A Tour Through the

"Mathematics of Signals, Networks, and Learning"

Lecture Notes for Spring 2025

Afonso S. Bandeira
bandeira@math.ethz.ch

Last update: June 4, 2025

In Preparation:

These notes are *in preparation* and being updated. This version was compiled on June 4, 2025.

Course overview

This is an introductory course to Mathematical aspects of Data Science, Machine Learning, Signal Processing, and Networks. One of the main goals of this course is to highlight the interaction between different areas of (pure) mathematics , in the context of problems motivated by Data Science and related topics. It is a "Theorem-Proof" based Mathematics course.

Course Coordinator: Petar Nizic-Nikolac

petar.nizic-nikolac@ifor.math.ethz.ch

The contents of the course will be updated as the semester progresses, and the following list is subject to many possible changes.

1. **Unsupervized Learning and Data Parsimony:**
   - Clustering and k-means
   - Singular Value Decomposition
   - Low Rank approximations and Eckart–Young–Mirsky Theorem
   - Dimension Reduction and Principal Component Analysis
   - Kernel PCA and Bochner's theorem
   - Sparsity and Compressed Sensing
   - Finite Frame Theory and Equiangular tight frames
   - The Paley ETF
   - Concentration inequalities and random low-coherence frames.
2. **Connections to graph theory**
   - Introduction to Spectral Graph Theory.
   - Paley ETF and the Paley graph
3. **Signal processing and Fourier analysis**
   - Shannon's sampling theorem and the Nyquist rate.

- Discrete Fourier transform

4. **Supervized Learning:**
    - Introduction to Classification and Generalization of Classifiers
    - Uniform convergence, the VC theorem and VC Dimension

Note for non-Mathematics students: this class requires a certain degree of mathematical maturity–including abstract thinking and the ability to understand and write proofs.

Please visit the Forum in the Moodle parge of the course for more information.

You will notice several questions along the way, separated into Challenges (and Exploratory Challenges).

- **Challenges** are well-defined mathematical questions, of varying level of difficulty. Some are very easy, and some are much harder than any homework problem.
- **Exploratory Challenges** are not necessarily well defined, but thinking about them should improve your understanding of the material.

We also include a few "Further Reading" references in case you are interested in learning more about a particular topic.

Several chapters of these lecture notes are a close adaptation of the ones of the previous years, both one by A. S. Bandeira and N. Zhivotovskiy [BZ22], and one by A. S. Bandeira and A. Maillard [BM22]. Some parts are adapted from the book [BSS23] with A. Singer and T. Strohmer. Thus, the credit for a significant part of this material goes to these collaborators of mine; as for any errors and typos, the credit is all mine. If you are looking for a more advanced course on this topic with lecture notes and many open problems, you can also read through [Ban16].

Besides the goal of serving as an introduction to the Mathematics of Data Science and related areas, the content selection also has the goal of illustrating interesting connections between different parts of Mathematics and some of their (a priori) surprising applications.

Important disclaimer – *This draft is in the making and subject to many future changes, adds and removals. Please excuse the lack of polishing and typos. If you find any typos or mistakes, please let us know! This draft was last updated on June 4, 2025.*

Contents

1 Introduction

We will study several areas of Signal Processing, Machine Learning and Analysis of Data, focusing on the mathematical aspects. We list below the areas we will consider (the list contains the subjects already covered, and will be updated as the course progresses).

- Unsupervised Learning: The most common instance in exploratory data analysis is when we receive data points without a priori known structure, think e.g. unlabeled images from a databased, genomes of a population, etc. The natural first question is to ask if we can learn the geometry of the data. Simple examples include: Does the data separate well into clusters? Does the data naturally live in a smaller dimensional space or manifold? Sometimes the dataset comes in the form of a network (or graph) and the same questions can be asked, an approach in this case is with Spectral Graph Theory which we will cover if time permits.

- Signal Processing: Often times, the data we observe come in a form of a signal $f(t)$, in which we can think of t as the time. After motivating Fourier analysis with Bochner's theorem in the previous part, we will understand how Fourier analysis allows to understand when we can uniquely reconstruct a signal from a discrete set of measurements, by proving Shannon's sampling theorem. We will then define the Discrete Fourier transform and overview some of its applications.

- Parsimony and Sparsity: Sometimes, the information/signal we are after has a particular structure. A common form of parsimony is sparsity in a particular linear dictionary, such as natural images in the Wavelet basis, or audio signals in the Fourier basis. We will present the basics of Compressed sensing of sparse vectors, and use it to motivate the construction of low-coherence frames.

- Finite frame theory: Motivated by the compressed sensing application above, we will introduce the notion of maximally low-coherence frames, or equiangular tight frames. Using elementary notions of number theory, we will present the construction of the Paley ETF, one of the few explicit constructions that exist for these objects.

- Supervised Learning: In this last part, we will introduce some basics of statistical learning theory, using the common problem of classification, i.e. learning an unknown classifier function from examples. As a textbook illustrative example, one can have in mind classifying correctly images of cats and dogs by generalizing from a finite sample of such images in which the label is given. We will introduce the notion of PAC learnability for finite classes of functions, and using tools of probability and concentration of measure, we will give guarantees on generalisation for possibly infinite classes of functions via the VC dimension.

2 Clustering and k-means

Clustering is one of the central tasks in machine learning. Given a set of data points, the purpose of clustering is to partition the data into a set of clusters where data points assigned to the same cluster correspond to similar data (for example, having small distance to each other if the points are in Euclidean space).

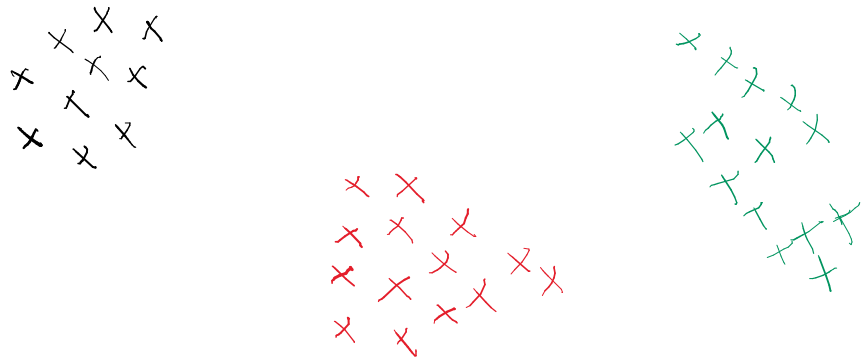

Figure 1: Examples of points which can be well separated in clusters.

k-means Clustering

One the most popular methods used for clustering is k-means clustering. Given $x_1, \ldots, x_n \in \mathbb{R}^p$, the k-means clustering partitions the data points in clusters $S_1, \ldots, S_k$ with centers $\mu_1, \ldots, \mu_k \in \mathbb{R}^p$ as the solution to:

$$\min_{\substack{\text{partition } S_1,\ldots,S_k \\ \mu_1,\ldots,\mu_k}} \sum_{l=1}^{k} \sum_{i \in S_l} \|x_i - \mu_l\|^2 . \tag{1}$$

A popular algorithm that attempts to minimize eq. (1) is Lloyd's Algorithm [Llo82] (this is also sometimes referred to as simply "the k-means algorithm"). It relies on the following two observations

Proposition 2.1 (*Properties of the k-means objective* –)

- Given a choice for the partition $S_1 \cup \cdots \cup S_k$, the centers that minimize (1) are given by

$$\mu_l = \frac{1}{|S_l|} \sum_{i \in S_l} x_i.$$

- Given the centers $\mu_1, \ldots, \mu_k \in \mathbb{R}^p$, the partition that minimizes (1) assigns each point x_i to the closest center μ_k.

Challenge 2.1. *Prove Proposition 2.1.*

We describe Lloyd's algorithm in Algorithm 2.1. Unfortunately, Lloyd's algorithm is not guaranteed to converge to the solution of (1). Indeed, it oftentimes gets stuck in local optima of (1). In fact, optimizing (1) is NP-hard and so there is no polynomial time algorithm that works in the worst-case (assuming the widely believed conjecture $P \neq NP$).

Algorithm 2.1 Lloyd's algorithm

It is an iterative algorithm that starts with an arbitrary choice of centers and iteratively alternates between

- Given centers $\mu_1, \ldots, \mu_k$, assign each point x_i to the cluster

$$l = \arg\min_{l=1,\ldots,k} \|x_i - \mu_l\| .$$

- Update the centers $\mu_l = \frac{1}{|S_l|} \sum_{i \in S_l} x_i$,

until no update is taken.

Challenge 2.2. *Show that Lloyd's algorithm 2.1 converges*[1] *(even if not always to the minimum of (1)).*

Challenge 2.3. *Can you find an example of points and starting centers for which Lloyd's algorithm does not converge to the optimal solution of (1)?*

Exploratory Challenge 2.4. *How would you try to "fix" Lloyd's Algorithm to avoid it getting stuck in the example you constructed in Challenge 2.3?*

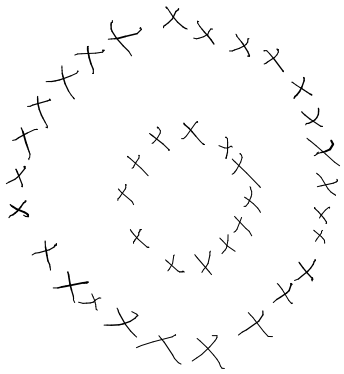

Figure 2: Because the solutions of k-means are always convex clusters, it is not able to handle some cluster structures.

While popular, k-means clustering has some potential issues:

- One needs to set the number of clusters a priori. A typical way to overcome this issue is to try the algorithm for different numbers of clusters.
- The way the formula (1) is defined needs the points to be defined in an Euclidean space. Often we are interested in clustering data for which we only have some measure of affinity between different data points, but not necessarily an embedding in $\mathbb{R}^p$ (this issue can be overcome by reformulating eq. (1) in terms of distances only — *you will do this on the first homework problem set.*).
- The formulation is computationally hard, so algorithms may produce suboptimal instances.

[1] In the sense that it stops after a finite number of iterations.

- The solutions of k-means are always convex clusters. This means that k-means cannot find clusters such as in Figure 2.

Further Reading 2.2. On the computational side, there are many interesting questions regarding when the k-means objective can be efficiently approximated, you can see a few open problems on this in [Ban16] (for example Open Problem 9.4).

3 The singular value decomposition

We recall here some useful facts and definitions on the singular value decomposition.

Data is often presented as a $d \times n$ matrix whose columns correspond to n data points in $\mathbb{R}^d$. Other examples include matrices of interactions where the entry (i, j) of a matrix contains information about an interaction, or similarity, between an item (or entity) i and j. In this context, the Singular Value Decomposition (SVD) is one of the most powerful tools to analyze matrix data.

Given a matrix $X \in \mathbb{R}^{n\times m}$, its Singular Value Decomposition is given by (U, Σ, V) such that

$$X = U\Sigma V^\top,$$

where $U \in O(n)$ and $V \in O(m)$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{n\times m}$ is a rectangular diagonal matrix, in the sense that $\Sigma_{ij} = 0$ for $i \neq j$, and whose diagonal entries are non-negative.

The diagonal entries $\sigma_1 \geq \sigma_2 \geq \ldots, \sigma_{\min\{n,m\}}$ of Σ are called the singular values[2] of X. Recall that unlike eigenvalues they must be real and non-negative. The columns $(u_k)_{k\in[n]}$ and $(v_\mu)_{\mu\in[m]}$ of respectively U and V are called the left and right singular vectors of X.

Proposition 3.1 (*Some basic properties of SVD*)

- $\mathrm{rk}(X)$ is equal to the number of non-zero singular values of X.
- If $n \leq m$, then the singular values of X are the square roots of the eigenvalues of $XX^\top$. If $m \leq n$ they are the square roots of the eigenvalues of $X^\top X$.

Challenge 3.1. *Prove this fact.*

The SVD can also be written in more economic ways. For example, if $\mathrm{rk}(X) = r \leq \min\{n, m\}$ then we can instead write

$$X = U\Sigma V^\top,$$

where $U^\top U = \mathrm{I}_{r\times r}$, $V^\top V = \mathrm{I}_{r\times r}$, and Σ is a non-singular $r \times r$ diagonal matrix matrix. Note that this representation only requires $r(n + m + 1)$ numbers, which if $r \ll \min\{n, m\}$ (i.e. if X is low-rank), is considerable savings when compared to the nm elements of X. It is also useful to write the SVD as

$$X = \sum_{k=1}^{r} \sigma_k u_k v_k^T,$$

where σ_k is the k-th largest singular value, and u_k and v_k are the corresponding left and right singular vectors.

[2]The most common convention is that the singular values are ordered in decreasing order, it is the convention we observe here.

4 Low-rank approximation of matrix data

A key observation in Machine Learning and Data Science is that (matrix) data is oftentimes well approximated by low-rank matrices. We will see examples of this phenomenon later in the course, and in the code simulations available on the class webpage.

In order to talk about what it means for a matrix B to approximate another matrix A, we need to have a notion of distance between matrices of the same dimensions, or equivalently a notion of norm of $A - B$. Let us start with some classical norms.

Definition 4.1 (*Spectral Norm*)

The spectral norm of $X \in \mathbb{R}^{n\times m}$ is given by

$$\|X\| \coloneqq \max_{\|v\|_2=1} \|Xv\|_2,$$

or equivalently $\|X\| \coloneqq \sigma_1(X)$.

Challenge 4.1. *Show that the two definitions above are equivalent.*

Another common matrix norm is the Frobenius (or Hilbert-Schmidt) norm.

Definition 4.2 (*Frobenius norm*)

The Frobenius norm of $X \in \mathbb{R}^{n\times m}$ is given by

$$\|X\|_F \coloneqq \left[\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}^2\right]^{1/2}.$$

Challenge 4.2. *Show that*

$$\|X\|_F^2 = \sum_{i=1}^{\min\{n,m\}} \sigma_i(X)^2 = \mathrm{Tr}[XX^\top].$$

Challenge 4.3. *Show that the spectral and Frobenius norms are indeed norms.*

Note that by solving Challenges 4.1 and 4.3 you have shown also that for any two matrices $X, Y \in \mathbb{R}^{n\times n}$,

$$\sigma_1(X+Y) \le \sigma_1(X) + \sigma_1(Y). \tag{2}$$

There is a natural generalization of the two norms above, the so called *Schatten p-norms*.

Definition 4.3 (*Schatten p-norm*)

Given a matrix $X \in \mathbb{R}^{n\times m}$ and $1 \le p \le \infty$, the Schatten p-norm of X is given by

$$\|X\|_{(S,p)} \coloneqq \left(\sum_{i=1}^{\min\{n,m\}} \sigma_i(X)^p\right)^{1/p} = \|\sigma(X)\|_p,$$

where $\sigma(X)$ corresponds to the vector whose entries are the singular values of X. Note that for $p = \infty$, this corresponds to the spectral norm and we often simply use $\|X\|$ without a subscript. Moreover, the Schatten 2-norm is the Frobenius norm, according to Challenge 4.2.

Challenge 4.4. *Show that the Schatten p-norm is a norm (proving triangular inequality for general $p \in [1, \infty]$ is non-trivial).*

Another key insight in this section is that, since the rank of a matrix X is the number of non-zero singular values, a natural rank-r approximation for a matrix X is to replace all singular values but the largest r singular values of X with zero. This is often referred to as the **truncated SVD**. Let us be more precise.

Definition 4.4 (*Truncated SVD*)

Let $X \in \mathbb{R}^{n\times m}$ and $X = U\Sigma V^\top$ be its SVD. We define $X_r = U_r\Sigma_r V_r^\top$ the truncated SVD of X by setting $U_r \in \mathbb{R}^{n\times r}$ and $V_r \in \mathbb{R}^{m\times r}$ to be, respectively, the first r columns of U and V; and $\Sigma_r \in \mathbb{R}^{r\times r}$ to be a diagonal matrix with the first r singular values of X (notice these are the largest ones, due to the way we defined SVD).

Warning: The notation X_r for low-rank approximations is not standard.

Note that $\mathrm{rk}(X_r) = r$ and $\sigma_1(X - X_r) = \sigma_{r+1}(X)$. It turns out that this way to approximate a matrix by a low-rank matrix is optimal is a very strong sense. This is captured by the celebrated Eckart–Young–Mirsky Theorem, which we will prove now, starting with a particular case.

