\sqrt{n}} e^{-t^2/2}\frac{dt}{2I_n} \leq \frac{1}{\sqrt{n}}e^{-\varepsilon^2 n/2}\int_0^{(\pi/2-\varepsilon)\sqrt{n}} e^{-t^2/2}\frac{dt}{2I_n} \\ &\leq \frac{1}{\sqrt{n}}e^{-\varepsilon^2 n/2}\int_0^{\infty} e^{-t^2/2}\frac{dt}{2I_n} = \frac{1}{2\sqrt{n}I_n}e^{-\varepsilon^2 n/2}\sqrt{\pi/2}. \end{aligned}$$

Integration by parts shows that $I_n = \frac{n-1}{n}I_{n-2}$. Since $(n-1)/\sqrt{n(n-2)} \geq 1$ for any $n \geq 3$, we get $\sqrt{n}I_n \geq \sqrt{n-2}I_{n-2}$ for any $n \geq 3$, so that

$$\sqrt{n}I_n \geq \min(I_1, \sqrt{2}I_2) = \min(1, \sqrt{2}\pi/4) = 1, \qquad \forall\, n \geq 1$$

In summary, $1-\mu(B(\pi/2+\varepsilon)) \leq e^{-\varepsilon^2 n/2}\sqrt{\pi/8}$. □

Theorem 11.5 implies a corresponding statement for Lipschitz functions. That is, Lipschitz functions on high-dimensional spheres are typically close to their average value.

For any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we denote $\|x\| := (x_1^2 + \cdots + x_n^2)^{1/2}$.

Theorem 11.6 (**Concentration of measure, Lipschitz function form**)**.** *Let $f\colon S^{n+1} \to \mathbb{R}$. Suppose that for all $x, y \in S^{n+1}$, $|f(x) - f(y)| \leq \|x - y\|$, so that f is 1-Lipschitz. Let μ denote normalized Haar measure on S^{n+1}. Then for all $\varepsilon > 0$,*

$$\mu\left(x \in S^{n+1}\colon \left|f(x) - \int_{S^{n+1}} f(y)d\mu(y)\right| \geq \varepsilon\right) \leq \sqrt{\frac{\pi}{2}}e^{-n\varepsilon^2/4}.$$

Proof. Let $m \in \mathbb{R}$ such that $\mu(x \in S^{n+1}\colon f(x) \leq m) \geq 1/2$ and $\mu(x \in S^{n+1}\colon f(x) \geq m) \geq 1/2$. Let $C := \{x \in S^{n+1}\colon f(x) \leq m\}$. Then $x \in C_\varepsilon$ if and only if $\exists\, y \in C$ with $\|x-y\|_2 \leq \varepsilon$. Since f is 1-Lipschitz, $|f(x) - f(y)| \leq \varepsilon$, so that $f(x) \leq m + \varepsilon$. Taking the contrapositive,

$$\{x \in S^{n+1}\colon f(x) > m + \varepsilon\} \subseteq S^{n+1} \smallsetminus A_\varepsilon$$

So, from Thm. 11.5, since $\mu(C) \geq 1/2$, we have

$$\mu\left(x \in S^{n+1}\colon f(x) > m + \varepsilon\right) \leq \sqrt{\pi/8}e^{-n\varepsilon^2/2}.$$

Similarly, $\mu\left(x \in S^{n+1}\colon f(x) < m - \varepsilon\right) \leq \sqrt{\pi/8}e^{-n\varepsilon^2/2}$. In conclusion,

$$\mu\left(x \in S^{n+1}\colon |f(x) - m| > \varepsilon\right) \leq 2\sqrt{\pi/8}e^{-n\varepsilon^2/2}. \qquad (*)$$

It remains to replace m with $\int_{S^{n+1}} f(y)d\mu(y)$. Consider $\mu \times \mu$ on $S^{n+1} \times S^{n+1}$. Observe

$$\begin{aligned}
&(\mu \times \mu)\big((x,y) \in S^{n+1} \times S^{n+1} \colon\ |f(x) - f(y)| \geq \varepsilon\big) \\
&\qquad \leq (\mu \times \mu)\left(\{|f(x) - m| \geq \varepsilon/2\} \cup \{|f(y) - m| \geq \varepsilon/2\}\right) \\
&\qquad \leq 2\mu(|f(x) - m| \geq \varepsilon/2) \leq 4\sqrt{\pi/8}e^{-n\varepsilon^2/2} \quad \text{, from } (*)
\end{aligned}$$

Let $\lambda > 0$. Then from Theorem 11.5, if $\lambda^2 := n/4$,

$$\begin{aligned}
&\int_{S^{n+1}\times S^{n+1}} e^{\lambda^2(f(x)-f(y))^2} d\mu(x)d\mu(y) \\
&\qquad = \int_0^\infty 2\lambda^2 t e^{\lambda^2 t^2} (\mu \times \mu)\big((x,y) \in S^{n+1} \times S^{n+1} \colon\ |f(x) - f(y)| \geq t\big)\, dt \\
&\qquad \leq 4\sqrt{\pi/8} \int_0^\infty \lambda^2 t e^{\lambda^2 t^2} e^{-nt^2/2} dt = \sqrt{\pi/8} \int_0^\infty t n e^{-nt^2/4} dt = 2\sqrt{\pi/8} = \sqrt{\pi/2}.
\end{aligned}$$

So, for this λ, Jensen's inequality in y implies that

$$\sqrt{\pi/2} \geq \int_{S^{n+1}\times S^{n+1}} e^{\lambda^2(f(x)-f(y))^2} d\mu(x)d\mu(y) \geq \int_{S^{n+1}} e^{\lambda^2(f(x)-\int_{S^{n+1}} f(y)d\mu(y))^2} d\mu(x).$$

Finally, by Chebyshev's inequality,

$$\begin{aligned}
\mu(x \in S^{n+1} \colon |f(x) - \int_{S^{n+1}} f d\mu| \geq \varepsilon) &= \mu(x \in S^{n+1} \colon e^{\lambda^2 |f(x) - \int_{S^{n+1}} f(y)d\mu(y)|^2} \geq e^{\lambda^2\varepsilon^2}) \\
&\leq e^{-\lambda^2\varepsilon^2} \int_{S^{n+1}} e^{\lambda^2 \left|f(x) - \int_{S^{n+1}} f(y)d\mu(y)\right|^2} d\mu(x) \leq \sqrt{\pi/2} e^{-\lambda^2\varepsilon^2}.
\end{aligned}$$

□

Theorem 11.7. ***(Spheres and Gaussians, Strong Version)*** *Let γ_n be the standard Gaussian measure on $\mathbb{R}^n$, and let $A \subseteq \mathbb{R}^n$ be a Borel set. Let $\sigma^N_{\sqrt{N}}$ denote the normalized Haar measure on $\sqrt{N} \cdot S^N$. Let $P_{N+1,n}$ (for $N \geq n$) be the standard linear projection from $\mathbb{R}^{N+1}$ onto $\mathbb{R}^n$, i.e. $P_{N+1,n}(x_1, \ldots, x_{N+1}) = (x_1, \ldots, x_n)$. Then*

$$\lim_{N\to\infty} \sigma^N_{\sqrt{N}}(P^{-1}_{N+1,n}(A) \cap (\sqrt{N} \cdot S^N)) = \gamma_n(A)$$