Lemma 4.5 (*Eckart–Young–Mirsky Theorem for Spectral norm*)

The truncated SVD provides the best low-rank approximation in spectral norm. In other words, let $X \in \mathbb{R}^{n\times m}$ and $r \leq \min\{n, m\}$. Let X_r be as in Definition 4.4. Then for any matrix B with $\mathrm{rk}(B) \leq r$ we have:

$$\|X - B\| \geq \|X - X_r\|.$$

Proof of Lemma 4.5 – The claim with $r = \min\{m, n\}$ is trivial, as then $X_r = X$. We assume $r < \min(m, n)$. Let $X = U\Sigma V^\top$ be the SVD of X. Since $\mathrm{rk}(B) = r$ there must exist a vector $w \neq 0$ in the span of the first $r+1$ right singular vectors $v_1, \ldots, v_{r+1}$ of X such that w is in the kernel of B. Without loss of generality let w have unit norm. Let us write $w = \sum_{k=1}^{r+1} \alpha_k v_k$. Since w is unit-norm and the v_k's are orthonormal we have $\alpha_k = v_k^\top w$ and $\sum_{k=1}^{r+1} \alpha_k^2 = 1$. Finally,

$$\|X - B\| \geq \|(X - B)w\|_2 = \|Xw\|_2 = \|\Sigma V^\top w\|_2 = \sqrt{\sum_{k=1}^{r+1} \sigma_k^2(X)\alpha_k^2} \geq \sigma_{r+1}(X) = \|X - X_r\|.$$

□

Challenge 4.5. *If you think the existence of the vector w in the start of the proof above is not obvious (or any other step), try to prove it.*

The inequality (2) is a particular case of a more general set of inequalities, the Weyl inequalities, named after Hermann Weyl (a brilliant Mathematician who spent many years at ETH). Here we focus on the inequalities for singular values, the more classical ones are for eigenvalues; it is worth noting also that these follow from the ones for eigenvalues since the singular values of X are the square-roots of the eigenvalues of $X^\top X$.

Theorem 4.6 (*Weyl inequalities for singular values*)

For all $X, Y \in \mathbb{R}^{n\times m}$:

$$\sigma_{i+j-1}(X + Y) \leq \sigma_i(X) + \sigma_j(Y),$$

for all $1 \leq i, j, \leq \min\{n, m\}$ satisfying $i + j - 1 \leq \min\{n, m\}$

Proof of Theorem 4.6 – Let X_{i-1} and Y_{j-1} be, respectively, the rank $i-1$ and $j-1$ approximation of X and Y (as in Definition 4.4). By eq. (2) we have

$$\sigma_1\left((X - X_{i-1}) + (Y - Y_{j-1})\right) \leq \sigma_1\left(X - X_{i-1}\right) + \sigma\left(Y - Y_{j-1}\right) = \sigma_i(X) + \sigma_j(Y).$$

Since $X_{i-1} + Y_{j-1}$ has rank at most $i + j - 2$, Lemma 4.5 implies that

$$\sigma_{i+j-1}(X+Y) = \sigma_1(X+Y-(X+Y)_{i+j-2}) \leq \sigma_1\left(X+Y-(X_{i-1}+Y_{j-1})\right).$$

Putting both inequalities together we get

$$\sigma_{i+j-1}(X+Y) \leq \sigma_1\left(X+Y-X_{i-1}-Y_{j-1}\right) \leq \sigma_i(X) + \sigma_j(Y).$$

□

Challenge 4.6. *There is another simple proof of this theorem based on the Courant-Fischer minimax variational characterization of singular values:*

$$\sigma_k(X) = \max_{V \subseteq R^m, \dim(V)=k} \min_{v \in V, \|v\|=1} \|Xv\|, \tag{3}$$

$$\sigma_{k+1}(X) = \min_{V \subseteq R^m, \dim(V)=m-k} \max_{v \in V, \|v\|=1} \|Xv\|. \tag{4}$$

Try to prove it that way.

We are now ready to prove the main theorem of this section:

Theorem 4.7 (*Eckart–Young–Mirsky Theorem*)

The truncated SVD provides the best low-rank approximation in any Schatten p-norm. Formally, let $X \in \mathbb{R}^{n\times m}$, $r \leq \min\{n,m\}$, and $1 \leq p \leq \infty$. Let X_r be the truncated SVD of X retaining the leading r singular values, see Definition 4.4. Then

$$X_r = \underset{\substack{B\in\mathbb{R}^{n\times m}\\ \mathrm{rk}(B)\leq r}}{\arg\min} \|X-B\|_{(S,p)}.$$

We have already proved this for $p = \infty$ (Lemma 4.5). The proof of the general result follows from Weyl's inequalities (Theorem 4.6).

Proof of Theorem 4.7 – Let $X \in \mathbb{R}^{n\times m}$, and B a matrix with $\mathrm{rk}(B) \leq r$. We use Weyl's inequalities for $X - B$ and B:

$$\sigma_{i+j-1}(X) \leq \sigma_i(X-B) + \sigma_j(B),$$

Taking $j = r + 1$, and $i > 1$ satisfying $i + (r + 1) - 1 \leq \min\{n, m\}$ we have

$$\sigma_{i+r}(X) \leq \sigma_i(X-B), \tag{5}$$

since $\sigma_{r+1}(B) = 0$. Note that:

$$\|X-B\|^p_{(S,p)} = \sum_{k=1}^{\min\{n,m\}} \sigma_k^p(X-B) \geq \sum_{k=1}^{\min\{n,m\}-r} \sigma_k^p(X-B).$$

And by eq. (5):

$$\sum_{k=1}^{\min\{n,m\}-r} \sigma_k^p(X-B) \geq \sum_{k=1}^{\min\{n,m\}-r} \sigma_{k+r}^p(X) = \sum_{k=r+1}^{\min\{n,m\}} \sigma_k^p(X) = \|X-X_r\|^p_{(S,p)}.$$

□

5 Principal Component Analysis

When given some high-dimensional data, a statistician often seeks to find out if this data can be approximately represented as lying in a smaller dimensional set, see Fig. 3. In general, this procedure

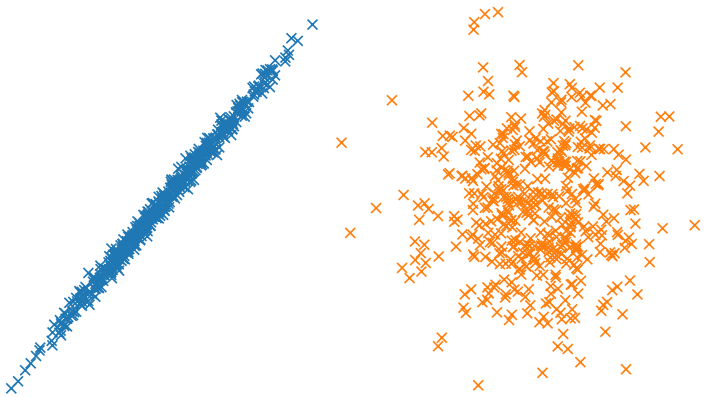

Figure 3: Two sets of data points. The blue points can clearly be well-approximated by a one-dimensional subspace (a line), while the orange points can not.

is referred to as *dimensionality reduction*: given a set of data points

$$y_1, \cdots, y_m \in \mathbb{R}^p, \tag{6}$$

we are hoping to find a good d-dimensional representation of $\{y_i\}_{i=1}^m$, ideally with $d \ll p$. The simplest such representation is given by a d-dimensional subspace: does there exists a set $z_1, \cdots, z_m$ of points which all lie on the same d-dimensional affine subspace, and such that z_i is "close" to y_i?

Let us simplify the setup slightly: we go from affine to linear subspace approximation by removing the empirical mean of $\{y_i\}_{i=1}^m$. More precisely, denoting $\mu \coloneqq (1/m)\sum_{i=1}^m y_i$, we will try to approximate $x_i \coloneqq y_i - \mu$ by points $\{z_i\}_{i=1}^m$ lying on a d-dimensional *linear* subspace.

To measure closeness of z_i to x_i, we will use the Euclidean norm. This leads us to look for

$$\underset{\{z_i\}_{i=1}^m}{\arg\min} \sum_{i=1}^m \|z_i - x_i\|_2^2, \tag{7}$$

in which the minimum is taken over all $\{z_i\}_{i=1}^m$ such that $\dim[\mathrm{Span}(\{z_i\})] \leq d$. Note that $\mathrm{Span}(\{z_i\})$ is also the image space of the matrix $Z \in \mathbb{R}^{p\times m}$, by defining

$$Z \coloneqq \begin{pmatrix} | & & | \\ z_1 & \cdots & z_m \\ | & & | \end{pmatrix} \quad \text{and} \quad X \coloneqq \begin{pmatrix} | & & | \\ x_1 & \cdots & x_m \\ | & & | \end{pmatrix}.$$

This allows to rewrite eq. (7) as:

$$\underset{\substack{Z\in\mathbb{R}^{p\times m} \\ \mathrm{rk}(Z)\leq d}}{\arg\min} \|Z - X\|_F^2. \tag{8}$$

We recognize exactly the setup of Theorem 4.7: the solution is given by the truncated SVD of X, that is

$$\underset{\substack{Z\in\mathbb{R}^{p\times m} \\ \mathrm{rk}(Z)\leq d}}{\arg\min} \|Z - X\|_F^2 = U_d \Sigma_d V_d^\top, \tag{9}$$

in which $X = U\Sigma V^\top$, and we used the notations of Definition 4.4. Coming back to our original task of approximating $\{y_i\}_{i=1}^m$, this means that the best d-dimensional representation is given by

$$y_i \sim z_i = U_d \beta_i + \mu, \tag{10}$$

where $\beta_i \coloneqq \Sigma_d v_i \in \mathbb{R}^d$, with v_i the i-th column of V_d. Eq. (10) is known as *Principal Component Analysis* (PCA). We refer to [BSS23] for an alternative derivation of PCA, not based on Eckart-Young's theorem.

Remark 5.1. Notice how in eq. (10) the left singular vectors U_d and the right singular vectors V_d have two different interpretations:

- The singular vectors U_d correspond to the basis in which to project the original points (after centering).
- The singular vectors V_d (or more precisely the vectors $\beta_i = \Sigma_d v_i$) then correspond to low dimensional coordinates for the points in this basis.

While centering the data points might seem arbitrary when looking for the best d-dimensional approximation, one can show that indeed this is the optimal choice:

Challenge 5.1. *Instead of centering the points at the start, we could have asked for the best approximation in the sense of picking $\beta_k \in \mathbb{R}^d$, a matrix U_d whose columns are a basis for a d-dimensional subspace, and $\mu \in \mathbb{R}^d$ such that (10) is the best possible approximation (in the sense of sum of squares of distances). Show that then $\mu = (1/m)\sum_{i=1}^m y_i$ the empirical mean.*

We end this section by two remarks:

Principal Component Analysis and sample covariance matrix – Another classical way to describe PCA (see, for example, Chapter 3.2 of [BSS23]) is to build the sample covariance matrix of the (centered) data, which is defined as:

$$\frac{1}{m-1} XX^\top = \frac{1}{m-1} \sum_{i=1}^m (y_i - \mu)(y_i - \mu)^\top.$$

PCA is then described as writing the data in the subspace generated by the leading eigenvectors of $XX^\top$. Notice that this is the same as above, since $XX^\top = U\Sigma V^\top (U\Sigma V^\top)^\top = U\Sigma^2 U^\top$, where $X = U\Sigma V^\top$ is the SVD of X. Thus the leading eigenvectors of $XX^\top$ correspond to the leading *left* singular vectors of X: writing the data in the subspace they generate is therefore exactly what we did in eq. (10)!

Principal Component Analysis and Gram matrix – While the basis in which we project the points x_i is given by the leading left singular vectors of X, we also saw that the leading right singular vectors were related to the coordinates in that basis. We note here that they correspond to eigenvectors of the *Gram matrix* of $\{x_i\}_{i=1}^m$, that is the matrix $M \in \mathbb{R}^{m \times m}$ whose entries are

$$M_{ij} \coloneqq \langle x_i, x_j \rangle. \tag{11}$$

Indeed, one has $M = X^\top X = V\Sigma^2 V^\top$, so the right singular vectors of X are the eigenvectors of M.

6 Kernel PCA

6.1 Kernel PCA

PCA aims to find the best low-dimensional linear representation of the data points. But what if the data indeed has some low-dimensional structure, but it is non-linear? For instance, think of Figure 2: PCA will not be able to find a representation of the data that can differentiate the two clusters.

A possible approach to overcome this limitation is to come back to eq. (11): one can interpret the matrix M as M_{ij} measuring affinity between point i and j; indeed $\langle x_i, x_j \rangle$ is larger if x_i and x_j are more similar. With this interpretation we can define versions of PCA with other notions of affinity

$$M_{ij} = K(x_i, x_j),$$

where the affinity function K is often called a Kernel. This is the idea behind *Kernel PCA*. Notice that this can be defined even when the data points are not in Euclidean space. Moreover in Kernel PCA we will consider the top eigenvectors of M: according to the previous section, this will give us a low-dimensional representation of the data, but not how this representation is built. This is often sufficient, as e.g. in clustering: we do not need to know how the low-dimensional representation is built as long as we can use it to cluster the data points!

Example: Gaussian Kernel – A common choice of Kernel is the so-called Gaussian kernel

$$K(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2/\varepsilon^2\right),$$

for $\varepsilon > 0$. The intuition of why one would use this notion of affinity is that it tends to ignore distances at a scale larger than ε; if data has a low dimensional structure embedded, with some curvature, in a larger dimensional ambient space then small distances in the ambient space should be similar to intrinsic distances, but larger distances are less reliable (recall Figure 2). In Fig. 4 we show how Kernel PCA with a Gaussian Kernel allows to efficiently cluster data similar to that of Fig. 2. See Chapter 5 in [BSS23] for more on this, and some other illustrative pictures.

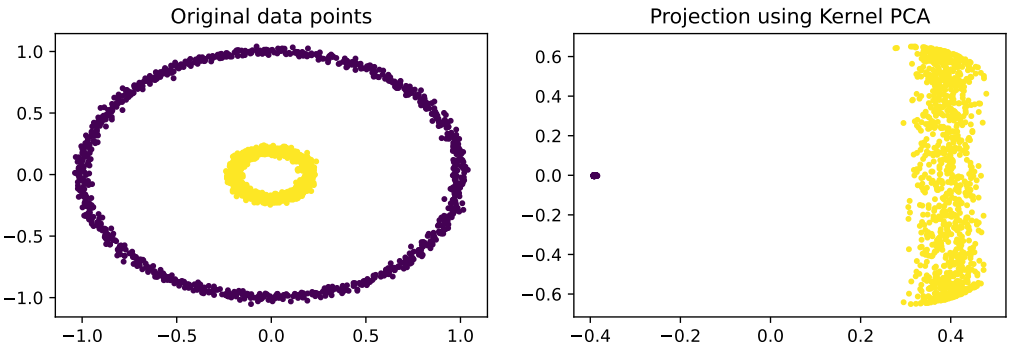

Figure 4: Clustering data points made of two (noisy) concentric circles using Kernel PCA with a Gaussian Kernel. When considering the two top eigenvectors of the Kernel matrix (on the right), we can easily cluster the data!

There is an alternative way of interpreting Kernel PCA: rather than seeing it as changing the affinity measure, we can also think of it as changing the way data points are represented. Then Kernel PCA is doing PCA on the new representation, i.e. $K(x_i, x_j) = \langle \phi_i, \phi_j \rangle$, where these new "representations" (ϕ_i) of the points (x_i) are often referred to as features, in Machine Learning. This observation will be further explored below.

Importantly, in order for the interpretation above to apply we need $M \succeq 0$ ($M \succeq 0$ means M is positive semidefinite, all eigenvalues are non-negative; we only use the notation $M \succeq 0$ for symmetric matrices). This motivates the definition of *positive definite kernels*:

Definition 6.1 (*Positive definite kernels*)

A kernel function K is *positive definite* if for any $n \geq 0$ and any $x_1, \cdots, x_n \in \mathbb{R}^p$, the matrix $(K(x_i, x_j))_{i,j=1}^n$ is symmetric and positive semi-definite.

Note the unfortunate choice of wording in this definition: a kernel is positive definite iff the associated matrices are positive *semi*-definite.

When this is the case, we can write the Cholesky decomposition of $M = (K(x_i, x_j))_{i,j=1}^n$ as

$$M = \Phi^\top \Phi,$$

for some matrix Φ. If ϕ_i is the i-th column of Φ then

$$M_{ij} = K(x_i, x_j) = \langle \phi_i, \phi_j \rangle,$$

for this reason ϕ_i is commonly referred to, in the Machine Learning community, referred to as the *feature* vector of i.

Challenge 6.1. *Show that the Gaussian Kernel $K(x, y) \coloneqq \exp\left(-\|x - y\|^2/\varepsilon^2\right)$ is positive definite.*

Further Reading 6.2. A very natural question is whether the feature vectors ϕ_i can be written as $\phi_i = \varphi(x_i)$, where the function φ (called the feature map) depends only on the kernel K and not on the data points. This turns out to be true, and is related to the celebrated *Mercer Theorem* (essentially a spectral theorem for positive definite kernels).

Exploratory Challenge 6.2. *Can you describe the feature map associated to the Gaussian Kernel?*

References – A more advanced introduction to kernel methods can be found e.g. in the lecture notes [Bac21], see also the references therein.

7 Graphs and Networks

In this section we will study networks, also called graphs.

Definition 7.1 (*Graph*)

A graph is a mathematical object consisting of a set of vertices V and a set of edges $E \subseteq \binom{V}{2}$. We will focus on undirected, unweighted, graphs. We say that $i \sim j$, i is connected to j, if $(i,j) \in E$. We assume graphs have no self-loops, i.e. $(i,i) \notin E$ for all i. These are usually called simple graphs.

In what follows the graph will have n nodes ($|V| = n$). It is sometimes useful to consider a weighted graph, in which an edge (i,j) has a non-negative weight w_{ij}. Essentially everything remains the same if considering weighted graphs, we focus on unweighted graphs to lighten the notation (See Chapter 4 in [BSS23] for a similar treatment that includes weighted graphs).

Definition 7.2 (*Degree and d-regular graph*)

The *degree* of a node i, $\deg(i)$, is the number of neighbors of node i. A graph is said to be d-regular if $\deg(i) = d$ for all $i \in V$.

A useful way to represent a graph is via its *adjacency matrix*. Given a graph $G = (V,E)$ on n nodes ($|V| = n$), we define its adjacency matrix $A \in \mathbb{R}^{n \times n}$ as the symmetric matrix with entries

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that for d-regular graphs, $A\mathbf{1} = d\mathbf{1}$, where $\mathbf{1}$ is the all-ones vector.

Proposition 7.3 (*Spectral norm of a d-regular graph*)

For A the adjacency of a d-regular graph, $\|A\| \leq d$.

Challenge 7.1. *Prove Proposition 7.3.*

We denote $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ the eigenvalues of A. Note that Proposition 7.3 means that, for a d-regular graph: $\lambda_1(A) = d$ and the corresponding leading eigenvector of A is $v_1 = \frac{1}{\sqrt{n}}\mathbf{1}$.

Remark 7.4. Note that, for a d-regular graph, the matrix $K = \mathrm{I}_n + \frac{1}{d}A$ is PSD. Motivated by the discussion a couple of weeks ago on Kernel PCA, it would be natural to do PCA on this matrix in a attempt to "draw" the graph in a low-dimensional space (after discarding the first "boring" principal component $v_1 = \frac{1}{\sqrt{n}}\mathbf{1}$). This has many names (they are slightly different variants that end up being the same in the case of regular graphs), it is known as, among other things, "Laplacian eigenmaps" and "Diffusion Maps" (see [BSS23]). If you have the idea to cluster the PCA projected data points using k-means, you basically rediscover "Spectral Clustering"! More below.

Challenge 7.2. *For which d-regular graphs do we have that $\lambda_2(A) = \lambda_1(A)$?*

Challenge 7.3. *For which d-regular graphs do we have that $|\lambda_n(A)| = |\lambda_1(A)|$?*

Exploratory Challenge 7.4. *For a $d-$regular graph, we call the* Fiedler *value*

$$f_G \coloneqq \max\{|\lambda_2(G)|, |\lambda_n(G)|\}.$$

Graphs with small Fiedler value are called Expanders, and are very important in many areas of Mathematics, Computer Science, and Engineering. Graphs for which $f_G \leq 2\sqrt{d-1}$ are called Ramanujan graphs. The first constructions were based on Number Theory. To this day, we still don't know that they exist for all degrees d, so here is a fascinating open problem:

- *Is it true that for all integers $d \geq 3$, and all integers n_0, there is a d-regular graph on $n \geq n_0$ nodes satisfying $f_G \leq 2\sqrt{d-1}$?*

Challenge 7.5. *It is true that $2\sqrt{d-1}$ in the Exploratory Challenge above is unimprovable (this is known as Alon-Boppana's Theorem). A weaker version of this theorem is (relatively) easy to show, try it: show that, for any d-regular graph, we have $f_G \geq \sqrt{d-1}$.*

A few definitions will be useful.

Definition 7.5 (*Cut and Connectivity*)

Given a subset $S \subseteq V$ of the vertices, we call $S^c := V \setminus S$ the complement of S and we define

$$\mathrm{cut}(S) := \sum_{i \in S} \sum_{j \in S^c} 1_{(i,j) \in E},$$

as the number of edges "cut" by the partition (S, S^c), where 1_X is the indicator of X. Furthermore, we say that a graph G is disconnected if there exists $\emptyset \subsetneq S \subsetneq V$ such that $\mathrm{cut}(S) = 0$.

It is useful to consider the following quadratic form in $\mathbb{R}^n$:

$$Q(x) := \sum_{(i,j) \in E} (x_i - x_j)^2.$$

Definition 7.6 (*Graph Laplacian*)

The symmetric matrix associated with this Quadratic Form is the celebrated *Graph Laplacian* L. In other words, $L \in \mathbb{R}^{n \times n}$ is the symmetric matrix that satisfies $Q(x) = x^\top L x$ for all $x \in \mathbb{R}^n$.

Definition 7.7 (*Degree Matrix*)

The diagonal matrix whose diagonal elements are the degrees of the graph G is known as the *Degree Matrix* D of G. It is given by

$$D_{ij} = \begin{cases} \deg(i) & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 7.8 (*Properties of the Graph Laplacian*)

Let $G = (V, E)$ be a d-regular graph, the following three definitions for $L_G \in \mathbb{R}^{n \times n}$ its graph Laplacian, are equivalent:

(*i*) L_G is the symmetric matrix such that, for all $x \in \mathbb{R}^n$

$$x^\top L_G x = \sum_{(i,j) \in E} (x_i - x_j)^2.$$

(*ii*) L_G is given by

$$L_G = \sum_{(i,j) \in E} (e_i - e_j)(e_i - e_j)^\top,$$

where e_i is the i-th element of the canonical basis.

(*iii*) $L_G = D - A$.

Challenge 7.6. *Prove Proposition 7.8.*

Proposition 7.9

The graph Laplacian L_G satisfies: $L_G \succeq 0$ and $L_G \mathbf{1} = 0$, where $\mathbf{1}$ denotes the all-ones vector.

Challenge 7.7. *Prove Proposition 7.9.*

Remark 7.10. Notice that, because in the definition of graph Laplacian, the matrix A appears with a negative sign. Therefore the largest eigenvalues of A become the smallest ones of L_G. Since $L_G \succeq 0$ the eigenvalues of L_G are usually ordered from smallest to largest $\lambda_1(L_G) \leq \cdots \leq \lambda_n(L_G)$. Note that $\lambda_1(L_G) = 0$.

Challenge 7.8. *Can you relate the dimension of the nullspace of L_G with the number of connected components of L_G?*

Remark 7.11. Note that when G is a d-regular graph, we have $L_G = dI - A$ and so $\lambda_i(L_G) = d - \lambda_i(A)$.

8 Graphs Cuts and Spectral Graph Theory

If $S \subset V$ and $x = \mathbb{1}_S - \mathbb{1}_{S^c}$ (a vector that takes the value 1 in S and -1 in S^c), then (show it !)

$$\text{cut}(S) = \frac{1}{4} x^\top L_G x$$

When n is an even number, the minimum bisection of a graph, MinBis_G is the minimum number of edges that are cut on a balanced partition of the nodes of the graph.

$$\text{MinBis}_G = \min_{\substack{S \subseteq V, \\ |S| = \frac{n}{2}}} \text{cut}(S) = \min_{\substack{S \subseteq V, \\ |S| = n/2, \\ x = \mathbb{1}_S - \mathbb{1}_{S^c}}} x^\top L_G x = \frac{1}{4} \min_{\substack{x \in \{\pm 1\}^n \\ x \perp \mathbf{1}}} x^\top L_G x. \tag{12}$$

Notice that, since $L_G \mathbf{1} = 0$, by the variational principle for the eigenvalues (Courant-Fisher)[3] we have

$$\lambda_2(A) = \min_{\substack{\|y\|_2 = 1 \\ y \perp \mathbf{1}}} y^\top L_G y. \tag{13}$$

For any $x \in \{\pm 1\}^n$ such that $x \perp \mathbf{1}$, we have that $z = \frac{x}{\sqrt{n}}$ satisfies the constraints in (13). This means that the search space of (13) is larger than the one in (12), we must have

$$\min_{\substack{\|y\|_2 = 1 \\ y \perp \mathbf{1}}} y^\top L_G y \geq \frac{1}{n} \min_{\substack{x \in \{\pm 1\}^n \\ x \perp \mathbf{1}}} x^\top L_G x.$$

We have just proved the following Theorem, which is a first instance of a rigorous connection between the geometry of G and the spectrum of L_G.

Theorem 8.1 (*Min Bisection and spectrum of the graph*)

For G a simple graph on n nodes (with n even) we have

$$\text{MinBis}_G = \min_{\substack{S \subseteq V, \\ |S| = \frac{n}{2}}} \text{cut}(S) \geq \frac{n}{4} \lambda_2(L_G).$$

Challenge 8.1. *[Courant-Fisher — variational formula for eigenvalues and eigenvectors] Let M be an $n \times n$ symmetric matrix. Let $\lambda_1(M) \leq \cdots \lambda_n(M)$ denotes the eigenvalues of M (with multiplicities)[4] and let $v_1, \ldots, v_n$ denote the corresponding eigenvectors. Show the following (including that the various* min *and* max *are attained):*

1. $\lambda_1(M) = \min_{v : \|v\| = 1} v^T M v$*, with the minimum attained by* $v = v_1$*.*
2. $\lambda_2(M) = \min_{v : \|v\| = 1, v \perp v_1} v^T M v$*, with the minimum attained by* $v = v_2$*.*
3. $\lambda_2(M) = \max_{u : \|u\| = 1} \min_{v : \|v\| = 1, v \perp u} v^T M v$*.*
4. *Derive (and prove) a similar variational formula for* λ_k*, for* $k > 2$*.*

The minimum bisection has the drawback that it only takes into account very specific cuts that separate the vertex set into two equally sized sets. We will now introduce different notions of cuts that attempt to find a middle ground between (i) not forcing the cuts to necessarily be bisections, (ii) not promoting very unbalanced cuts, like taking one vertex to one side of the partition, and the rest to the other.

We define the Ratio Cut as follows.

[3] See Challenge 8.1.

[4] Note that we are ordering the eigenvalues non-decreasingly, we could have also ordered them non-increasingly and derive a variational principle that way, mutatis mutandis.

Definition 8.2 (*Ratio Cut*)

Let $G = (V, E)$ be a graph. Given a vertex partition (S, S^c), the Ratio Cut of S is defined as:

$$\mathrm{Rcut}(S) \coloneqq \frac{\mathrm{cut}(S)}{|S|} + \frac{\mathrm{cut}(S)}{|S^c|}.$$

We call Ratio Cut of G the minimal $\mathrm{Rcut}(S)$ over non-trivial[5] partitions:

$$\mathrm{Rcut}_G \coloneqq \min_{\emptyset \subsetneq S \subsetneq V} \mathrm{Rcut}(S).$$

Note that we can rewrite $\mathrm{Rcut}(S) = \frac{\mathrm{cut}(S)}{\min\{|S|,|S^c|\}}$. If a partition (S, S^c) is very unbalanced, then $\frac{1}{\min\{|S|,|S^c|\}}$ will be small, which discourages very unbalanced cuts.

Recall that we ordered the eigenvalues of $L_G = D - A$ as:

$$0 = \lambda_1(L_G) \leq \lambda_2(L_G) \leq \cdots \leq \lambda_n(L_G).$$

We will show the following relationship between the second eigenvalue (also called spectral gap) $\lambda_2(L_G)$ and the ratio cut:

Theorem 8.3 (*Ratio cut and spectral gap*)

Let $G = (V, E)$ be a simple graph. Then

$$\lambda_2(L_G) \leq \mathrm{Rcut}_G.$$

Remark 8.4. Notice that Theorem 8.3 implies Theorem 8.1. Moreover, we recover that disconnected graphs have $\lambda_2(L_G) = 0$ (and the converse is also true, see Challenges 7.2 and 7.8).

Proof of Theorem 8.3 – The key idea in this proof is that of a *relaxation* — when a complicated minimization problem is lower bounded by taking the minimization over a larger, but simpler, set. By the Courant-Fischer variational principal of eigenvalues (see Challenge 8.1) and Proposition 7.8 we know that

$$\lambda_2(L_G) = \min_{\substack{\|z\|=1,\\ z \perp \mathbf{1}_n}} z^\top L_G z = \min_{\substack{\|z\|=1,\\ z \perp \mathbf{1}_n}} \sum_{(i,j)\in E} (z_i - z_j)^2.$$

The key argument is that the Ratio Cut will correspond to the same minimum when we restrict the vector z to be of the form $z = a\mathbb{1}_S + b\mathbb{1}_{S^c}$, i.e. $z \in \{a, b\}^n$ for some $a, b \in \mathbb{R}$. More precisely, for a non-trivial subset $S \subset V$, let us consider the vector $y \in \mathbb{R}^n$ such that

$$y_i = \begin{cases} a & \text{if } i \in S \\ b & \text{if } i \in S^c. \end{cases},$$

which we can also write as $y = a1_S + b1_{S^c}$ (where 1_S corresponds to the indicator of S). For the constraints $\|y\| = 1$ and $y \perp \mathbf{1}_n$ to be satisfied we must have

$$\begin{cases} a^2|S| + b^2|S^c| &= 1, \\ a|S| + b|S^c| &= 0, \end{cases} \tag{14}$$

and therefore $a = [|S^c|/(n|S|)]^{1/2}$ and $b = -[|S|/(n|S^c|)]^{1/2}$ (up to a global sign change). Note that we used $|S| + |S^c| = n$. The rest of the proof proceeds by computing $y^\top L_G y$.

$$y^\top L_G y = \sum_{(i,j)\in E} (y_i - y_j)^2 = (a-b)^2 \sum_{i\in S} \sum_{j \in S^c} 1_{(i,j)\in E}$$

$$= \frac{\text{cut}(S)}{n}\left[\sqrt{\frac{|S^c|}{|S|}} + \sqrt{\frac{|S|}{|S^c|}}\right]^2 = \frac{\text{cut}(S)}{n}\left[\frac{|S^c|}{|S|} + \frac{|S|}{|S^c|} + 2\right]$$
$$= \frac{\text{cut}(S)}{n}\left[\frac{|S^c|}{|S|} + \frac{|S|}{|S^c|} + \frac{|S|}{|S|} + \frac{|S^c|}{|S^c|}\right]$$
$$= \text{cut}(S)\left[\frac{1}{|S|} + \frac{1}{|S^c|}\right] = R(S).$$

Finally we have:

$$\lambda_2(L_G) = \min_{\substack{\|y\|=1,\\ y\perp \mathbf{1}_n}} \sum_{(i,j)\in E}(y_i - y_j)^2 \leq \min_{\substack{\|y\|=1,\, y\perp \mathbf{1}_n \\ y\in\{a,b\}^n \text{ for } a,b\in\mathbb{R}}} \sum_{(i,j)\in E}(y_i - y_j)^2 = \min_{\emptyset\subsetneq S\subsetneq V} R(S). \tag{15}$$

□

When the graph has nodes with a large range of different degrees it is sometimes better to measure the size of a set of nodes S not by $|S|$ but by the sum of the degrees of the nodes in S.

Definition 8.5

Given a subset of nodes $S \subset V$ of a graph, we defined $\text{vol}(S) = \sum_{i\in S}\deg(i)$. We use $\text{vol}(G) = \text{vol}(V)$ to denote the volume of all the vertices.