Proof. Let $\{g_i\}$ be iid standard Gaussian random variables, and define $R_N^2 := g_1^2 + \cdots + g_N^2$. By rotation invariance, $\frac{\sqrt{N}}{R_{N+1}}(g_1, \ldots, g_{N+1})$ has the same distribution as $\sigma^N_{\sqrt{N}}$. So, by projecting both sides, $\frac{\sqrt{N}}{R_{N+1}}(g_1, \ldots, g_n)$ has the same distribution as $P_{N+1,n}(\sigma^N_{\sqrt{N}}) = \sigma^N_{\sqrt{N}}(P^{-1}_{N+1,n}(\cdot))$ (for $N \geq n$). In the previous sentence, we calculated both mentioned measures as pushforward measures, under the map $P_{N+1,n} \colon \mathbb{R}^{N+1} \to \mathbb{R}^n$.

$R_N^2/N \to 1$ a.s. by the Strong Law of Large Numbers, giving a weak convergence result. However, we need to be more precise. Note that $R_n^2, R_{N+1}^2 - R_n^2$ and $(g_1, \ldots, g_n)/R_n$ are independent of each other. The first two are independent by definition, and the third is independent from the others by, say, invoking polar coordinates. (The third term is the "angular" part of the Gaussian vector, and the first is the "radial" part.) Therefore, R_n^2/R_{N+1}^2 is independent of $(g_1, \ldots, g_n)/R_n$.

Now, write $R_n^2/R_{N+1}^2 = 1/(1 + (g_{n+1}^2 + \cdots + g_{N+1}^2)/(g_1^2 + \cdots + g_n^2))$ and observe that the ratio of g_i's is a ratio of two independent chi squared distributions. A computation then shows that R_n^2/R_{N+1}^2 has beta distribution with parameters $n/2$ and $(N+1-n)/2$. The corresponding distribution function is therefore

$$\frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{N-n+1}{2}\right)}x^{\frac{n}{2}-1}(1-x)^{\frac{N-n+1}{2}-1}$$

Combining our results and using the definition of the pushforward, we have

$$\begin{aligned}
&\sigma^N_{\sqrt{N}}(P^{-1}_{N+1,n}(A)\cap(\sqrt{N}\cdot S^N)) = \mathbb{P}(P_{N+1,n}(\sigma^N_{\sqrt{N}})\in A)\\
&= \mathbb{P}\left(\frac{\sqrt{N}}{R_{N+1}}(g_1,\ldots,g_n)\in A\right) = \mathbb{P}\left(\left(N\frac{R_n^2}{R_{N+1}^2}\right)^{1/2}\cdot\frac{1}{R_n}(g_1,\ldots,g_n)\in A\right)\\
&= \frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{N-n+1}{2}\right)}\int_{S^{n-1}}\int_0^1 1_A(\sqrt{N}tx)t^{\frac{n}{2}-1}(1-t)^{\frac{N-n+1}{2}-1}dtd\sigma^{n-1}(x)\\
&= \frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{N-n+1}{2}\right)}\frac{2}{N^{n/2}}\int_{S^{n-1}}\int_0^{\sqrt{N}} 1_A(ux)u^{n-1}\left(1-\frac{u^2}{N}\right)^{\frac{N-n+1}{2}-1}dud\sigma^{n-1}(x)
\end{aligned}$$

In the last line we used the change of variables $u=\sqrt{N}t$. Applying the Dominated Convergence Theorem and letting $N\to\infty$, the last quantity converges to

$$\frac{1}{2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right)}\int_{S^{n-1}}\int_0^\infty 1_A(ux)u^{n-1}e^{-u^2/2}dud\sigma^{n-1}(x)$$

which is $\gamma_n(A)$ in polar coordinates, as desired. (We used the formula $\Gamma(n+1)=n\Gamma(n)$ to get the correct constant in front of the integral.) □

We can now verify Gaussian isoperimetry via Thms. 11.7 and 11.4.

Theorem 11.8. *(**Gaussian Isoperimetric inequality**) For $A\subseteq\mathbb{R}^n$ define $d\gamma_n(x):=e^{-\|x\|_2^2/2}dx/(2\pi)^{n/2}$. Define $\gamma(\partial A):=\liminf_{\varepsilon\to0^+}(\gamma(A_\varepsilon)-\gamma(A))/\varepsilon$. Then among all sets with the same (Gaussian) volume, one with minimal (Gaussian) boundary measure is the half space.*

Proof. [Led96] Define $\Phi(t):=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^t e^{-x^2/2}dx$, $t\in\mathbb{R}$. We may assume that $\gamma_n(A)=\Phi(a)$ for some $a\in\mathbb{R}$. It suffices to show that

$$\gamma_n(A_r)\geq\Phi(a+r)$$

That is, we want $\Phi^{-1}(\gamma_n(A_r))\geq\Phi^{-1}(\gamma_n(A))+r$. Now, let $b\in\mathbb{R}$, $b<a$. Since $\Phi(a)=\gamma_n(A)$, we have

$$\gamma_n(A)=\Phi(a)>\Phi(b)=\gamma_1((-\infty,b])$$

So, for N large enough ($N\geq n$), Thm. 11.7 gives

$$\sigma^N_{\sqrt{N}}(P^{-1}_{N+1,n}(A)\cap(\sqrt{N})S^N)>\sigma^N_{\sqrt{N}}(P^{-1}_{N+1,1}((-\infty,b])\cap(\sqrt{N})S^N)\qquad(*)$$

Note that $P^{-1}_{N+1,n}(A_r)\cap(\sqrt{N})S^N\supseteq(P^{-1}_{N+1,n}(A)\cap(\sqrt{N})S^N)_r$, where the subscript on the left denotes a neighborhood in $\mathbb{R}^n$, and the subscript on the right denotes a neighborhood on $(\sqrt{N})S^N$ with respect to geodesic distance. This set containment follows since the map $P^{-1}_{n+1,n}$ increases distances. Note also that $P^{-1}_{N+1,1}((-\infty,b])\cap(\sqrt{N})S^N$ is a geodesic ball on