We will now define the Normalized Cut.

Definition 8.6 (*Normalized Cut*)

Let $G = (V, E)$ be a graph without isolated nodes[6]. Given a vertex partition (S, S^c), the Normalized Cut of S is defined as:

$$\text{Ncut}(S) \coloneqq \frac{\text{cut}(S)}{\text{vol}(S)} + \frac{\text{cut}(S)}{\text{vol}(S^c)}.$$

We call Normalized Cut of G the minimal $\text{Ncut}(S)$ over non-trivial partitions:

$$\text{Ncut}_G \coloneqq \min_{\emptyset\subsetneq S\subsetneq V}\text{Ncut}(S).$$

In order to show an analogue of Theorem 8.3 for the Normalized Cut, we need to defined a Normalized Laplacian.

Definition 8.7 (*Normalized Graph Laplacian*)

Let $G = (V, E)$ be a graph, the normalized graph Laplacian is given by

$$\mathcal{L}_G = D^{-\frac{1}{2}} L_G D^{-\frac{1}{2}},$$

where L_G is the graph laplacian and D the diagonal degree matrix. Note that $\mathcal{L}_G$ is symmetric. We order the eigenvalues of $\mathcal{L}_G$ as $0 = \lambda_1(\mathcal{L}_G) \leq \lambda_2(\mathcal{L}_G) \leq \cdots \leq \lambda_n(\mathcal{L}_G)$.

Theorem 8.8 (*Normalized cut and spectral gap*)

Let $G = (V, E)$ be a simple graph without isolated nodes. Then

$$\lambda_2(\mathcal{L}_G) \leq \text{Ncut}_G.$$

The proof is similar to the one of Theorem 8.3, but we will organize it in two Lemmas for ease of exposition. Theorem 8.8 easy follows from Lemmas 8.9 and 8.10.

Lemma 8.9

Let $G = (V, E)$ be a simple graph. Then

$$\mathrm{Ncut}_G = \min_{\substack{y^T D y = 1,\, y \perp D\mathbf{1}_n \\ y \in \{a,b\}^n \text{ for } a,b \in \mathbb{R}}} y^T L_G y.$$

This proof differs from that of Theorem 8.3 by the fact that the normalization is done with respect to the volumes of the subsets and not the number of nodes. This corresponds to normalizing each node by its degree. The normalized cut also corresponds to minimizing $y^T L_G y$ but for a vector $y = a1_S + b1_{S^c}$ (where 1_S is the indicator of the set S) but rather than constraining to $\sum_i y_i = 0$ and $\sum_i y_i^2 = 1$, one constraints to $\sum_i \deg(i) y_i = 0$ and $\sum_i \deg(i) y_i^2 = 1$ (which can be rewritten as $\mathbf{1}Dy = 0$ and $y^T D y = 1$). For $y = a1_S + b1_{S^c}$, this corresponds to a normalization analogous to (14), but normalized with respect to the volumes:

$$\begin{cases} a^2 \operatorname{vol}(S) + b^2 \operatorname{vol}(S^c) & = 1, \\ a \operatorname{vol}(S) + b \operatorname{vol}(S^c) & = 0. \end{cases} \tag{16}$$

This forces $a = [\operatorname{vol}(S^c)/(\operatorname{vol}(G)\operatorname{vol}(S))]^{1/2}$ and $b = -[\operatorname{vol}(S)/(\operatorname{vol}(G)\operatorname{vol}(S^c))]^{1/2}$ (up to a global sign change). We used $\operatorname{vol}(G) = \operatorname{vol}(S) + \operatorname{vol}(S^c)$.

Similar to the proof of Theorem 8.8 we have that, for $y_S = \left(\frac{\operatorname{vol}(S^c)}{\operatorname{vol}(G)\operatorname{vol}(S)}\right)^{\frac{1}{2}} 1_S + \left(\frac{\operatorname{vol}(S)}{\operatorname{vol}(G)\operatorname{vol}(S^c)}\right)^{\frac{1}{2}} 1_{S^c}$,

$$\begin{aligned}
\min_{\substack{y^T D y = 1,\, y \perp D\mathbf{1}_n \\ y \in \{a,b\}^n \text{ for } a,b \in \mathbb{R}}} y^T L_G y &= \min_{\emptyset \subsetneq S \subsetneq V} \sum_{(i,j) \in E} ((y_S)_i - (y_S)_j)^2 \\
&= \min_{\emptyset \subsetneq S \subsetneq V} \left(\left(\frac{\operatorname{vol}(S^c)}{\operatorname{vol}(G)\operatorname{vol}(S)} \right)^{\frac{1}{2}} - \left(\frac{\operatorname{vol}(S)}{\operatorname{vol}(G)\operatorname{vol}(S^c)} \right)^{\frac{1}{2}} \right)^2 \sum_{i \in S} \sum_{j \in S^c} 1_{(i,j) \in E} \\
&= \min_{\emptyset \subsetneq S \subsetneq V} \frac{\operatorname{cut}(S)}{\operatorname{vol}(G)} \left[\sqrt{\frac{\operatorname{vol}(S^c)}{\operatorname{vol}(S)}} + \sqrt{\frac{\operatorname{vol}(S)}{\operatorname{vol}(S^c)}} \right]^2 = \frac{\operatorname{cut}(S)}{\operatorname{vol}(G)} \left[\frac{\operatorname{vol}(S^c)}{\operatorname{vol}(S)} + \frac{\operatorname{vol}(S)}{\operatorname{vol}(S^c)} + 2 \right] \\
&= \min_{\emptyset \subsetneq S \subsetneq V} \frac{\operatorname{cut}(S)}{\operatorname{vol}(G)} \left[\frac{\operatorname{vol}(S^c)}{\operatorname{vol}(S)} + \frac{\operatorname{vol}(S)}{\operatorname{vol}(S^c)} + \frac{\operatorname{vol}(S)}{\operatorname{vol}(S)} + \frac{\operatorname{vol}(S^c)}{\operatorname{vol}(S^c)} \right] \\
&= \min_{\emptyset \subsetneq S \subsetneq V} \operatorname{cut}(S) \left[\frac{1}{\operatorname{vol}(S)} + \frac{1}{\operatorname{vol}(S^c)} \right] = \min_{\emptyset \subsetneq S \subsetneq V} \mathrm{Ncut}(S) = \mathrm{Ncut}_G.
\end{aligned}$$

□

Lemma 8.10

Let $G = (V, E)$ be a simple graph without isolated nodes. Then

$$\lambda_2(\mathcal{L}_G) = \min_{\substack{y^T D y = 1 \\ y \perp D\mathbf{1}_n \\ y \in \mathbb{R}^n}} y^T L_G y.$$

This Lemma is also a consequence of the Courant-Fischer variational principal of eigenvalues (see Challenge 8.1) and Proposition 7.8. Namely, consider $z = D^{\frac{1}{2}} y$, and notice that D is invertible. We have,

$$\min_{\substack{y^T D y = 1 \\ y \perp D\mathbf{1}_n \\ y \in \mathbb{R}^n}} y^T L_G y = \min_{\substack{z^T z = 1 \\ z \perp D^{\frac{1}{2}}\mathbf{1}_n \\ z \in \mathbb{R}^n}} z^T D^{-\frac{1}{2}} L_G D^{-\frac{1}{2}} z$$

$$= \min_{\substack{\|z\|=1 \\ z \perp D^{\frac{1}{2}}\mathbf{1}_n \\ z \in \mathbb{R}^n}} z^T \mathcal{L}_G z$$
$$= \lambda_2(\mathcal{L}_G),$$

where the last line follows from Courant-Fischer and the fact that $D^{\frac{1}{2}}\mathbf{1}_n$ is the eigenvector corresponding to the smallest eigenvalue of $\mathcal{L}_G$ (as $\mathcal{L}_G \mathbf{1}_n = 0$).

□

Remark 8.11. Note that if the graph is d-regular then $D = dI$ and Theorems 8.3 and 8.8 coincide.

There are (at least) two consequential ideas of this result:

1. The way cuts of partitions are measured in $\mathrm{Rcut}(S)$ or $\mathrm{Ncut}(S)$ promotes somewhat balanced partitions (so that neither S nor S^c are too small), this turns out to be beneficial to avoid trivial solutions such as partition a graph by splitting just one node from all the others. In a sense, one measures size by number of vertices and the other by number of edges.

2. There is an important algorithmic consequence of (15): when we want to cluster a network in two groups, what we want to minimize is the RHS of (15), this is unfortunately computationally intractable (in fact, it is known to be NP-hard). However, the LHS of the inequality is a spectral problem and so computationally tractable. This is the idea behind the popular algorithm of *Spectral clustering* (Algorithm 8.1).

Algorithm 8.1 Spectral Clustering

Given a d-regular graph $G = (V, E)$, let v_2 be the eigenvector corresponding to the second smallest eigenvalue of the Laplacian $\mathcal{L}_G$. Let $\phi_2 = D^{-\frac{1}{2}} v_2$ Given a threshold $\tau \in \mathbb{R}$ (one can try all different possibilities, or run k-means in the entries of v_2 for $k = 2$), set S to be the minimum of

$$S_\tau = \{i \in V : \phi_2(i) \leq \tau\}. \tag{17}$$

Algorithm 8.1 should be thought about as "projecting" the nodes of the graph in a one-dimensional space, before trying to cluster them using this one-dimensional projection.

Remark 8.12. With this interpretation in mind, Algorithm 8.1 can be generalized to cluster data into $k > 2$ clusters. In that case one considers the $k - 1$ eigenvectors (from the 2nd to the kth) and to each node i we associate the $k - 1$ dimensional representation

$$[\phi_2(i), \phi_3(i), \cdots, \phi_k(i)]^\top,$$

and use k-means on this representation. Many different normalizations exist, and using v_2 or ϕ_2 corresponds to two popular choices (corresponding to Rcut or Ncut).

Remark 8.13 (Spectral clustering and Kernel PCA)**.** Spectral clustering should not appear "magical": in the particular case of d-regular graphs it is simply doing k-means on the representation given by kernel PCA, in which the kernel matrix is $K = d\mathrm{I}_n + A$ (which is PSD, and has the same eigenvectors as the Laplacian), that is – up to a shift – the adjacency matrix of the graph. And the adjacency is the most natural "affinity" kernel one can design from a graph, so it is very natural to use it to do kernel PCA! See also Remark 7.4. This is oftentimes referred to as "Diffusion Maps", "Laplacian Eigenmaps", or "Spectral Embedding", see for example Chapter 5 in [BSS23].

8.1 Cheeger's inequality and guarantees for spectral clustering

A natural question is whether one can give a guarantee for spectral clustering: "Does Algorithm 8.1 produce a partition whose ratio cut is comparable with R_G?" We give the proof of such a guarantee in the next section, but we will briefly describe it below, and highlight its relation to the celebrated Cheeger's Inequality.

For this is it is best to define a slight adaptation of the notion of normalized cut, known as Cheeger cut.

Definition 8.14 (*Cheeger cut*)

Let G be a graph and S a non-trivial subset of V, then

$$h(S) = \frac{\mathrm{cut}(S)}{\min\{\mathrm{vol}(S), \mathrm{vol}(S^c)\}} = \max\left\{\frac{\mathrm{cut}(S)}{\mathrm{vol}(S)}, \frac{\mathrm{cut}(S)}{\mathrm{vol}(S^c)}\right\},$$

and

$$h_G = \min_{\emptyset \subsetneq S \subsetneq V} h(S).$$

We also define the smallest $h(S)$ obtained by (17):

$$\alpha_G = \min_{\tau \in \mathbb{R}} h(S_\tau),$$

where S_τ is given by (17).

The following is a consequence of these definitions

Proposition 8.15

We have $\mathrm{Ncut}_G \leq h_G \leq 2\mathrm{Ncut}_G$ and $h_G \leq \alpha_G$.

The Spectral clustering algorithm will give a Cheeger cut of α_G. In the next section we will show (Lemma 9.1) that $\alpha_G \leq \sqrt{2\lambda_2(\mathcal{L}_G)}$. This has a couple of interesting consequencies. Together with Theorem 8.8 it implies the celebrated Cheeger's inequality.

Theorem 8.16 (*Cheeger's Inequality*)

Let $G = (V, E)$ be a simple graph. The following holds:

$$\frac{1}{2}\lambda_2(L_G) \leq h_G \leq \sqrt{2\lambda_2(L_G)}.$$

It also implies a guarantee on the performance of Spectral clustering.

Corollary 8.17 (*Spectral clustering guarantee*)

There is a threshold $\tau \in \mathbb{R}$ in Algorithm 8.1 producing a partition S such that

$$h(S) \leq \sqrt{2\lambda_2(L_G)} \leq \sqrt{4h_G}.$$

Remark 8.18. Cheeger's inequality can be easily adapted to weighted graphs, see for example [BSS23].

Cheeger's inequality was first established for manifolds by Jeff Cheeger in 1970 [Che70], the graph version is due to Noga Alon and Vitaly Milman [Alo86, AM85] in the mid 80s.

9 The proof of Cheeger's Inequality

In this section we will prove the following Lemma (there are several known proofs of it, see [Chu10] or [BSS23], here we follow a proof Amit Singer presented in a course I took in Princeton; any errors/typos are my own).

Lemma 9.1

Let G be a simple graph without isolated nodes. Then

$$\alpha_G \leq \sqrt{2\lambda_2(\mathcal{L}_G)},$$

where α_G is as Definition 8.14 and $\mathcal{L}_G$ as in Definition 8.7.

Proof

Given $v \in \mathbb{R}^n$ let

$$\mathcal{R}(v) = \frac{v^T L_G v}{v^T D v}$$

denote the Rayleight quotient. Let z be the eigenvector of $\mathcal{L}_G$ associated with the eigenvalue $\lambda_2(\mathcal{L}_G)$. Then $y = D^{-\frac{1}{2}} z$ satisfies[7]

$$\mathcal{R}(y) = \frac{y^T L_G y}{y^T D y} = \frac{z^T \mathcal{L}_G z}{z^T z} = \lambda_2(\mathcal{L}_G), \tag{18}$$

and $y \perp D^{\frac{1}{2}}\mathbf{1}$. The goal is to find a rounding of y (as in (17)). We will start by ordering the nodes, WLOG, we take the ordering such that

$$y_1 \geq \cdots \geq y_n,$$

which allows us to write $\alpha_G = \min_i h_{S_i}$ where $S_i = \{1, \cdots, i\}$.

Let r be the largest integer so that $\mathrm{vol}(S_r) \leq \frac{1}{2}\mathrm{vol}(G)$ (r is a midpoint, in terms of volume). This allow us to write

$$h_{S_i} = \begin{cases} \frac{\mathrm{cut}(S_i)}{\mathrm{vol}(S_i)} & \text{if } i \leq r, \\ \frac{\mathrm{cut}(S_i)}{\mathrm{vol}(V \setminus S_i)} & \text{if } i > r. \end{cases} \tag{19}$$

The goal of the proof is to show that we can "round" the entries of y to only two distinct values. Intuitively, making the separation at the index r sounds reasonable (even if it may not be the optimal solution) so we start with creating two vectors, a y^- and a vector y^+, correspond to the indices before and after r, and we will analyse them separately. More precisely, we define

$$y_i^+ = \begin{cases} y_i - y_r & \text{if } y_i > y_r, \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

$$y_i^- = \begin{cases} -y_i + y_r & \text{if } y_i < y_r, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

We will need four useful statements, two equalities and two inequalities, we will leave them as challenges.

Challenge 9.1. *For any $i \in [n]$, we have $y_i - y_r = y_i^+ - y_i^-$.*

Challenge 9.2. *For any $i, j \in [n]$, we have $|y_i - y_j| = |y_i^+ - y_j^+| + |y_i^- - y_j^-|$.*

[7] Note that, in practice, one computes z such that $\mathcal{R}(z) \leq \lambda_2(\mathcal{L}_G) + \varepsilon$, for small ε. The argument presented here depends gracefully on ε, since it sollely depends on $\mathcal{R}(y)$.

Challenge 9.3. *For any positive real numbers α, β, γ, and δ, we have*

$$\frac{\alpha+\beta}{\gamma+\delta} \geq \min\left\{\frac{\alpha}{\gamma}, \frac{\beta}{\delta}\right\}.$$

Challenge 9.4. *Let* deg *denote the vector of degrees of the graph. Since $y \perp$ deg we have, for any $\alpha \in \mathbb{R}$,*

$$\sum_{i=1}^{n} \deg_i y_i^2 \leq \sum_{i=1}^{n} \deg_i (y_i - \alpha)^2.$$

The rest of the proof is separated in two parts, the first is to show that $\lambda_2(\mathcal{L}) = \mathcal{R}(y) \geq \min\{\mathcal{R}(y^+), \mathcal{R}(y^-)\}$; the second lower bounds $\mathcal{R}(y^+)$ and $\mathcal{R}(y^-)$ by $\alpha_G^2/2$.

For the first part, we have (the justification for each line is given below),

$$\begin{aligned}
\lambda_2(\mathcal{L}_G) &= \mathcal{R}(y) = \frac{y^T L y}{y^T D y} && (22)\\
&= \frac{\sum_{i<j} A_{ij}(y_i - y_j)^2}{\sum_{i=1}^n \deg_i y_i^2} && (23)\\
&= \frac{\sum_{i<j} A_{ij}(|y_i^+ - y_j^+| + |y_i^- - y_j^-|)^2}{\sum_{i=1}^n \deg_i y_i^2} && (24)\\
&\geq \frac{\sum_{i<j} A_{ij}(|y_i^+ - y_j^+|^2 + |y_i^- - y_j^-|^2)}{\sum_{i=1}^n \deg_i y_i^2} && (25)\\
&\geq \frac{\sum_{i<j} A_{ij}(|y_i^+ - y_j^+|^2 + |y_i^- - y_j^-|^2)}{\sum_{i=1}^n \deg_i (y_i - y_r)^2} && (26)\\
&= \frac{\sum_{i<j} A_{ij}(|y_i^+ - y_j^+|^2 + |y_i^- - y_j^-|^2)}{\sum_{i=1}^n \deg_i (y_i^+ - y_i^-)^2} && (27)\\
&= \frac{\sum_{i<j} A_{ij}(|y_i^+ - y_j^+|^2 + |y_i^- - y_j^-|^2)}{\sum_{i=1}^n \deg_i ((y_i^+)^2 + (y_i^-)^2)} && (28)\\
&= \frac{\sum_{i<j} A_{ij}|y_i^+ - y_j^+|^2 + \sum_{i<j} A_{ij}|y_i^- - y_j^-|^2}{\sum_{i=1}^n \deg_i (y_i^+)^2 + \sum_{i=1}^n \deg_i (y_i^-)^2} && (29)\\
&\geq \min\left\{\frac{\sum_{i<j} A_{ij}|y_i^+ - y_j^+|^2}{\sum_{i=1}^n \deg_i (y_i^+)^2}, \frac{\sum_{i<j} A_{ij}|y_i^- - y_j^-|^2}{\sum_{i=1}^n \deg_i (y_i^-)^2}\right\} && (30)\\
&= \min\left\{\mathcal{R}(y^+), \mathcal{R}(y^-)\right\}, && (31)
\end{aligned}$$

where for (23) recall that $A_{ij} = 1_{(i,j)\in E}$; (24) follows from Challenge 9.2; (25) follows from the fact that $(|\alpha| + |\beta|)^2 \geq |\alpha|^2 + |\beta|^2$; (26) follows from Challenge 9.4; (27) follows from Challenge 9.1; (28) follows by noting that $y_i^+ y_i^- = 0$; and (30) follows from Challenge 9.3.

Now we proceed by bounding $\mathcal{R}(y^+)$ below and leave doing the same for $\mathcal{R}(y^-)$ as a challenge

Challenge 9.5. *Use the same argument we use below to show $\mathcal{R}(y^-) \geq \frac{\alpha_G^2}{2}$.*

We have

$$\begin{aligned}
\mathcal{R}(y^+) &= \frac{\sum_{i<j} A_{ij}|y_i^+ - y_j^+|^2}{\sum_{i=1}^n (y_i^+)^2 \deg_i} && (32)\\
&= \frac{\sum_{i<j} A_{ij}|y_i^+ - y_j^+|^2}{\sum_{i=1}^n (y_i^+)^2 \deg_i} \frac{\sum_{i<j} A_{ij}|y_i^+ + y_j^+|^2}{\sum_{i<j} A_{ij}|y_i^+ + y_j^+|^2} && (33)\\
&\geq \frac{\sum_{i<j} A_{ij}|y_i^+ - y_j^+|^2}{\sum_{i=1}^n (y_i^+)^2 \deg_i} \frac{\sum_{i<j} A_{ij}|y_i^+ + y_j^+|^2}{2\sum_{i<j} A_{ij}(|y_i^+|^2 + |y_j^+|^2)} && (34)
\end{aligned}$$

$$= \frac{\sum_{i<j} A_{ij}|y_i^+ - y_j^+|^2}{\sum_{i=1}^n (y_i^+)^2 \deg_i} \frac{\sum_{i<j} A_{ij}|y_i^+ + y_j^+|^2}{2\sum_{i=1}^n (y_i^+)^2 \deg_i} \tag{35}$$

$$\geq \frac{\left(\sum_{i<j} A_{ij}((y_i^+)^2 - (y_j^+)^2)\right)^2}{2\left(\sum_{i=1}^n (y_i^+)^2 \deg_i\right)^2} \tag{36}$$

where (34) follows from $(\alpha+\beta)^2 \leq 2(\alpha^2+\beta^2)$; and (36) from the Cauchy-Schwarz inequality on $\left\{A_{ij}^{\frac{1}{2}}|y_i^+ + y_j^+|^2\right\}_{i<j}$ and $\left\{A_{ij}^{\frac{1}{2}}|y_i^+ - y_j^+|^2\right\}_{i<j}$.

For the next step, we will use the following identity, noting that the entries of y are in non-decreasing order:

$$\sum_{i<j} A_{ij}((y_i^+)^2 - (y_j^+)^2) = \sum_{i<j} A_{ij} \sum_{\ell=i}^{j-1} \left((y_\ell^+)^2 - (y_{\ell+1}^+)^2\right) = \sum_{\ell=1}^{n-1} \left(\sum_{i\leq\ell}\sum_{j>\ell} A_{ij}\right) \left((y_\ell^+)^2 - (y_{\ell+1}^+)^2\right).$$

Since $\mathrm{cut}(S_\ell) = \sum_{i\leq\ell}\sum_{j>\ell} A_{ij}$ we have

$$\sum_{i<j} A_{ij}((y_i^+)^2 - (y_j^+)^2) = \sum_i \mathrm{cut}(S_i)((y_i^+)^2 - (y_{i+1}^+)^2)$$

Thus, we have

$$\mathcal{R}(y^+) \geq \frac{\left(\sum_i \mathrm{cut}(S_i)((y_i^+)^2 - (y_{i+1}^+)^2)\right)^2}{2\left(\sum_i (y_i^+)^2 \deg_i\right)^2} \tag{37}$$

$$= \frac{\left(\sum_i h_{S_i} \mathrm{vol}(S_i)((y_i^+)^2 - (y_{i+1}^+)^2)\right)^2}{2\left(\sum_i (y_i^+)^2 \deg_i\right)^2} \tag{38}$$

$$\geq \frac{\alpha_G^2}{2} \frac{\left(\sum_i \mathrm{vol}(S_i)((y_i^+)^2 - (y_{i+1}^+)^2)\right)^2}{\left(\sum_i (y_i^+)^2 \deg_i\right)^2} \tag{39}$$

$$= \frac{\alpha_G^2}{2} \frac{\left(\sum_i (y_i^+)^2 (\mathrm{vol}(S_i) - \mathrm{vol}(S_{i-1}))\right)^2}{\left(\sum_i (y_i^+)^2 \deg_i\right)^2} \tag{40}$$

$$\geq \frac{\alpha_G^2}{2} \frac{\left(\sum_i (y_i^+)^2 \deg_i\right)^2}{\left(\sum_i (y_i^+)^2 \deg_i\right)^2} \tag{41}$$

$$= \frac{\alpha_G^2}{2}, \tag{42}$$

where (38) follows from the fact that the summand in the denominator is only non-zero for $i \leq r$; (39) follows from the definition of α_G; and (40) by the so called Summation-by-parts formula (see Challenge 9.6.

ASB: TBC

Challenge 9.6 (Summation-by-parts)**.** *Show the equality in (40). Do you see the analogy with integration-by-parts? Can you come up with a general summation-by-parts formula?*

□

10 Parsimony, compressed sensing and sparse recovery

10.1 Parsimony

In this section 10.1 of the notes we recall some general observations we made in the very first lecture.

Parsimony is an important principle in machine learning. The key idea is that oftentimes one wants to learn (or recover) an object with a particular structure. It is also important in supervised learning, the key idea there being that classifiers (or regression rules, as you will see in a Statistics course) that are simple are in theory more likely to generalize to unseen data. We may see some of these phenomena in the very last part of the course, see also the notes of the last years [BZ22].

Observations of this type date back at least to eight centuries ago, the most notable instance being William of Ockham's celebrated *Occam's Razor*: "Entia non-sunt multiplicanda praeter necessitatem (Entities must not be multiplied beyond necessity)", which is today used as a synonym for parsimony. One example discussed in last year's notes [BZ22] is recommendation systems, in which the goal is to make recommendations of a product to users based both on the particular user scores of other items, and the scores other users gives to items. The score matrix whose rows correspond to users, columns to items, and entries to scores is known to be low rank and this form of parsimony is key to perform "matrix completion", meaning to recover (or estimate) unseen scores (matrix entries) from the ones that are available.

A simpler form of parsimony is *sparsity* (i.e. having few non-zero entries). Not only is sparsity present in many problems, including signal and image processing, but the mathematics arising from its study are crucial also to solve problems such as matrix completion. In what follows we will use image processing as the driving motivation.

Sparse recovery – Most of us have noticed how saving an image in JPEG dramatically reduces the space it occupies in our hard drives (as opposed to file types that save the pixel value of each pixel in the image, e.g. TIFF of BMP). The idea behind these compression methods is to exploit known structure in the images; although our cameras will record the pixel value (even three values in RGB) for each pixel, it is clear that most collections of pixel values will not correspond to pictures that we would expect to see. This special structure tends to exploited via sparsity. Indeed, natural images are known to be sparse in certain bases (such as the wavelet base) and this is the core idea behind JPEG (actually, JPEG2000; JPEG uses a different basis). There is an example illustrating this in the jupyter notebook accompanying the class.

Let us think of $x \in \mathbb{R}^N$ as the signal corresponding to the image already in the basis for which it is sparse, meaning that it has few non-zero entries. We use the notation $\|x\|_0$ for the number of non-zero entries of x, it is common to refer to this as the ℓ_0 norm, even though it is not actually a norm. Let us assume that $x \in \mathbb{R}^N$ is s-sparse, i.e. $\|x\|_0 \leq s$. Usually we will assume $s \ll N$. This means that, when we take a picture, our camera makes N measurements (each corresponding to a pixel) but then, after an appropriate change of basis, it keeps only $s \ll N$ non-zero coefficients and drops the others.

This motivates the question: "If only a few degrees of freedom are kept after compression, why not measure in a more efficient way and take considerably less than N measurements?". This question is in the heart of Compressed Sensing. It is particularly important in MRI imaging as less measurements potentially means less measurement time. The following book is a great reference on Compressed Sensing [FR13].

10.2 Compressed Sensing and Sparse Recovery

More precisely, given a s-sparse vector $x \in \mathbb{K}^n$ (with $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$), we take M linear measurements $y_i = a_i^\top x$, with $s < M \ll N$, and measurement (or sensing) vectors $\{a_i\}_{i=1}^M$. Our goal is to recover x

from the underdetermined system:

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} = \begin{bmatrix} & & \\ & \Phi & \\ & & \end{bmatrix} \begin{bmatrix} \\ x \\ \\ \\ \end{bmatrix}.$$

Here, $\Phi \in \mathbb{R}^{M \times N}$ is the matrix whose i-th row is a_i. Since the system is underdetermined and we know x is sparse, the natural thing to try in order to recover x is to solve

$$\begin{array}{ll} \min & \|z\|_0 \\ \text{s.t.} & \Phi z = y, \end{array} \tag{43}$$

and hope that the optimal solution z corresponds to the signal in question x.

Remark 10.1. There is another useful way to think about (43), which we will discuss later in the section on finite frame theory. We can think of the columns of Φ as a redundant "dictionary". In that case, the goal becomes to represent a vector $y \in \mathbb{K}^M$ as a linear combination of the dictionary elements. To leverage the redundancy a common choice is to use the sparsest representation, corresponding to solving problem (43).

Definition 10.2 (*Spark*)

The spark of a matrix Φ is the minimum number of columns of the matrix that make up a linearly dependent set.

Challenge 10.1. *For a matrix Φ, show that* $\text{spark}(\Phi) \leq \text{rk}(\Phi) + 1$. *Can you prove it in a single line?*

We can give a first guarantee for the solution of eq. (43) to be actually x:

Proposition 10.3

If x is s-sparse and $\text{spark}(\Phi) > 2s$ then x is the unique solution to (43) for $y = \Phi x$.

Proof of Prop 10.3 – Assume that there exists $x' \neq x$ such that $y = \Phi x'$ and $\|x'\|_0 \leq \|x\|_0 \leq s$. Then $\Phi[x - x'] = 0$. Since $\|x - x'\|_0 \leq 2s$ and $x - x' \neq 0$, this implies that there is a set of at most $2s$ columns of Φ which are linearly dependent, in contradiction with our assumptions. □

Challenge 10.2. *Can you construct Φ with large spark and small number of measurements M?*

There are two significant issues with (43), stability (as the ℓ_0 norm is very brittle) and computation. In fact, (43) is known to be a computationally hard problem in general (provided $\texttt{P} \neq \texttt{NP}$). Instead, the approach usually taken in sparse recovery is to consider a convex relaxation of the ℓ_0 norm, the ℓ_1 norm: $\|z\|_1 \coloneqq \sum_{i=1}^N |z_i|$. Figure 5 depicts how the ℓ_1 norm can be seen as a convex relaxation of the ℓ_0 norm and how it promotes sparsity. This motivates one to consider the following optimization problem (surrogate to (43)):

$$\begin{array}{ll} \min & \|z\|_1 \\ \text{s.t.} & \Phi z = y, \end{array} \tag{44}$$

For (44) to be useful, two things are needed: (i) the solution of it needs to be meaningful (hopefully to coincide with x) and (ii) (44) should be efficiently solvable. We first consider (ii) in Section 10.2.1, and then discuss (i) in Section 10.2.2.

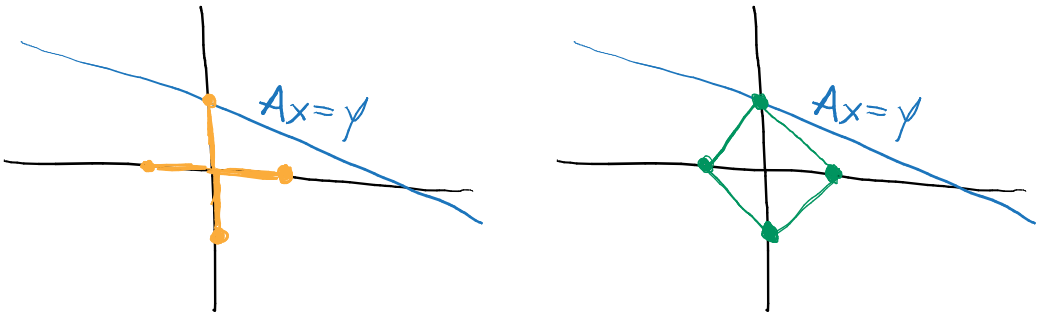

Figure 5: A two-dimensional depiction of ℓ_0 and ℓ_1 minimization. In ℓ_1 minimization (the picture of the right), one inflates the ℓ_1 ball (the diamond) until it hits the affine subspace of interest, this image conveys how this norm promotes sparsity, due to the pointy corners on sparse vectors.