$(\sqrt{N})S^N$. This gives us two things. First, we can apply the spherical isoperimetric inequality. Second, we have $(P_{N+1,1}^{-1}((-\infty,b])\cap(\sqrt{N})S^N)_r = P_{N+1,1}^{-1}((-\infty,b+r(N)])$, where

$$r(N) = -\sqrt{N}\cos(\cos^{-1}(b/\sqrt{N}) + r/\sqrt{N}) + b, \qquad b<0$$

with a similar equality for $b>0$. In either case, the cosine addition formula shows that $\lim_{N\to\infty} r(N) = r$. Combining all of these observations, we have

$$\begin{aligned}
&\sigma_{\sqrt{N}}^N(P_{N+1,n}^{-1}(A_r)\cap(\sqrt{N})S^N) \\
&\quad\geq \sigma_{\sqrt{N}}^N((P_{N+1,n}^{-1}(A)\cap(\sqrt{N})S^N)_r) \quad \text{, by set containment} \\
&\quad\geq \sigma_{\sqrt{N}}^N((P_{N+1,n}^{-1}((-\infty,b])\cap(\sqrt{N})S^N)_r) \quad \text{, by Thm. 11.4, and } (*) \\
&\quad= \sigma_{\sqrt{N}}^N(P_{N+1,n}^{-1}((-\infty,b+r(N)])\cap(\sqrt{N})S^N)
\end{aligned}$$

So, letting $N\to\infty$ and applying Thm. 11.7, we get

$$\gamma_n(A_r) \geq \Phi(b+r)$$

And since $b<a$ is arbitrary, we get $\gamma_n(A_r)\geq\Phi(a+r)$ as desired. □

Proof. [Led96] Instead of using the isoperimetric inequality on the sphere, we argue more directly, as in Thm. 11.3, and we use Brunn-Minkowski (for Gaussian space, Thm. 11.9). Let $\lambda\in(0,1)$ and let $B\subseteq\mathbb{R}^n$ be the Euclidean ball with radius $r/(1-\lambda)$, so $B=(r/(1-\lambda))B_2^n$. From Gaussian Brunn-Minkowski (Thm. 11.9) we have

$$\begin{aligned}
\Phi^{-1}(\gamma_n(\lambda A)_r) = \Phi^{-1}(\gamma_n(\lambda A + rB_2^n)) &= \Phi^{-1}(\gamma_n(\lambda A+(1-\lambda)B)) \\
&\geq \lambda\Phi^{-1}(\gamma_n(A)) + (1-\lambda)\Phi^{-1}(\gamma_n(B)) \\
&= \lambda\Phi^{-1}(\gamma_n(A)) + (1-\lambda)\varepsilon(\lambda,r)\Phi^{-1}(\gamma_1(-\infty,r/(1-\lambda)]) \\
&= \lambda\Phi^{-1}(\gamma_n(A)) + \varepsilon(\lambda,r)r
\end{aligned}$$

where $\varepsilon(\lambda,r)\to1$ as $\lambda\to1$, for fixed r. So, letting $\lambda\to1$ gives $\Phi^{-1}(\gamma_n(A_r))\geq\Phi^{-1}(\gamma_n(A))+r$, as desired. □

Proof. [Bob97] We begin with a functional form of the isoperimetric inequality. We prove this via a "two-point" inequality, which is leveraged to the whole Gaussian space, via the central limit theorem. With Φ as above, and $\phi(t):=\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$, define $I:=\phi\circ\Phi^{-1}$. We want to prove

$$\gamma(\partial A)\geq I(\gamma(A)) \qquad (*)$$

This inequality is equivalent to $\gamma(A_r)\geq\Phi(\Phi^{-1}(\gamma(A))+r)$. To see this, differentiate the latter inequality to get $(*)$. Conversely, if $\gamma(\partial A)\geq I(\gamma(A))$, for all A, we can conclude that $\gamma(A_r)\geq\Phi(\Phi^{-1}(\gamma(A))+r)-\varepsilon$, for all small $\varepsilon>0$, with ε independent of A, but with r small and dependent on A. Now, let $p\in[0,1], r\in\mathbb{R}$, and consider $R_r(p):=\Phi(\Phi^{-1}(p)+r)$. Note that $R_{r_1+r_2}(p)=R_{r_1}(R_{r_2}(p))$ and $A_{r_1+r_2}=(A_{r_1})_{r_2}$. Thus, we can improve the inequality to $\gamma(A_r)\geq\Phi(\Phi^{-1}(\gamma(A))+r)-\varepsilon$ with r,ε both independent of A, since r small implies (via addition), the same inequality for r large. Letting $\varepsilon\to0$ then gives our result.

Now, we will deduce $(*)$ from the following functional inequality

$$I(\int f d\gamma) \leq \int \sqrt{I(f)^2+|\nabla f|^2}d\gamma \qquad (**)$$

where $f\colon \mathbb{R}^n \to [0,1]$. Surprisingly, we can deduce $(**)$ from the following inequality (which we will not prove)

$$I\left(\frac{a+b}{2}\right) \le \frac{1}{2}\sqrt{(I(a))^2 + \left|\frac{a-b}{2}\right|^2} + \frac{1}{2}\sqrt{(I(b))^2 + \left|\frac{a-b}{2}\right|^2}$$

where $0 \le a,b \le 1$. For $f\colon \{-1,1\} \to [0,1]$, write $f(-1) = a$ and $f(1) = b$, so that we re-write this inequality as

$$I(\mathbb{E}f) \le \mathbb{E}\sqrt{(I(f))^2 + |\nabla f|^2} \qquad (\dagger)$$

Here the expected value is taken with respect to the Bernoulli measure $\mu := (1/2)\delta_{-1} + (1/2)\delta_1$, and $|\nabla f| = |f(1) - f(-1)|/2$. To get $(**)$ from $(\dagger)$, we apply the Central Limit Theorem.

We first show that $(\dagger)$ tensorizes. That is, $(\dagger)$ holds for $f\colon \{-1,1\}^n \to [0,1]$ with measure μ^n on $\{-1,1\}^n$, and $|\nabla f(x)|^2 := \frac{1}{4}\sum_{j=1}^n |f(x) - f(s_j(x))|^2$, with $s_j(x) := (x_1, \ldots, x_{j-1}, -x_j, x_{j+1}, \ldots, x_n)$. Let $x \in \{-1,1\}^n$, and now let $f\colon \{-1,1\}^{n+1} \to [0,1]$. Let $f_0(x) := f(x,-1)$, $f_1(x) := f(x,1)$, $p_0 := \mu(\{-1\})$, $p_1 := \mu(\{1\})$, $a_0 := \int f_0 d\mu^n$, $a_1 := \int f_1 d\mu^n$. Then $\int f d\mu^{n+1} = p_0 a_0 + p_1 a_1$. Also,

$$|\nabla f(x,-1)|^2 = |\nabla f_0(x)|^2 + (1/4)\,|f_0(x) - f_1(x)|^2$$

$$|\nabla f(x,1)|^2 = |\nabla f_1(x)|^2 + (1/4)\,|f_0(x) - f_1(x)|^2$$

$$\begin{aligned}\int \sqrt{I(f)^2 + |\nabla f|^2} d\mu^{n+1} = p_0 &\int \sqrt{I(f_0)^2 + |\nabla f_0|^2 + \frac{1}{4}|f_0 - f_1|^2} d\mu^n \\ &+ p_1 \int \sqrt{I(f_1)^2 + |\nabla f_1|^2 + \frac{1}{4}|f_0 - f_1|^2} d\mu^n \quad (\ddagger)\end{aligned}$$