10.2.1 Computational efficiency

To address computational efficiency we will focus on the real case ($\mathbb{K} = \mathbb{R}$) and formulate (44) as a Linear Program (and thus show that it is efficiently solvable). Let us define ω^+ as the positive part of x and ω^- as the negative part of x, meaning that $x = \omega^+ - \omega^-$ and, for each i, either ω_i^- or ω_i^+ is zero. Note that, in that case (for $x \in \mathbb{R}^N$),

$$\|x\|_1 = \sum_{i=1}^{N} [\omega_i^+ + \omega_i^-] = \mathbf{1}^\top (\omega^+ + \omega^-).$$

Therefore, we are led to consider:

$$\begin{array}{ll} \min & \mathbf{1}^\top (\omega^+ + \omega^-) \\ \text{s.t.} & \Phi (\omega^+ - \omega^-) = y \\ & \omega^+ \geq 0 \\ & \omega^- \geq 0, \end{array} \tag{45}$$

which is a linear program. It is not difficult to see (prove it!) that the optimal solution of (45) will indeed satisfy that, for each i, either ω_i^- or ω_i^+ is zero and the program above is indeed equivalent to (44). Since linear programs are efficiently solvable [VB04], this means that (44) can be solved efficiently.

Remark 10.4. While (44) does not correspond to a linear program in the Complex case $\mathbb{K} = \mathbb{C}$ it is nonetheless efficient to solve, the key property is that it is a convex problem, but a general discussion about convexity is outside the scope of this course.

10.2.2 Exact recovery via ℓ_1 minimization

The goal now is to show that, under certain conditions, the solution of (44) for $y = \Phi x$ indeed coincides with x. There are several approaches to this, we refer to [BSS23] for a few alternatives. Here we will discuss a deterministic approach based on the notion of *coherence*.

Let $S = \text{supp}(x)$ (i.e. $x_i \neq 0 \Leftrightarrow i \in S$) and suppose that $z \neq x$ is an optimal solution of the ℓ_1 minimization problem (44) with $y = \Phi x$. Let $v := z - x$, so $z = v + x$ and notice that we must have:

$$\|v + x\|_1 \leq \|x\|_1 \quad \text{and} \quad \Phi(v + x) = \Phi x,$$

so that $\Phi v = 0$. For a vector $u \in \mathbb{R}^N$, we define $u_S = (u_i)_{i \in S} \in \mathbb{R}^{|S|}$, and we let $\|u\|_S := \|u_S\|_1 = \sum_{i \in S} |u_i|$. We have:

$$\|x\|_S = \|x\|_1 \geq \|v + x\|_1 = \|v + x\|_S + \|v\|_{S^c} \geq \|x\|_S - \|v\|_S + \|v\|_{S^c},$$

where the last inequality follows by the reverse triangular inequality. This means that $\|v_S\|_1 \geq \|v_{S^c}\|_1$, but since $|S| \ll N$ it is unlikely for Φ to have vectors in its nullspace that are this concentrated on such few entries. This motivates the following definition.

Definition 10.5 (*Null Space Property*)

Φ is said to satisfy the s-Null Space Property if, for all $v \in \ker(\Phi)\backslash\{0\}$ (the nullspace of Φ) and all $|S| = s$ we have

$$\|v_S\|_1 < \|v_{S^c}\|_1.$$

In the argument above, we have shown that if Φ satisfies the Null Space Property for s, then x will indeed be the unique optimal solution to (44). In fact, the converse also holds

Theorem 10.6 (*NSP and ℓ_1 recovery*)

The following are equivalent for $\Phi \in \mathbb{K}^{M\times N}$:

1. For any s-sparse vector $x \in \mathbb{K}^N$, x is the unique optimal solution of (44) for $y = \Phi x$.
2. Φ satisfies the s-Null Space Property.

Challenge 10.3. *We proved* (1) $\Leftarrow$ (2) *in Theorem 10.6. Can you prove* (1) $\Rightarrow$ (2)*?*

We now prove the main Theorem of this section, which gives a sufficient condition for exact recovery via ℓ_1 minimization based on the notion of *worst-case coherence* of a matrix, or more precisely of its columns. We need first to introduce this notion.

Definition 10.7 (*Worst-case coherence*)

Given a set of vectors $\phi_1, \ldots, \phi_N \in \mathbb{K}^M$ such that $\|\phi_k\|_2 = 1$ for all $k \in [N]$ we call the worst-case coherence (sometimes also called dictionary coherence) the quantity

$$\mu := \max_{i\neq j} |\langle \phi_i, \phi_j \rangle|.$$

We call the worst-case coherence of a matrix the worst-case coherence of its column vectors.

We are now ready to state our main theorem:

Theorem 10.8 (*Low coherence and ℓ_1 recovery*)

If the worst-case coherence μ of a matrix Φ with unit norm column vectors satisfies

$$s < \frac{1}{2}\left(1 + \frac{1}{\mu}\right), \tag{46}$$

then Φ satisfies the s-NSP.

Proof of Theorem 10.8 – If $\mu = 0$ then the columns of Φ form an orthonormal basis of $\mathbb{R}^M$, thus $\ker(\Phi) = \emptyset$ and so it must satisfy the NSP for any s.

We now focus on $\mu > 0$. Let $v \in \ker(\Phi)\backslash\{0\}$ and $k \in [N]$, recall that ϕ_k is the k-th column of Φ, we have

$$\sum_{l=1}^{N} v_l \phi_l = 0,$$

and so $v_k\phi_k = -\sum_{l\neq k} v_l\phi_l$. Since $\|\phi_k\| = 1$ we have (recall $\phi_k^* = \overline{\phi_k}^\top$)

$$v_k = \phi_k^*\left(-\sum_{l\neq k} v_l\phi_l\right) = -\sum_{l\neq k} v_l(\phi_k^*\phi_l).$$

Thus,

$$|v_k| \leq \left| \sum_{l \neq k} v_l (\phi_k^* \phi_l) \right| \leq \mu \sum_{l \neq k} |v_l| = \mu(\|v\|_1 - |v_k|).$$

This means that for all $k \in [N]$ we have

$$(1+\mu)|v_k| \leq \mu \|v\|_1.$$

Finally, for $S \subset [N]$ of size s we have

$$\|v_S\|_1 = \sum_{k \in S} |v_k| \leq s \frac{\mu}{1+\mu} \|v\|_1 < \frac{1}{2} \|v\|_1,$$

where the last inequality follows from the hypothesis (46) of the Theorem. Since $\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1$ this completes the proof. □

In the next lectures we will study matrices with low worst-case coherence.

Remark 10.9. Different approaches are based on probability theory, and roughly follow the following path: since due to Theorem 10.6 recovery is formulated in terms of certain vectors not belonging to the nullspace of Φ, if one draws Φ from an ensemble of random matrices the problem reduces to understanding when a random subspace (the nullspace of the random matrix) avoids certain vectors, this is the subject of the celebrated "Gordon's Escape through a Mesh Theorem" (see [BSS23]), you can see versions of this approach also at [CRPW12] or, for an interesting approach based on Integral Geometry [ALMT14].

11 Finite frame theory and the Welch bound

Motivated by Theorem 10.8 we will now try to build low-coherence matrices. In order to do so we first introduce some basic elements of finite dimensional frame theory. For a reference on this topic, see for example the first chapter of the book [Chr16]. Recall that $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. We also change slightly notations with respect to the previous section: the usual vector x will live in $\mathbb{K}^d$ instead of $\mathbb{K}^N$, and we will denote $\phi_1, \cdots, \phi_N \in \mathbb{K}^d$ the frame vectors.

11.1 Finite frame theory

If $N = d$ and $\phi_1, \ldots, \phi_d \in \mathbb{K}^d$ are a basis then any point $x \in \mathbb{K}^d$ is uniquely identified by the inner products $c_k \coloneqq \langle x, \phi_k \rangle$. In particular if $\phi_1, \ldots, \phi_d \in \mathbb{K}^d$ form an orthonormal basis this representation satisfies a Parseval identity: $\|[\langle x, \phi_k \rangle]_{k=1}^d\| = \|x\|$. Using this identity on $x - y$ yields:

$$\left\| \{\langle x - y, \phi_k \rangle\}_{k=1}^d \right\| = \|x - y\| \qquad (\forall x, y \in \mathbb{K}^d). \tag{47}$$

This identity ensures *stability* in the representation: when we perturb x slightly, we only change slightly the inner products representation $\{\langle x, \phi_k \rangle\}_{k=1}^d$. But what about when the set of vectors $\{\phi_1, \cdots, \phi_N\}$ is not an orthonormal basis? In particular when $N > d$?

Redundancy – For instance, in signal processing and communication it is useful to include *redundancy*. Indeed, if instead of a basis one considers a "redundant" spanning set $\phi_1, \ldots, \phi_N \in \mathbb{K}^d$ with $m > d$ a few advantages arise: for example, if in a communication channel one of the coefficients b_k gets erased, it might still be possible to reconstruct x. Such sets are sometimes called **redundant dictionaries** or **overcomplete dictionaries**.

Stability – Still, it is important to keep some form of stability of the type of the Parseval identity (47). While this is particularly important for infinite dimensional vector spaces (more precisely Hilbert spaces) we will focus our exposition on finite dimensions.

Definition 11.1 (*Frame*)

A set $\phi_1, \ldots, \phi_N \in \mathbb{K}^d$ is called a frame of $\mathbb{K}^d$ if there exist constants $0 < A \leq B$ such that, for all $x \in \mathbb{K}^d$:

$$A\|x\|^2 \leq \sum_{k=1}^N |\langle x, \phi_k \rangle|^2 \leq B\|x\|^2.$$

A and B are called respectively the lower and upper frame bound. The largest possible value of A and the lowest possible value of B are called the optimal frame bounds.

Challenge 11.1. *Show that $\phi_1, \ldots, \phi_N \in \mathbb{K}^d$ is a frame if and only if it spans all of $\mathbb{K}^d$.*

Further Reading 11.2. In infinite dimensions the situation is considerably more delicate than suggested by Challenge 11.1, and it is tightly connected with the notion of stable sampling from signal processing. You can see, e.g., [Chr16].

Given a frame $\phi_1, \ldots, \phi_N \in \mathbb{K}^d$, let

$$\Phi \coloneqq \begin{bmatrix} | & & | \\ \phi_1 & \cdots & \phi_N \\ | & & | \end{bmatrix}. \tag{48}$$

The following are classical definitions in the frame theory literature (although for finite dimensions the objects are essentially just matrices involving Φ and so the definitions are not as important; also note that we are doing a slight abuse of notation using the same notation for a matrix and the linear operator it represents – it will be clear from context which object we mean.)

Definition 11.3

Given a frame $\phi_1, \dots, \phi_N \in \mathbb{K}^d$, we give the following definitions.

- The operator $\Phi : \mathbb{K}^N \to \mathbb{K}^d$ corresponding to the matrix Φ, meaning $\Phi(c) = \sum_{k=1}^N c_k \phi_k$, is called the *Synthesis Operator*.
- Its adjoint operator $\Phi^* : \mathbb{K}^d \to \mathbb{K}^N$ corresponding to the matrix $\Phi^* = \overline{\Phi}^\top$, meaning $\Phi^*(x) = \{\langle x, \phi_k \rangle\}_{k=1}^N$, is called the *Analysis Operator*.
- The self-adjoint operator $S : \mathbb{K}^d \to \mathbb{K}^d$ given by $S = \Phi\Phi^*$ is called the *Frame Operator*.

Challenge 11.2. *Show that $S \succeq 0$ and that S is invertible.*

The following are interesting (and useful) definitions:

Definition 11.4 (*Tight frame*)

A frame is called a tight frame if the frame bounds can be taken to be equal $A = B$.

Challenge 11.3. *What can you say about the Frame Operator S for a tight frame?*

We recall now the definition of worst-case coherence, which we already gave in the matrix setting, now in the language of frames (see Definition 10.7):

Definition 11.5 (*Worst-case coherence*)

(*i*) A frame $\phi_1, \dots, \phi_N \in \mathbb{K}^d$ is said to be unit normed (or unit norm) if for all $k \in [N]$ we have $\|\phi_k\| = 1$.

(*ii*) Given a unit norm frame $\phi_1, \dots, \phi_N \in \mathbb{K}^d$ we call the worst-case coherence (sometimes also called dictionary coherence) the quantity

$$\mu := \max_{i \neq j} |\langle \phi_i, \phi_j \rangle|.$$

Challenge 11.4. *In a very similar way one can define the spark of a frame as the spark of the matrix whose $i-th$ column is given by ϕ_i, see Definition 10.2. Can you give a relationship between the spark and the worst-case coherence of a frame?*

11.2 The Welch bound

Let us now come back to our original motivation: with Theorem 10.8 in mind, in this section we study the worst-case coherence of frames with the goal of understanding how much savings (in measurements) one can achieve with the technique described in Section 10. We start with a lower bound, due to Welch [Wel74].

Theorem 11.6 (*Welch Bound*)

Let $\phi_1, \dots, \phi_N \in \mathbb{K}^d$ be a unit norm frame, with $N \geq d$. Let μ be its worst case coherence

$$\mu = \max_{i \neq j} |\langle \phi_i, \phi_j \rangle| .$$

Then

$$\mu \geq \sqrt{\frac{N-d}{d(N-1)}}.$$

Proof of Theorem 11.6 – Note that we can assume $\mathbb{K} = \mathbb{C}$: indeed, the theorem for $\mathbb{K} = \mathbb{R}$ will then follow by simply viewing real vectors as elements of $\mathbb{C}^d$.

Let G be the Gram matrix of the vectors, $G_{ij} := \langle \phi_j, \phi_i \rangle = \phi_i^* \phi_j$. In other words, $G = \Phi^* \Phi$. It is positive semi-definite and its rank is at most d. Let $\lambda_1, \ldots, \lambda_d$ denote the largest eigenvalues of G, in particular this includes all non-zero ones. We have

$$(\mathrm{Tr}[G])^2 = \Big(\sum_{k=1}^{d} \lambda_k \Big)^2 \leq d \sum_{k=1}^{d} \lambda_k^2 = d \sum_{k=1}^{N} \lambda_k^2 = d \|G\|_F^2,$$

where the inequality follows from Cauchy-Schwarz between the vectors with the λ_k's and the all-ones vector. Note that since the vectors ϕ_i are unit normed, $\mathrm{Tr}(G) = \sum_{i=1}^{N} \|\phi_i\|^2 = N$, thus

$$\sum_{i,j=1}^{N} |\langle \phi_i, \phi_j \rangle|^2 = \|G\|_F^2 \geq \frac{1}{d} (\mathrm{Tr}[G])^2 = \frac{N^2}{d}.$$

Also,

$$\sum_{i,j=1}^{N} |\langle \phi_i, \phi_j \rangle|^2 = \sum_{i=1}^{N} |\langle \phi_i, \phi_i \rangle|^2 + \sum_{i \neq j}^{N} |\langle \phi_i, \phi_j \rangle|^2 = N + \sum_{i \neq j}^{N} |\langle \phi_i, \phi_j \rangle|^2 \leq N + (N^2 - N)\mu^2.$$

Putting everything together gives:

$$\mu \geq \sqrt{\frac{\frac{N^2}{d} - N}{(N^2 - N)}} = \sqrt{\frac{N - d}{d(N-1)}}.$$

□

Remark 11.7. Notice that in the proof above there were two inequalities used, if we track the cases when they are "equality" we can see for which frames the Welch bound is tight. The Cauchy-Schwarz inequality is tight when the vector consisting in the first d eigenvalues of G is a multiple of the all-ones vector, which is the case exactly when Φ is a Tight Frame (recall Definition 11.4). The second inequality is tight when all the terms in the sum $\sum_{i \neq j}^{N} |\langle \phi_i, \phi_j \rangle|^2$ are equal. The frames that satisfy these properties are called ETFs – Equiangular Tight Frames.

12 Equiangular Tight Frames (ETFs)

12.1 Definition and maximal size

In this section we continue the analysis started in the last sections, by studying equiangular tight frames, which are tight frames with the lowest possible worst-case coherence.

Definition 12.1 (*Equiangular Tight Frame*)

A unit-normed tight frame $\phi_1, \cdots, \phi_N \in \mathbb{K}^d$ is called an Equiangular Tight Frame (ETF) if there exists $\mu \geq 0$ such that, for all $i \neq j$,

$$|\langle \phi_i, \phi_j \rangle| = \mu. \tag{49}$$

Remark 12.2. Note that as described in Remark 11.7, the only possible value of μ for an ETF is given by the Welch bound, i.e. $\mu = \sqrt{(N-d)/(d[N-1])}$.

Proposition 12.3 (*Maximum size of an ETF*)

Let $\phi_1, \ldots, \phi_N$ by an equiangular tight frame in $\mathbb{K}^d$. Then:

- If $\mathbb{K} = \mathbb{C}$ then $N \leq d^2$.
- If $\mathbb{K} = \mathbb{R}$ then $N \leq \frac{d(d+1)}{2}$.

Proof of Proposition 12.3 – We start with a remark. Note that any real matrix $M \in \mathbb{R}^{d \times d}$ can be written as a vector, called $\mathrm{vec}(M) \in \mathbb{R}^{d^2}$, which collects all its entries[8]. Moreover

$$\langle \mathrm{vec}(M_1), \mathrm{vec}(M_2) \rangle = \mathrm{Tr}[M_1 M_2^\top].$$

Note that the vectors $\mathrm{vec}(M)$ for M symmetric actually live in a subspace of dimension $d(d+1)/2$. Similarly any complex matrix $M \in \mathbb{C}^{d \times d}$ can be written as a complex vector $\mathrm{vec}(M) \in \mathbb{C}^{d^2}$, and the inner product is $\langle \mathrm{vec}(M_1), \mathrm{vec}(M_2) \rangle = \mathrm{Tr}[M_1 M_2^*]$. And the vectors $\mathrm{vec}(M)$ for M Hermitian actually live in a *real* subspace of dimension d^2.

We now come back to the proof, both in the real and complex case. Let $\psi_i := \mathrm{vec}(\phi_i \phi_i^*)$. It is easy to check that these are unit-norm vectors in $\mathbb{K}^{d^2}$, and moreover, their inner products are (for $i \neq j$):

$$\psi_i^* \psi_j = \langle \mathrm{vec}(\phi_j \phi_j^*), \mathrm{vec}(\phi_i \phi_i^*) \rangle = \mathrm{Tr}((\phi_i \phi_i^*)(\phi_j \phi_j^*)^*) = |\langle \phi_i, \phi_j \rangle|^2 = \mu^2.$$

This means that their Gram matrix H is given by

$$H = (1-\mu^2)\mathrm{I}_N + \mu^2 \mathbf{1}\mathbf{1}^\top.$$

Since $\mu < 1$ we have $\mathrm{rk}(H) = N$. However, we can also write $H = \Psi^* \Psi$, for the matrix $\Psi \in \mathbb{K}^{d^2 \times N}$ whose i-th column is given by ψ_i. Therefore, $\mathrm{rk}(H) \leq \mathrm{rk}(\Psi)$ (here the rank means the dimension over $\mathbb{R}$ of the image space of the matrix). But as we saw in the remark above, the image space of Ψ has real dimension:

- For $\mathbb{K} = \mathbb{R}$, at most $\frac{1}{2} d(d+1)$.
- For $\mathbb{K} = \mathbb{C}$ at most d^2.

Thus $N \leq d^2$ for $\mathbb{K} = \mathbb{C}$, and $N \leq \frac{d(d+1)}{2}$ for $\mathbb{K} = \mathbb{R}$. □

Further Reading 12.4. Equiangular Tight Frames in $\mathbb{C}^d$ with $N = d^2$ are important objects in Quantum Mechanics, where they are called `SIC-POVM`: Symmetric, Informationally Complete, Positive Operator-Valued Measure. It is a central open problem to prove that they exist in all dimensions d, see Open Problem 6.3. in [Ban16] (the conjecture that they do exist is known as Zauner's Conjecture).

[8] Here it is not important how the indexing of the entries is done as long as consistent throughout.

12.2 First examples of low-coherence frames

Building ETFs with many vectors is a very non-trivial task. We will devote the next Section 21 to the construction of an ETF that arises from Number Theory, and you will also highlight in Exercice Class (if time permits) connections with spectral graph theory (a very nice example of how different fields of mathematics interact when studying "data science"!).

On the other hand, there are simple families of vectors with worst case coherence $\mu \sim 1/\sqrt{d}$.

Definition 12.5

An $d \times d$ real matrix H is called an Hadamard matrix if all entries are ± 1 and all columns are orthogonal.

The columns $(h_1, \cdots, h_d)$ of $\frac{1}{\sqrt{d}}H$ form an orthonormal basis on $\mathbb{R}^d$ (in other words, $\frac{1}{\sqrt{d}}H$ is an orthogonal matrix). Any h_k has an inner product of $|\langle h_k, e_l\rangle| = 1/\sqrt{d}$ with any element e_l of the canonical basis. This means that the $d \times 2d$ matrix

$$\Phi := \left[\mathrm{I}_d \;\; \frac{1}{\sqrt{d}}H\right] \tag{50}$$

has worst case coherence $1/\sqrt{d}$ (to be compared with the Welch bound of $1/\sqrt{2d-1}$ here). Theorem 10.8 guarantees then that, for Φ given by (50), ℓ_1 minimization achieves exact recovery for sparsity levels

$$s < \frac{1}{2}\left(1 + \sqrt{d}\right).$$

Exploratory Challenge 12.1. *We do not know for which values of d there exist $d \times d$ Hadamard matrices (See Example 12.6 for $d = 2^k$). It is relatively straightforward to show (try it!) that it must be that $4|d$. However, it is still an open problem (and a great one!) to show that for every dimension that is a multiple of 4, there exists an Hadamard matrix of that size (see Conjecture 14 in our blog [BKMR25], available at https://randomstrasse101.math.ethz.ch/)*

Example 12.6. There is an elegant known construction of Hadamard matrices for $d = 2^k$ due to Silverster. Let H_0 be the 1×1 matrix $H_0 = 1$. Construct H_k, for $k \geq 1$, recursively as the $2^k \times 2^k$ matrix given by

$$H_k = H_0 \otimes H_{k-1} = \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix}.$$

Challenge 12.2. *Show that H_k is a $2^k \times 2^k$ Hadamard matrix for all $k \geq 1$.*

For $\mathbb{C}$ it is easier to construct unitary matrices with entries having $\frac{1}{\sqrt{d}}$ modulus.

Example 12.7 (Discrete Fourier Transform)**.** We will see below (see Definition 20.1) the Discrete Fourier transform matrix $F \in \mathbb{C}^{d\times d}$:

$$F_{jk} = \frac{1}{\sqrt{d}} \exp\left[-2\pi i(j-1)(k-1)/d\right].$$

The columns $(F_1, \cdots, F_d)$ of F form an orthonormal basis on $\mathbb{C}^d$ (in other words, F is a unitary matrix, see Chalenge 20.1). Similarly to the Hadamard example,[9] any F_k has an inner product of $|\langle F_k, e_l\rangle| = 1/\sqrt{d}$ with any element e_l of the canonical basis. This means that the $d \times 2d$ matrix

$$\Phi := [\mathrm{I}_d \; F] \tag{51}$$

has worst case coherence $1/\sqrt{d}$.

[9] In fact, the Hadamard matrix construction above is the Fourier Transform in the group $(\mathbb{Z}/2\mathbb{Z})^{\otimes k}$ (also known as the Binary Fourier Transform), if you are interested in learning more on Fourier Transforms on other groups, do an online search on Fourier Analysis on Locally Compact Abelian Groups.

Remark 12.8. While the DFT construction above has a "redundancy" coefficient $N/d = 2$, there are many constructions of unit norm frames with low coherence, with redundancy coefficient much larger than 2. There is a all field of research involving these constructions, see for instance this article listing all constructions known in 2016 [FM15]. You can also take a look at the PhD thesis of Dustin Mixon [Mix12] which describes part of this field, and discusses connections to Compressed Sensing; Dustin Mixon also has a blog in part devoted to these questions [Mix]).

Exploratory Challenge 12.3 is meant to be solved later in the course, and shows that even randomly picked vectors do quite well (it requires some of the probability tools introduced later on).

Exploratory Challenge 12.3. *Towards the end of the course, equiped with a few more tools of probability (in particular concentration inequalities), you'll be able to show that by taking a frame made up of random (independent) vectors in the unit norm sphere, the coherence is comparable to the Welch bound. More precisely, this challenge is showing that N such vectors in d dimensions will have worst-case coherence* $\text{polylog}(N)/\sqrt{d}$*, where* $\text{polylog}(N)$ *means a polynomial of the logarithm of N (you will also work out the actual dependency).*

Further Reading 12.9. Challenge 12.3 along Theorem 10.8 show that for matrices consisting of random (independent) columns, sparse recovery with ℓ_1 minimization is possible up to sparsity levels $s \lesssim \sqrt{d}/\text{polylog}(N)$. It turns out that one can actually perform it for much larger levels of sparsity $s \lesssim d/\log(N)$, which up to logarithmic factors is the same as the number of measurements! Proving this however is outside the scope of this course, as it requires heavier probability Theory machinery. Interestingly, matching this performance with deterministic constructions seems notoriously difficult, in fact there is only one known construction "breaking the square-root bottleneck". You can read more about this in Open Problem 5.1. in [Ban16] (and references therein).

12.3 Mutually Unbiased Bases (MUBs)

Definition 12.10 (*Mutually Unbiased Bases*)

Construction (50) suggest the notion of *Mutually Unbiased Bases* (MUBs). Two orthonormal bases $v_1, \ldots, v_d$ and $u_1, \ldots, u_d$ of $\mathbb{C}^d$ are called mutually unbiased if for all i, j we have $|v_i^* u_j| = 1/\sqrt{d}$. A set of $k \geq 2$ bases is called mutually unbiased if the bases are pairwise mutually unbiased.

Challenge 12.4. *Show that a matrix formed with two orthonormal bases (such as (50)) cannot have worst case coherence smaller than $1/\sqrt{d}$. This motivates the definition above as the "most possibly unbiased" bases.*

Further Reading 12.11. Mutually Unbiased basis are an important object in quantum mechanics, communication, and signal processing, however there is still much that is not understood about them. A very nice and natural question to ask is: "what is the maximum number of bases that can be made mutually unbiased in $\mathbb{C}^d$?" Let us denote this number $\mathcal{M}(d)$. Remarkably, very little is known about $\mathcal{M}(d)$, besides a general bound $\mathcal{M}(d) \leq d+1$, and that this bound is achievable if d is the power of a prime number.

Exploratory Challenge 12.5 (Open Problem)**.** *How many mutually unbiased bases exist in $d = 6$ dimensions ? The best known upper bound is $\mathcal{M}(6) \leq 6+1 = 7$ (see above), and the best known lower bound is $\mathcal{M}(6) \geq 3$. See Open Problem 6.2. in [Ban16].*

13 Solving contaminated linear systems

In this section we will illustrate the power of redundancy and ℓ_1 minimization to solve a linear system where the measurement vector has suffered corruptions. Let $w \in \mathbb{C}^d$ be a signal of interest, and Φ a $d \times N$ full row rank matrix whose columns are a frame $\phi_1, \ldots, \phi_N \in \mathbb{C}^d$ (Φ is the Analysis Operator of the frame).[10]

We are interested in a situation in which we receive linear measurements $c = \Phi^* w$ (note that $N \geq d$ so these measurements are over-determined, not under-determined like in Compressed Sensing) but for which part of these measurements have been corrupted, without any assumptions on the corruption, except for the fact that it is only on at most s entries of $\Phi^* w \in \mathbb{C}^N$. We can model this by receiving a vector $c \in \mathbb{C}^N$ given by

$$c = \Phi^* w + e,$$

where e represents the corruptions/errors, we know that $\|e\|_0 \leq s$, but it is otherwise arbitrary and unknown. We want to recover w. If we can recover e then we can simply solve $\Phi^* w = c - e$ (which we can because that system is over-determined). Henceforth, we will focus on recovering e.

In 2005, Candes and Tao [CT05] proposed the following approach: The idea is to take a matrix Ψ whose nullspace is the same as (or at least contains) the column space of Φ^*. In that case $\Psi\Phi^* = 0$. This means that $\Psi c = \Psi\Phi^* w + \Psi e$ and so $\Psi c = \Psi e$. Thus we can try to recover e by solving

$$\begin{array}{ll} \min & \|h\|_1 \\ \text{s.t.} & \Psi h = \Psi c, \end{array} \tag{52}$$

We know that if $\|e\|_0 \leq s$ and Ψ satisfies the s-NSP then e is the unique optimizer of (52).

Definition 13.1 (*Notation for column and row space of a matrix*)

Let $\mathrm{col}(A)$ and $\mathrm{row}(A)$ denote, respectively the column and row space of a matrix A.

Since we can take Ψ such that $\ker(\Psi) = \mathrm{col}(\Phi^*)$ we immediately get the following theorem.

Theorem 13.2

Let Φ be a $d \times N$ matrix with full row rank and let $s < d$. If, for all $v \in \mathrm{col}(\Phi)$ and all $S \subset [N]$ such that $|S| \leq s$, we have $\|v_S\|_1 < \|v_{S^c}\|_1$, then the procedure described above exactly recovers w from $c = \Phi^* w + e$, where e are arbitrary s-sparse "corruptions".

Further Reading 13.3. In the same spirit as in Further Reading 12.9 it is possible to handle a linear fraction of corruptions using random constructions (and analysis based on Probability Theory). In this section we will focus on deterministic constructions based on bounding worst-case coherence, as in previous sections.

We will focus on the case in which Φ is a Unit-norm tight frame (UNTF) for the sake of exposition. One way to build Ψ is via what is known as the Naimark Complement [CFM+13], let us describe the construction.

Henceforth we assume $N > d$ (avoiding the degenerate case $N = d$). In Challenge 11.3 you showed that $\Phi\Phi^* = \frac{N}{d}\mathrm{I}_{d\times d}$.[11] This means that that the matrix $\sqrt{\frac{d}{N}}\Phi$ has d orthonormal columns in $\mathbb{C}^N$ (or $\mathbb{R}^N$ if working over the reals). It is not hard to see (e.g. by Gram-Schmidt) that we can extend it (by adding columns) into an $N \times N$ unitary matrix. In order words, that there exists $\sqrt{\frac{N-d}{N}}\Psi \in \mathbb{C}^{(N-d)\times N}$

[10] We give the exposition for the complex case, but you can focus in the real case and simply replace Φ^* by Φ^T throughout this Section if you prefer.

[11] If you simply noted that it is a multiple of the identity matrix, but did not compute the coefficient in front, you can do it, for example by noticing that $\mathrm{Tr}(\Phi\Phi^*) = \Phi^*\Phi$, or by computing $\|\Phi\|_F^2$.

such that

$$U = \begin{bmatrix} \sqrt{\frac{d}{N}}\Phi \\ \sqrt{\frac{N-d}{N}}\Psi \end{bmatrix},$$

where the $\sqrt{\frac{N-d}{N}}$ scaling was picked with foresight.[12] is an $N \times N$ unitary matrix, with orthonormal columns $u_1, \ldots, u_N$.

Notice that, for any $i \in [N]$,

$$1 = \|u_i\|^2 = \frac{d}{N}\|\phi_i\|^2 + \frac{N-d}{N}\|\psi_i\|^2 = \frac{d}{N} + \frac{N-d}{N}\|\psi_i\|^2,$$

where $\psi_1, \ldots, \psi_N$ are the columns of Ψ. This shows that the columns of Ψ are unit-normed.

Note also that

$$\mathrm{I}_{N\times N} = UU^* = \begin{bmatrix} \frac{d}{N}\Phi\Phi^* & \sqrt{\frac{d}{N}}\sqrt{\frac{N-d}{N}}\Phi\Psi^* \\ \sqrt{\frac{d}{N}}\sqrt{\frac{N-d}{N}}\Psi\Phi^* & \frac{N-d}{N}\Psi\Psi^* \end{bmatrix},$$

which implies that $\Psi\Phi^* = 0$ and that $\Psi\Psi^* = \frac{N}{N-d}\mathrm{I}_{(N-d)\times(N-d)}$. This shows that Ψ is a unit-norm tight frame and that its kernel coincides with the column space of Φ.

Let us now compute the worst-case coherence of Ψ. Let $i \neq j \in [N]$, we have

$$\frac{N-d}{N}\langle \psi_i, \psi_j\rangle = \langle u_i, u_j\rangle - \frac{d}{N}\langle \phi_i, \phi_j\rangle = -\frac{d}{N}\langle \phi_i, \phi_j\rangle.$$

Thus $\mu(\Psi)$, the worst-case coherence of Ψ satisfies $\frac{N-d}{N}\mu(\Psi) = \frac{d}{N}\mu(\Phi)$, or equivalently $\mu(\Psi) = \frac{d}{N-d}\mu(\Phi)$, where $\mu(\Phi)$ denotes the worst-case coherence of the frame Φ.

Theorem 10.8 guarantees that as long as $s \leq \frac{1}{2}\left(1 + \frac{1}{\mu(\Psi)}\right)$, then Ψ satisfies the s-NSP. This implies the following theorem.