By applying (the proof of) Jensen's inequality twice, we have

$$\sqrt{(\int u)^2 + (\int v)^2} \le \int \sqrt{u^2 + (\int v)^2} \le \int \sqrt{u^2 + v^2} \qquad (***)$$

Specifically, we use that $\phi(x,y) := \sqrt{x^2 + y^2}$ is convex in one argument, when the other argument is fixed. Now, let $\ell_v(u) = au + b$ be a linear function such that $\ell_v(u) \le \phi(u, \int v)$ and $\ell_v(\int u) = \phi(\int u, \int v)$. Then

$$\phi(\int u, \int v) = \ell_v(\int u) = a\int u + b = \int \ell_v(u) \le \int \phi(u, \int v)$$

Now, let $\ell_u(v) = av + b$ be a linear function such that $\ell_u(v) \le \phi(u,v)$ and $\ell_u(\int v) = \phi(u, \int v)$. Then

$$\phi(u, \int v) = \ell_u(\int v) = a\int v + b = \int \ell_u(v) \le \int \phi(u,v)$$

We now use $(***)$ twice, with $u_0 = \sqrt{I(f_0)^2 + |\nabla f_0|^2}$, $v = (f_0 - f_1)/2$, and $u_1 = \sqrt{I(f_1)^2 + |\nabla f_1|^2}$. Then (‡) implies

$$\int \sqrt{I(f)^2 + |\nabla f|^2} d\mu^{n+1} \geq p_0 \sqrt{(\int u_0 d\mu^n)^2 + (\int v d\mu^n)^2}$$
$$+ p_1 \sqrt{(\int u_1 d\mu^n)^2 + (\int v d\mu^n)^2}$$

By the inductive assumption of (†), $\int u_0 d\mu^n \geq I(a_0)$, $\int u_1 d\mu^n \geq I(a_1)$. By definition of v, $\int v d\mu^n = (a_0 - a_1)/2$. Therefore,

$$\int \sqrt{I(f)^2 + |\nabla f|^2} d\mu^{n+1} \geq p_0 \sqrt{I(a_0)^2 + (1/4)(a_0 - a_1)^2}$$
$$+ p_1 \sqrt{I(a_1)^2 + (1/4)(a_0 - a_1)^2}$$

By the $n = 1$ case of (†),

$$p_0 \sqrt{I(a_0)^2 + (1/4)(a_0 - a_1)^2} + p_1 \sqrt{I(a_1)^2 + (1/4)(a_0 - a_1)^2}$$
$$\geq I(p_0 a_0 + p_1 a_1) = I(\int f d\mu^{n+1})$$

So, the inductive step is complete, and (†) is proven.

We now show how to get Gaussian isoperimetry from (†). Let $f \colon \mathbb{R}^n \to [0,1]$ be a smooth function. Let $x_1, \ldots, x_k \in \{-1,1\}^n$, and define $f_k(x_1, \ldots, x_k) := f((x_1 + \cdots + x_k)/\sqrt{k})$. By the Central Limit Theorem, as $k \to \infty$,

$$\int_{\{-1,1\}^{nk}} f_k d\mu^{nk} \to \int_{\mathbb{R}^n} f d\gamma_n$$

Also,

$$|\nabla f_k(x_1, \ldots, x_k)|^2 = \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{k} |f_k(x_1, \ldots, x_k) - f_k(x_1, \ldots, s_i(x_j), \ldots, x_k)|^2$$
$$= \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{k} \left| f((x_1 + \cdots + x_k)/\sqrt{k}) - f((x_1 + \cdots + s_i(x_j) + \cdots + x_k)/\sqrt{k}) \right|^2$$
$$= \|\nabla f((x_1 + \cdots + x_k)/\sqrt{k})\|_2^2 + O(1/\sqrt{k})$$

Since f is smooth, the error $O(1/\sqrt{k})$ is uniform over all $(x_1, \ldots, x_k)$. So, using the Central Limit Theorem again, as $k \to \infty$,

$$\int_{\{-1,1\}^{nk}} \sqrt{I(f_k)^2 + |\nabla f_k|^2} d\mu^{nk} \to \int_{\mathbb{R}^n} \sqrt{I(f)^2 + |\nabla f|^2} d\gamma_n$$

So, we have the analogue of (†) for the Gaussian measure. To finally get Gaussian isoperimetry, let f approximate 1_A for a set A, and note that $I(0) = I(1) = 0$. □

Theorem 11.9. *(**Gaussian Brunn-Minkowski**, [Bor03]) As above, let $\Phi(t) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}dx$, let $\gamma = \gamma_n$ be the standard Gaussian measure on $\mathbb{R}^n$, let $A, B \subseteq \mathbb{R}^n$ be Borel sets, and let $\lambda \in (0,1)$. Then*

$$\Phi^{-1}(\gamma_n(\lambda A + (1-\lambda)B)) \geq \lambda\Phi^{-1}(\gamma_n(A)) + (1-\lambda)\Phi^{-1}(\gamma_n(B))$$

(Compare to Euclidean Brunn-Minkowski, Thm. 11.2.)

Proof. We use a method of heat flows, together with the maximum principle. Let $u = T_t f$ be a solution of the heat equation $\partial u/\partial t = \frac{1}{2}\Delta u$. Suppose u has range $[0,1]$. Note that $d\Phi/dt = \phi := e^{-t^2/2}/\sqrt{2\pi}$. Define $U := \Phi^{-1}(u)$. Straightforward calculations show that

$$\frac{\partial u}{\partial t} = \phi(U)\frac{\partial U}{\partial t}, \qquad \nabla u = \Phi(U)\nabla U, \qquad \Delta u = \phi(U)(\nabla U - U\left|\nabla U\right|^2)$$

$$\frac{\partial U}{\partial t} = \frac{1}{2}\Delta U - \frac{1}{2}U\left|\nabla U\right|^2 \qquad (*)$$

Let $A, B \subseteq \mathbb{R}^n$ be compact subsets, and let $\varepsilon \in (0,1)$, $\delta \in (0,\varepsilon)$. Let F be a smooth bump function adapted to an epsilon neighborhood of A. Let $f := \delta + (1-\varepsilon)F$, so that $\delta \leq f \leq \alpha := \delta + 1 - \varepsilon < 1$. Similarly define g for B. Define

$$\kappa := \max\{\Phi(\lambda\Phi^{-1}(\alpha) + (1-\lambda)\Phi^{-1}(\delta)), \Phi(\lambda\Phi^{-1}(\delta) + (1-\lambda)\Phi^{-1}(\alpha))\}$$

Crucially, $\kappa \to 0$ as $\delta \to 0$ (since $\Phi^{-1}(0) = -\infty$). Now, let h such that $\kappa \leq h \leq \alpha$, such that $h = \alpha$ on $\lambda A_\varepsilon + (1-\lambda)B_\varepsilon$ and $h = \kappa$ off an ε-neighborhood of $\lambda A_\varepsilon + (1-\lambda)B_\varepsilon$. Observe that the following inequality holds pointwise for $x, y \in \mathbb{R}^n$.

$$\Phi^{-1}(h(\lambda x + (1-\lambda)y)) \geq \lambda\Phi^{-1}(f(x)) + (1-\lambda)\Phi^{-1}(g(y)) \qquad (**)$$

This inequality is clear for $x \in A, y \in B$ and for $x \notin A, y \notin B$, since we reduce to $\alpha \geq \alpha$ or $\kappa \geq \delta$. For $x \in A, y \notin B$, the inequality holds by definition of κ. (And similarly for $x \notin A, y \in B$).

Now, by letting $\delta \to 0$ and then $\varepsilon \to 0$ in the following inequality, we will complete the theorem.

$$\Phi^{-1}(\int hd\gamma) \geq \lambda\Phi^{-1}(\int fd\gamma) + (1-\lambda)\Phi^{-1}(\int gd\gamma) \qquad (\ddagger)$$

We prove ($\ddagger$) via the more general inequality

$$\Phi^{-1}(T_t h(\lambda x + (1-\lambda)y)) \geq \lambda\Phi^{-1}(T_t f(x)) + (1-\lambda)\Phi^{-1}(T_t g(y)) \qquad (\clubsuit)$$

where T_t is the heat semigroup operator.

Surprisingly, the inequality ($\clubsuit$) is almost a symbolic exercise. Define

$$C(t,x,y) := \Phi^{-1}(T_t h(\lambda x + (1-\lambda)y)) - \lambda\Phi^{-1}(T_t f(x)) + (1-\lambda)\Phi^{-1}(T_t g(y))$$

Then $(**)$ expresses $C(0,x,y) \geq 0$ and our desired inequality ($\ddagger$) corresponds to $C(1,0,0) \geq 0$. Moreover, by construction, C vanishes as $\left|x\right|, \left|y\right| \to \infty$. Since we know $C(0,x,y) \geq 0$, in order to prove ($\clubsuit$) it suffices to show via a maximum principle that $C(t,x,y) \geq 0$ for all $x, y \in \mathbb{R}^n$, $t \geq 0$.

Let $\xi := (t,x)$, $\eta := (t,y)$ and $\zeta := (t, \lambda x + (1-\lambda y))$. Then

$$
\begin{aligned}
\nabla_x C &= \lambda[(\nabla(\Phi^{-1}(T_t h)))(\zeta) - (\nabla(\Phi^{-1}(T_t f)))(\xi)] \\
\nabla_y C &= (1-\lambda)[(\nabla(\Phi^{-1}(T_t h)))(\zeta) - (\nabla(\Phi^{-1}(T_t g)))(\eta)] \\
\Delta_x C &= \lambda^2(\Delta(\Phi^{-1}(T_t h)))(\zeta) - \lambda(\Delta\Phi^{-1}(T_t f))(\xi) \\
\Delta_y C &= (1-\lambda)^2(\Delta(\Phi^{-1}(T_t h)))(\zeta) - (1-\lambda)(\Delta\Phi^{-1}(T_t g))(\eta)
\end{aligned}
$$

Also,

$$\sum_{1\le i\le n} \frac{\partial^2 C}{\partial x_i \partial y_i} = \lambda(1-\lambda)(\Delta(\Phi^{-1}(T_t h)))(\zeta)$$

Let $\mathcal{E} := \frac{1}{2}\left[\Delta_x + 2\sum_{1\le i\le n}\frac{\partial^2}{\partial x_i \partial y_i} + \Delta_y\right]$. Then, using $(*)$,

$$
\begin{aligned}
\mathcal{E}C &= \frac{1}{2}\left[(\Delta(\Phi^{-1}(T_t h)))(\zeta) - \lambda(\Delta(\Phi^{-1}(T_t f)))(\xi) - (1-\lambda)(\Delta(\Phi^{-1}(T_t g)))(\eta)\right] \\
&= \frac{\partial}{\partial t}(\Phi^{-1}(T_t h))(\zeta) + \frac{1}{2}(\Phi^{-1}(T_t h))(\zeta)\left|(\nabla(\Phi^{-1}(T_t h)))(\zeta)\right|^2 \\
&\quad - \lambda\frac{\partial}{\partial t}(\Phi^{-1}(T_t f))(\xi) - \frac{\lambda}{2}(\Phi^{-1}(T_t f))(\xi)\left|(\nabla(\Phi^{-1}(T_t f)))(\xi)\right|^2 \\
&\quad - (1-\lambda)\frac{\partial}{\partial t}(\Phi^{-1}(T_t g))(\eta) - \frac{1-\lambda}{2}(\Phi^{-1}(T_t g))(\eta)\left|(\nabla(\Phi^{-1}(T_t g)))(\eta)\right|^2 \\
&=: \frac{\partial C}{\partial t} + \Psi(t,x,y)
\end{aligned}
$$

Now, write

$$
\begin{aligned}
&\left|(\nabla(\Phi^{-1}(T_t f)))(\xi)\right|^2 = \left|(\nabla(\Phi^{-1}(T_t h)))(\zeta)\right|^2 \\
&\quad + \sum_{1\le i\le n}\left[\frac{\partial\Phi^{-1}(T_t f)}{\partial x_i}(\xi) + \frac{\partial\Phi^{-1}(T_t h)}{\partial x_i}(\zeta)\right]\left[\frac{\partial\Phi^{-1}(T_t f)}{\partial x_i}(\xi) - \frac{\partial\Phi^{-1}(T_t h)}{\partial x_i}(\zeta)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\left|(\nabla(\Phi^{-1}(T_t g)))(\eta)\right|^2 = \left|(\nabla(\Phi^{-1}(T_t h)))(\zeta)\right|^2 \\
&\quad + \sum_{1\le i\le n}\left[\frac{\partial\Phi^{-1}(T_t g)}{\partial x_i}(\eta) + \frac{\partial\Phi^{-1}(T_t h)}{\partial x_i}(\zeta)\right]\left[\frac{\partial\Phi^{-1}(T_t g)}{\partial x_i}(\eta) - \frac{\partial\Phi^{-1}(T_t h)}{\partial x_i}(\zeta)\right]
\end{aligned}
$$

Then by these equalities and the definition of Ψ, we can write $\Psi(t,x,y) =: \frac{1}{2}\left|(\nabla(\Phi^{-1}(T_t h)))(\zeta)\right|^2 C - b(t,x,y)\cdot\nabla_{(x,y)}C$. In summary,

$$\mathcal{E}C + b(t,x,y)\cdot\nabla_{(x,y)}C = \frac{\partial C}{\partial t} + \frac{1}{2}\left|(\nabla(\Phi^{-1}(T_t h)))(\zeta)\right|^2 C \qquad (***)$$

Note that $\mathcal{E}$ is elliptic, since it can be written as $\mathcal{E} = \frac{1}{2}(\nabla_x,\nabla_y)^T A(\nabla_x,\nabla_y)$, where A is a block matrix of the form $\begin{pmatrix} id & id \\ id & id \end{pmatrix}$, and each id is an $n\times n$ identity matrix.

We now argue by contradiction. By the definitions of f,g,h, we know that $\inf_{0\le t\le T} C(t,x,y)$ is non-negative as $|x|+|y|\to\infty$. We make the contrary assumption that $C(t,x,y) < 0$ for some $(t,x,y)\in[0,T]\times\mathbb{R}^n\times\mathbb{R}^n$. Then there exists $\varepsilon > 0$ such that $\varepsilon t + C(t,x,y)$ has a strictly negative minimum in $[0,T]\times\mathbb{R}^n\times\mathbb{R}^n$. Suppose this minimum occurs at $P = (t_0,x_0,y_0)$

with $t_0 > 0$. Note that $t_0 > 0$, since $t_0 = 0$ cannot occur, since ($\clubsuit$) reduces to ($**$) at $t = 0$. Then, using that P is a minimum and $\mathcal{E}$ is an elliptic operator,

$$C(P) < 0, \quad \frac{\partial C}{\partial t}(P) + \varepsilon \leq 0, \quad \nabla_{(x,y)} C(P) = 0, \quad \mathcal{E}C(P) \geq 0$$

However, these inequalities contradict ($***$). We conclude that $C(t, x, y) \geq 0$, so ($\clubsuit$) holds, as desired. □

We begin with Gross' Log-Sobolev inequality. Let μ be a probability measure on some measure space Ω. Given a non-negative measurable function f on Ω with $\int f \log(1+f) < \infty$, define the **entropy** of f as

$$\text{Ent}_\mu(f) := \int f \log f d\mu - (\int f d\mu) \cdot \log(\int f d\mu) = \int f \log \left(\frac{f}{\int f d\mu} \right) d\mu$$

Note that $\text{Ent}_\mu(f) \geq 0$ by Jensen's inequality (since $g(x) = x \log x$, $x \geq 0$ is convex). Also, a short calculation shows that $\text{Ent}_\mu(\lambda f) = \lambda \text{Ent}_\mu(f)$, i.e. the entropy is homogeneous of order 1.

Theorem 11.10. *(**Log-Sobolev Inequality**, Gross, [Led01]) Let $f: \mathbb{R}^n \to \mathbb{R}$ be smooth. As usual, let $\gamma := (1/(2\pi)^{n/2}) e^{-\|x\|_2^2/2} dx$ be the standard Gaussian measure on $\mathbb{R}^n$, and let $|\nabla f| := \|\nabla f\|_2$. Then*

$$Ent_\gamma(f^2) \leq 2 \int |\nabla f|^2 d\gamma$$

Proof. Recall that that the operator $L := \Delta - x \cdot \nabla$ on $\mathbb{R}^n$ has the associated semigroup $\{P_t\}_{t \geq 0}$ with representation

$$P_t f(x) = \int f(e^{-t} x + (1 - e^{-2t})^{1/2} y) d\gamma(y), \quad t \geq 0, x \in \mathbb{R}^n$$

This semigroup is known as the Ornstein-Uhlenbeck semigroup. Now, let f be smooth and non-negative with $\varepsilon \leq f \leq 1/\varepsilon$ for some $\varepsilon > 0$. Via abstract semigroup theory or direct verification with the above formula for P_t, we have: $P_t = e^{tL}$, $(\partial/\partial t) P_t f = L P_t f$, $P_0 f = f$, $P_\infty f := \lim_{t \to \infty} P_t f = \int f d\gamma$. Also, recall the Gaussian integration by parts formula $\int f(-Lg) d\gamma = \int \nabla f \cdot \nabla g d\gamma$, via real variable integration by parts.

Lastly, we need the formula $\int P_t f(x) d\gamma(x) = \int f d\gamma$. This follows immediately from the fact that, if X, Y are independent standard Gaussians, then $\alpha X + \sqrt{1 - \alpha^2} Y$ is also a standard Gaussian. Thus,

$$\int P_t f(x) d\gamma(x) = \mathbb{E} f(e^{-t} X + \sqrt{1 - e^{-2t}} Y) = \mathbb{E} f(X') = \int f d\gamma$$

where X' is another standard Gaussian. Alternately,

$$\frac{\partial}{\partial t} \int P_t f = \int \frac{\partial}{\partial t} P_t f = \int L P_t f = -\int P_t f \cdot (\nabla 1) = 0$$

In either case, we are ready to begin the proof.

$$
\begin{aligned}
\mathrm{Ent}_\gamma(f) &= \int f\log f - (\int f)\log(\int f) \\
&= \int (P_0 f\log P_0 f) - P_\infty f\log P_\infty f \\
&= \int_{\mathbb{R}^n} (-\int_0^\infty \frac{\partial}{\partial t}(P_t \log P_t)dt)d\gamma \quad \text{, integrating by parts} \\
&= -\int_0^\infty \int_{\mathbb{R}^n} \frac{\partial}{\partial t} P_t \log P_t d\gamma dt \\
&= -\int_0^\infty \left(\int LP_t f\log P_t f d\gamma + \int LP_t f d\gamma \right) dt \quad \text{, chain rule} \\
&= -\int_0^\infty \left(-\int \nabla P_t f\cdot \nabla \log P_t f) - \int \nabla P_t f \cdot (\nabla 1) \right) dt \quad \text{, } L \text{ int. by parts} \\
&= \int_0^\infty \int \frac{|\nabla P_t f|^2}{P_t f} d\gamma dt
\end{aligned}
$$

Now, from the formula for P_t, we see that $\nabla P_t f = e^{-t}P_t(\nabla f)$. So, using the formula for P_t and Minkowski's inequality we have $|\nabla P_t f| \leq e^{-t}P_t(|\nabla f|)$. Note that we may apply Cauchy-Schwarz to $P_t(fg)(x)$ for fixed x on $L_2(\gamma)$, so that $P_t(fg) \leq \sqrt{P_t(f^2)}\sqrt{P_t(g^2)}$. Therefore

$$|\nabla P_t f| \leq e^{-t}P_t(|\nabla f|) = e^{-t}P_t(|\nabla f|\sqrt{f}/\sqrt{f}) \leq e^{-t}\sqrt{P_t f}\sqrt{P_t(|\nabla f^2|/f)}$$

that is,

$$\frac{|\nabla P_t f|^2}{P_t f} \leq e^{-2t} P_t\left(\frac{|\nabla f|^2}{f}\right)$$

So, combining this with our above equality for entropy gives

$$\mathrm{Ent}_\gamma(f) \leq \int_0^\infty e^{-2t}\int P_t \left(\frac{|\nabla f|^2}{f}\right) d\gamma dt = \frac{1}{2}\int \frac{|\nabla f|^2}{f} d\gamma$$

using that $\int P_t g d\gamma = \int g d\gamma$, and then integrating in t. Letting $f = g^2$ and applying the chain rule completes the proof. □

Remark 11.11. The above applies equally well to measures $d\mu = e^{-U}dx$ where $\mathrm{Hess}U(x) \geq cId$. The result is then

$$\mathrm{Ent}_\mu(f^2) \leq \frac{2}{c}\int |\nabla f|^2 d\mu$$

12. Appendix: Notation

Let n, m be a positive integers. Let A, B be sets contained in a universal set Ω.

$\mathbb{N} = \{1, 2, \ldots\}$ denotes the set of natural numbers

$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ denotes the set of integers

$\mathbb{Q} = \{a/b\colon a, b, \in \mathbb{Z}, b \neq 0\}$ denotes the set of rational numbers

$\mathbb{R}$ denotes the set of real numbers

$\mathbb{C} = \{a + b\sqrt{-1}\colon a, b \in \mathbb{R}\}$ denotes the set of complex numbers

$\in$ means "is an element of." For example, $2 \in \mathbb{R}$ is read as "2 is an element of $\mathbb{R}$."

$\forall$ means "for all"

$\exists$ means "there exists"

$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n)\colon x_i \in \mathbb{R}\, \forall\, 1 \leq i \leq n\}$

$f\colon A \to B$ means f is a function with domain A and range B. For example,

$f\colon \mathbb{R}^2 \to \mathbb{R}$ means that f is a function with domain $\mathbb{R}^2$ and range $\mathbb{R}$

$\emptyset$ denotes the empty set

$A \subseteq B$ means $\forall\, a \in A$, we have $a \in B$, so A is contained in B

$A \smallsetminus B := \{a \in A\colon a \notin B\}$

$A^c := \Omega \smallsetminus A$, the complement of A in Ω

$A \cap B$ denotes the intersection of A and B

$A \cup B$ denotes the union of A and B

$A \Delta B := (A \smallsetminus B) \cup (B \smallsetminus A)$

$\mathbb{P}$ denotes a probability law on Ω

Let $n \geq m \geq 0$ be integers. We define

$$\binom{n}{m} := \frac{n!}{(n-m)!m!} = \frac{n(n-1)\cdots(n-m+1)}{m(m-1)\cdots(2)(1)}.$$

Let $a_1, \ldots, a_n$ be real numbers. Let n be a positive integer.

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_{n-1} + a_n.$$

$$\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdots a_{n-1} \cdot a_n.$$

$\min(a_1, a_2)$ denotes the minimum of a_1 and a_2.

$\max(a_1, a_2)$ denotes the maximum of a_1 and a_2.

The min of a set of nonnegative real numbers is the smallest element of that set. We also define $\min(\emptyset) := \infty$.

Let $A \subseteq \mathbb{R}$.

$$\sup A \text{ denotes the supremum of } A\text{, i.e. the least upper bound of } A.$$
$$\inf A \text{ denotes the infimum of } A\text{, i.e. the greatest lower bound of } A.$$

Let $X\colon \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mu)$.

$$\begin{aligned} \mathbb{E}(X) &\text{ denotes the expected value of } X \\ \|X\|_p &:= (\mathbb{E}\,|X|^p)^{1/p}, \text{ denotes the } L_p\text{-norm of } X \text{ when } 1 \leq p < \infty \\ \|X\|_\infty &:= \inf\{c > 0\colon \mathbb{P}(|X| \leq c) = 1\}, \text{ denotes the } L_\infty\text{-norm of } X \\ \mathrm{var}(X) &= \mathbb{E}(X - \mathbb{E}(X))^2, \text{ the variance of } X \\ \sigma_X &= \sqrt{\mathrm{var}(X)}, \text{ the standard deviation of } X \end{aligned}$$

Let $A \subseteq \Omega$.

$\mathbb{E}(X|A) := \mathbb{E}(X1_A)/\mathbb{P}(A)$ denotes the expected value of X conditioned on the event A.

$$1_A\colon \Omega \to \{0,1\}, \text{ denotes the indicator function of } A\text{, so that}$$
$$1_A(\omega) = \begin{cases} 1 & \text{, if } \omega \in A \\ 0 & \text{, otherwise.} \end{cases}$$

Let $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the standard inner product on $\mathbb{R}^n$, so that for any $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, we have $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$. We also denote $\|x\| := (\sum_{i=1}^n x_i^2)^{1/2}$ as the standard norm on $\mathbb{R}^n$.

Let X be a random variable on a sample space Ω, so that $X\colon \Omega \to \mathbb{R}$. Let $\mathbb{P}$ be a probability law on Ω. Let $x, t \in \mathbb{R}$.

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega\colon X(\omega) \leq x\}) \\ &\qquad \text{the Cumulative Distibution Function of } X. \\ M_X(t) &= \mathbb{E}e^{tX} \text{ denotes the Moment Generating Function of } X \text{ at } t \in \mathbb{R} \end{aligned}$$

Let $g, h\colon \mathbb{R} \to \mathbb{R}$. Let $t \in \mathbb{R}$.

$$(g * h)(t) = \int_{-\infty}^{\infty} g(x)h(t-x)dx \text{ denotes the convolution of } g \text{ and } h \text{ at } t \in \mathbb{R}$$

Let $f, g\colon \mathbb{R} \to \mathbb{C}$. We use the notation $f(t) = o(g(t))$, $\forall\, t \in \mathbb{R}$ to denote $\lim_{t\to\infty} \left|\frac{f(t)}{g(t)}\right| = 0$. We use the notation $f(t) = O(g(t))$ to denote that $\exists\, c > 0$ such that $|f(t)| \leq c\,|g(t)|$ for all $t \in \mathbb{R}$. We write $f(t) = \Omega(g(t))$ when $\exists\, c > 0$ such that $|f(t)| \geq c\,|g(t)|$ for all $t \in \mathbb{R}$. We write $f(t) = \Theta(g(t))$ when $f(t) = O(g(t))$ and $g(t) = O(f(t))$.

References

[Aar11] Scott Aaronson, *A linear-optical proof that the permanent is #p-hard*, Electronic Colloquium on Computational Complexity (ECCC) **18** (2011), 43.

[ABSS97] Sanjeev Arora, László Babai, Jacques Stern, and Z. Sweedyk, *The hardness of approximate optima in lattices, codes, and systems of linear equations*, J. Comput. Syst. Sci. **54** (1997), no. 2, 317–331.

[ACKS15] Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop, *The hardness of approximation of euclidean k-means*, Preprint, arXiv:1502.03316, 2015.

[ANFSW0] Sara. Ahmadian, Ashkan. Norouzi-Fard, Ola. Svensson, and Justin. Ward, *Better guarantees for k-means and euclidean k-median by primal-dual algorithms*, SIAM Journal on Computing **0** (0), no. 0, FOCS17–97–FOCS17–156.

[Ant95] Martin Anthony, *Classification by polynomial surfaces*, Discrete Applied Mathematics **61** (1995), no. 2, 91 – 103.

[BFKV98] Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh Vempala, *A polynomial-time algorithm for learning noisy linear threshold functions*, Algorithmica **22** (1998), no. 1/2, 35–52.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities*, Oxford University Press, Oxford, 2013, A nonasymptotic theory of independence, With a foreword by Michel Ledoux. MR 3185193

[Bob97] S. G. Bobkov, *An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in gauss space*, Ann. Probab. **25** (1997), no. 1, 206–214.

[Bor03] Christer Borell, *The Ehrhard inequality*, C. R. Math. Acad. Sci. Paris **337** (2003), no. 10, 663–666. MR 2030108 (2004k:60102)

[Bou02] Olivier Bousquet, *A Bennett concentration inequality and its application to suprema of empirical processes*, C. R. Math. Acad. Sci. Paris **334** (2002), no. 6, 495–500. MR 1890640

[CAS19] Vincent Cohen-Addad and Karthik Srikanta, *Inapproximability of Clustering in Lp-metrics*, FOCS'19 - 60th Annual IEEE Symposium on Foundations of Computer Science (Baltimore, United States), November 2019.

[CEM$^+$15] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu, *Dimensionality reduction for k-means clustering and low rank approximation*, Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '15, ACM, 2015, pp. 163–172.

[Cha05] Sourav Chatterjee, *An error bound in the Sudakov-Fernique inequality*, Preprint, arXiv.0510424, 2005.

[Che09] Ke Chen, *On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications*, SIAM Journal on Computing **39** (2009), no. 3, 923–947.

[CW09] Kenneth L. Clarkson and David P. Woodruff, *Numerical linear algebra in the streaming model*, Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '09, ACM, 2009, pp. 205–214.

[DKS18] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart, *Learning geometric concepts with nasty noise*, Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018, 2018, pp. 1061–1073.

[ES16] Ronen Eldan and Ohad Shamir, *The power of depth for feedforward neural networks*, Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016, 2016, pp. 907–940.

[FGKP06] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami, *New results for learning noisy parities and halfspaces*, 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings, 2006, pp. 563–574.

[FGRW12] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu, *Agnostic learning of monomials by halfspaces is hard*, SIAM J. Comput. **41** (2012), no. 6, 1558–1590. MR 3029261

[FMS07] Dan Feldman, Morteza Monemizadeh, and Christian Sohler, *A ptas for k-means clustering based on weak coresets*, Proceedings of the Twenty-third Annual Symposium on Computational Geometry (New York, NY, USA), SCG '07, ACM, 2007, pp. 11–18.

[FS97] Yoav Freund and Robert E Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, J. Comput. Syst. Sci. **55** (1997), no. 1, 119–139.

[Gal14] François Le Gall, *Powers of tensors and fast matrix multiplication*, Preprint, arXiv:1401.7714. ISAAC 2014., 2014.

[GOR$^+$21] Fabrizio Grandoni, Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Rakesh Venkat, *A refined approximation for Euclidean k-means*, preprint, arXiv:2107.07358, 2021.

[HMT11] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Review **53** (2011), no. 2, 217–288.

[HPM04] Sariel Har-Peled and Soham Mazumdar, *On coresets for k-means and k-median clustering*, Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '04, ACM, 2004, pp. 291–300.

[HRST17] Johan Håstad, Benjamin Rossman, Rocco A. Servedio, and Li-Yang Tan, *An average-case depth hierarchy theorem for boolean circuits*, J. ACM **64** (2017), no. 5, 35:1–35:27.

[JP78] David S. Johnson and Franco P. Preparata, *The densest hemisphere problem*, Theor. Comput. Sci. **6** (1978), 93–107.

[JSV04] Mark Jerrum, Alistair Sinclair, and Eric Vigoda, *A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries*, J. ACM **51** (2004), no. 4, 671–697.

[KKMS08] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio, *Agnostically learning halfspaces*, SIAM J. Comput. **37** (2008), no. 6, 1777–1805.

[KKP17] Daniel Kane, Sushrut Karmalkar, and Eric Price, *Robust polynomial regression up to the information theoretic limit*, 58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017, 2017, pp. 391–402.

[KMN$^+$04] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, *A local search approximation algorithm for k-means clustering*, Comput. Geom. **28** (2004), no. 2-3, 89–112. MR 2062789 (2005a:68210)

[KR08] Subhash Khot and Oded Regev, *Vertex cover might be hard to approximate to within 2-ε*, J. Comput. Syst. Sci. **74** (2008), no. 3, 335–349.

[KV94] Michael J. Kearns and Umesh V. Vazirani, *An introduction to computational learning theory*, MIT Press, Cambridge, MA, USA, 1994.

[KV09] Ravi Kannan and Santosh Vempala, *Spectral algorithms*, Foundations and Trends in Theoretical Computer Science **4** (2009), no. 3-4, 157–288.

[KW16] Daniel M. Kane and Ryan Williams, *Super-linear gate and super-quadratic wire lower bounds for depth-two and depth-three threshold circuits*, Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '16, ACM, 2016, pp. 633–643.

[Led96] Michel Ledoux, *Isoperimetry and Gaussian analysis*, Lectures on probability theory and statistics (Saint-Flour, 1994), Lecture Notes in Math., vol. 1648, Springer, Berlin, 1996, pp. 165–294. MR 1600888 (99h:60002)

[Led01] ———, *The concentration of measure phenomenon*, Mathematical Surveys and Monographs, vol. 89, American Mathematical Society, Providence, RI, 2001. MR 1849347 (2003k:28019)

[LOGT12] James R. Lee, Shayan Oveis Gharan, and Luca Trevisan, *Multi-way spectral partitioning and higher-order Cheeger inequalities*, STOC'12—Proceedings of the 2012 ACM Symposium on Theory of Computing, ACM, New York, 2012, pp. 1117–1130. MR 2961569

[Mah11] Michael W. Mahoney, *Randomized algorithms for matrices and data*, Found. Trends Mach. Learn. **3** (2011), no. 2, 123–224.

[Mat00] J. Matoušek, *On approximate geometric k-clustering*, Discrete Comput. Geom. **24** (2000), no. 1, 61–84. MR 1765234 (2001e:52036)

[MMR18] Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn, *Performance of johnson-lindenstrauss transform for k-means and k-medians clustering*, CoRR **abs/1811.03195** (2018).

[Sch90] Robert E. Schapire, *The strength of weak learnability*, Machine Learning **5** (1990), no. 2, 197–227.

[Tal96] Michel Talagrand, *New concentration inequalities in product spaces*, Invent. Math. **126** (1996), no. 3, 505–563. MR 1419006

[Tel15] Matus Telgarsky, *Representation benefits of deep feedforward networks*, CoRR **abs/1509.08101** (2015).

[Ver18] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2018.

USC Mathematics, Los Angeles, CA
E-mail address: `stevenmheilman@gmail.com`