Theorem 13.4

Let Φ be a $d \times N$ whose columns are a unit-norm tight frame, let $s < d < N$. If

$$s \leq \frac{1}{2d}\left(d + \frac{N-d}{\mu(\Phi)}\right),$$

then the procedure described above exactly recovers w from $c = \Phi^* w + e$, where e are arbitrary s-sparse "corruptions".

Challenge 13.1. *Work out how many corruptions are allowed in Examples 12.6 and 12.7Show that a Naimark complement of an ETF is also an ETF.*

Exploratory Challenge 13.2. *Can you come up with interesting examples where $N \neq 2d$?*

Exploratory Challenge 13.3. *Can you work out what are the regimes of s, d, N allowed when Φ has random columns (similarly to Exploratory Challenge 12.3).*

Challenge 13.4. *Show that a Naimark complement of an ETF is also an ETF.*

Challenge 13.5. *(When) is the procedure above the same as atempting to recover w as the optimal solution of (53)?*

$$\min_{z\in\mathbb{C}^d} \quad \|\Phi^* z - c\|_1. \tag{53}$$

(compare with the classical least squares method, which would correspond to $\min_{z\in\mathbb{C}^d} \|\Phi^* z - c\|_2$).

[12] As you will see, this scaling will make it so that Ψ will have unit-norm columns, and needs not be normalized later on.

14 Elements of classification theory

A lot of these chapters on classification theory were adapted from chapters written by Nikita Zhivotovskiy in [BZ22]. The errors/typos are all mine.

A bit of history – In this last part of the lectures, which we will cover for approximately three weeks, we introduce some of the fundamentals of learning theory. We start with the theory of classification, a foudational topic in Statistical Machine Learning. The direction we are discussing in this part of the course was initiated by Vladimir Vapnik and Alexey Chervonenkis in the mid-60s and independently by Leslie Valiant in the mid-80s. The results of Vapnik and Chervonenkis led to what we now call Vapnik–Chervonenkis Theory. From the early 70s to the present day, their work has an ongoing impact on Machine Learning, Statistics, Empirical Process Theory, Computational Geometry and Combinatorics. In parallel, the work of Valiant looked at a similar problem from a more computational perspective. In particular, Valiant developed the theory of Probably Approximately Correct (PAC) Learning that led, among other contributions, to his 2010 Turing Award.

Classification model – The statistician is given a sequence of independent identically distributed observations

$$X_1, \dots, X_n$$

taking values in $\mathcal{X} \subseteq \mathbb{R}^p$, each distributed according to some unknown distribution P[13]. In practice, X_i might be seen as an image or a feature vector. For example, consider the problem of health diagnostics. In this case these vectors can describe some medical information such as age, weight, blood pressure and so on. An important part of our analysis is that the dimension of the space will not play any role, and classification is possible even in abstract measurable spaces.

Contrary to e.g. clustering tasks we have considered previously, classification models belong to the realm of *supervised learning*, meaning that the statistician also observe labels associated to the observations

$$f^*(X_1), \dots, f^*(X_n),$$

where f^* is a (unknown to her) *target* classifier[14] mapping $\mathcal{X} \to \{0, 1\}$. These labels will depend on the application, they can represent cat/dogs when classifying images, spam/not spam when classifying ls, disease/no disease when diagnosing a patient... Moreover, we restrict the classifier to have value in $\{0, 1\}$, but what we will describe can be generalized for a finite number of classes.

Classification task – Using the labeled sample

$$S_n = \{(X_1, f^*(X_1)), \dots, (X_n, f^*(X_n))\}, \tag{54}$$

the statistician's aim is to construct a measurable classifier $\widehat{f} : \mathcal{X} \to \{0, 1\}$ that can be used to classify any element $x \in \mathcal{X}$, e.g. a new image. The *risk* (or the *error*) of a classifier $f : \mathcal{X} \to \{0, 1\}$ is defined by the probability of making an error on a random sample:

$$R(f) \coloneqq \mathbb{P}(f(X) \neq f^*(X)),$$

where $X \sim P$. With this definition in mind, the statistician wants to find a classifier that has risk as small as possible. Besides the labeled sample S_n, a second important information is available to her: f^* belongs to some known class $\mathcal{F}$ of (measurable) classifiers[15] mapping $\mathcal{X}$ to $\{0, 1\}$.

Definition 14.1 (*Consistent classifier* –)

We say that a classifier $\widehat{f} : \mathcal{X} \to \{0, 1\}$ is *consistent* with the sample S_n if for all $i \in \{1, \cdots, n\}$:

$$\widehat{f}(X_i) = f^*(X_i).$$

[13] We assume that there is a probability space $(\mathcal{X}, F, P)$, where F is a Borel σ-algebra.

[14] Moreover, we always assume standard measurability assumption on f^* so that e.g., $\{f^*(x) = 1\}$ is measurable.

[15] Note that in the Computer Science literature these classifiers are sometimes called *concepts*.

Performance of consistent classifiers – Which strategy could the statistican adopt? Given that the sample S_n is basically the only information in her possession, the most natural way is to choose $\widehat{f} \in \mathcal{F}$ consistent with S_n and use it as a guess, hoping that it will be close to the true classifier f^*. Hopefully, if the number of samples n is large enough, this will be true. However, since the sample S_n is random, we cannot guarantee this with certainty. Instead, we may only say that $\widehat{f}$ is close to f^* *with high probability*: this would mean intuitively that for a large fraction of all random realizations of the sample S_n, any classifier consistent with a particular realization of the sample has a small risk.

Theorem 14.2 (*Risk of consistent classifiers* –)

Assume that $f^* \in \mathcal{F}$, and that $\mathcal{F}$ is finite. For the confidence parameter $\delta \in (0,1)$ and the precision parameter $\varepsilon \in (0,1)$, assume that we have

$$n \geq \left\lceil \frac{\log |\mathcal{F}|}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta} \right\rceil.$$

Then (the probability being over the law of $X_1, \cdots, X_n$):

$$\mathbb{P}(\forall \widehat{f} \in \mathcal{F} \text{ consistent with the sample } S_n \,:\, R(\widehat{f}) < \varepsilon) \geq 1 - \delta. \tag{55}$$

Equivalently, with probability at least $1-\delta$, any classifier f such that $R(f) > \varepsilon$ cannot be consistent with the sample S_n.

Proof of Theorem 14.2 – Let us denote $\mathcal{F}_\varepsilon := \{f \in \mathcal{F} \,:\, R(f) \geq \varepsilon\} \subseteq \mathcal{F}$, and fix any $f \in \mathcal{F}_\varepsilon$. If no such function exists, the claim follows. By independence of $(X_i)_{i=1}^n$:

$$\begin{aligned}
\mathbb{P}[f \text{ is consistent with } S_n] &= \mathbb{P}(f(X_i) = f^*(X_i) \text{ for } i = 1, \ldots, n) \\
&= \prod_{i=1}^n \mathbb{P}_{X_i}(f(X_i) = f^*(X_i)) \\
&= \prod_{i=1}^n (1 - \mathbb{P}_{X_i}(f(X_i) \neq f^*(X_i))) \\
&\leq (1-\varepsilon)^n \\
&\leq \exp(-n\varepsilon),
\end{aligned} \tag{56}$$

where in the last line we used $1 - x \leq \exp(-x)$. We recall the union bound:

Proposition 14.3 (*Union bound*)

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. For any countable sequence of events $(A_n)_{n\geq 1} \in \mathcal{F}$:

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n \right) \leq \sum_{n \geq 1} \mathbb{P}(A_n).$$

By Proposition 14.3 and eq. (56) we have

$$\begin{aligned}
&\mathbb{P}(\text{there is } f \in \mathcal{F} \text{ with } R(f) \geq \varepsilon \text{ and such that } f(X_i) = f^*(X_i) \text{ for } i = 1, \ldots, n) \\
&= \mathbb{P}\left(\bigcup_{f \in \mathcal{F}_\varepsilon} \{f(X_i) = f^*(X_i) \text{ for } i = 1, \ldots, n\} \right) \\
&\leq \sum_{f \in \mathcal{F}_\varepsilon} \mathbb{P}(f(X_i) = f^*(X_i) \text{ for } i = 1, \ldots, n) \\
&\leq |\mathcal{F}_\varepsilon| \exp(-n\varepsilon) \leq |\mathcal{F}| \exp(-n\varepsilon),
\end{aligned} \tag{57}$$

Since we want this probability to be smaller than δ, we see that if

$$n \geq \left\lceil \frac{\log |\mathcal{F}|}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta} \right\rceil,$$

then by eq. (57) we can guarantee that, with probability at least $1-\delta$, any classifier $f \in \mathcal{F}$ consistent with the samples has its risk smaller than ε. □

Remark 14.4 (Risk bound)**.** One may rewrite the result of Theorem 14.2 as a *risk bound*. That is, we first fix the sample size n and want to estimate the precision ε of any consistent classifier. More precisely, eq. (55) implies that

$$\mathbb{P}\left(\sup_{\substack{\widehat{f} \in \mathcal{F} \\ \widehat{f} \text{ consistent with } S_n}} R(\widehat{f}) \leq \frac{\log |\mathcal{F}|}{n} + \frac{1}{n} \log \frac{1}{\delta} \right) \geq 1 - \delta.$$

PAC learnability The result of Theorem 14.2 inspires the following definition. In what follows, PAC stands for *Probably Approximately Correct*. Indeed, we showed that for any finite class $\mathcal{F}$, any consistent classifier is approximately correct (i.e. has risk $\leq \varepsilon$) with high probability.

Definition 14.5 (*PAC learnability* –)

A (possibly infinite) class $\mathcal{F}$ of classifiers is *PAC-learnable* with the *sample complexity* $n(\delta, \varepsilon)$ if there is a mapping

$$A : \bigcup_{m=0}^{\infty} (\mathcal{X} \times \{0,1\})^m \to \{0,1\}^{\mathcal{X}}$$

called the *learning algorithm* (given a sample S of any size it outputs a classifier $A(S)$) that satisfies the following property: for

(*i*) *every* distribution P on $\mathcal{X}$,

(*ii*) *every* $\delta, \varepsilon \in (0,1)$, and

(*iii*) *every* target classifier $f^* \in \mathcal{F}$,

if the sample size n is greater or equal than $n(\delta, \varepsilon)$, then

$$\mathbb{P}_{(X_1,\ldots,X_n)} \left(R(A(S_n)) \leq \varepsilon \right) \geq 1 - \delta.$$

Remark 14.6 (Measurability of classifiers)**.** When considering finite classes we have little problems with measurability and we may only request that for all $f \in \mathcal{F}$ the set $\{f(x) = 1\}$ is measurable. The notion of PAC-learnability allows infinite classes. In this case the question of measurability is more subtle. However, as a rule of thumb, one may argue that measurability issues will almost never appear in the analysis of real-life algorithms. In particular, starting from the late 80's there is a useful and formal notion of *well-behaved* classes: these are essentially the classes for which these measurability issues do not appear. See also Remark 16.4 in Section 16.

An immediate outcome of Theorem 14.2 is the following result.

Corollary 14.7 (*PAC learnability of finite classes*)

Any finite class $\mathcal{F}$ is PAC learnable with the sample complexity

$$n(\delta, \varepsilon) = \left\lceil \frac{\log |\mathcal{F}|}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta} \right\rceil .$$

Moreover, to learn this class, we simply need to output any consistent classifier $\widehat{f}$.

An important limitation of Theorem 14.2 is that it only deals with finite classes. Working only with discrete spaces of solutions is somewhat impractical: many modern machine learning techniques are based on optimization (e.g. gradient descent) methods that require that the class $\mathcal{F}$ is parametrized in a relatively smooth way by $\mathbb{R}^p$. One of our main goals for the remaining of the class will therefore be to see what we can say about PAC learnability of possibly infinite class of functions, and will culminate with the characterization of PAC learnability via a property known as the Vapnik-Chervonenkis dimension in Section 17. First we will need to introduce an important result of probability theory called Hoeffding's inequality.

15 Hoeffding's inequality

In this section, we prove a very useful bound for the sum of many i.i.d. random variables. In Section 15.2 (which is not covered in the lectures), we show that this allows to prove that randomized constructions of frames have low coherence. Let us recall first a classical result of probability theory:

Proposition 15.1 (*Markov's inequality*)

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and let $X : \Omega \to \mathbb{R}$ be a nonnegative random variable. Then for all $t > 0$:

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

An immediate consequence of Markov's inequality (by applying it to $|X - \mathbb{E}X|^2$) is Chebyshev's inequality:

Proposition 15.2 (*Chebyshev's inequality*)

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and let $X : \Omega \to \mathbb{R}$ be a random variable with finite mean $\mathbb{E}[X]$. Then for all $t > 0$:

$$\mathbb{P}[|X - \mathbb{E}X| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}.$$

15.1 Concentration inequalities

Given a random variable X, we define the real-valued function M_X (also called moment generating function, or MGF) as

$$M_X(\lambda) := \mathbb{E}\exp(\lambda X),$$

whenever this expectation exists. For example, it is standard to verify that if X is distributed according to the normal law with mean 0 and variance σ^2, then for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}\exp(\lambda X) = \exp(\lambda^2\sigma^2/2).$$

We now show that a similar upper bounds holds for any zero-mean bounded random variable. This result is originally due to Wassily Hoeffding. In this proof we use Jensen's inequality.

Proposition 15.3 (*Jensen's Inequality*)

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function and X is a random variable for which $\mathbb{E}X$ and $\mathbb{E}\varphi(X)$ exist. Then $\varphi(\mathbb{E}X) \leq \mathbb{E}\,\varphi(X)$.

Since φ is a convex function, it has supporting lines, i.e. for all $\alpha \in \mathbb{R}$ there exists $\beta \in \mathbb{R}$, and a line $\beta(x-\alpha)$ such that $\varphi(\alpha)+\beta(x-\alpha) \leq \varphi(x)$ for all $x \in \mathbb{R}$. Taking $\alpha = \mathbb{E}X$ we have $\varphi(\mathbb{E}X)+\beta(X-\mathbb{E}X) \leq \varphi(X)$ and hence

$$\mathbb{E}\varphi(X) \geq \mathbb{E}\big[\varphi(\mathbb{E}X) + \beta(X - \mathbb{E}X)\big] = \varphi(\mathbb{E}X).$$

□

Lemma 15.4 (*Hoeffding's lemma*)

Let X be a zero mean random variable ($\mathbb{E}X = 0$) such that $X \in [a, b]$ almost surely. Then for all $\lambda \in \mathbb{R}$:

$$M_X(\lambda) \leq \exp(\lambda^2(b-a)^2/8).$$

Remark 15.5. Random variables X such that the MGF of $X - \mathbb{E}X$ is upper bounded by the MGF of a Gaussian random variable are usually called *sub-Gaussian* random variables. A consequence of Lemma 15.4 is that all bounded random variables are sub-Gaussian.

In the lecture, we will prove a weaker version of Lemma 15.4, in which the denominator 8 is replaced by 2. For completeness, we include the proof of the weaker upper bound (proven in the class) and of the stronger result.

Proof of Lemma 15.4 (weaker) – This weaker proof is interesting because it uses the idea of *symmetrization*, which is often useful in probability theory, and which we will encouter in Section 16. Let us denote X' an independent copy of the random variable X, and $\mathbb{E}'$ the expectation with respect to X' only. Since $\mathbb{E}X = 0$, we have

$$M_X(\lambda) = \mathbb{E}\exp(\lambda(X - \mathbb{E}'X')) = \mathbb{E}\exp(\lambda\mathbb{E}'[X - X']) \overset{\text{(a)}}{\leq} \mathbb{E}\mathbb{E}'\exp(\lambda[X - X']).$$

We used Jensen's inequality in (a). Note that $X - X'$ is a symmetric random variable, thus its distribution is equal to the one of

$$X - X' \overset{\mathrm{d}}{=} \varepsilon(X - X'),$$

in which $\varepsilon \in \{\pm 1\}$ is a Rademacher random variable with $\mathbb{P}[\varepsilon = 1] = 1/2$, independent of (X, X'). Thus we have (writing now $\mathbb{E}$ for the expectation with respect to all X, X' and ε):

$$M_X(\lambda) \leq \mathbb{E}\exp(\varepsilon\lambda[X - X']).$$

This is the core idea of the symmetrization method, as we can now inverse the order of expectation by Fubini's theorem, and perform first the expectation over ε:

$$M_X(\lambda) \leq \mathbb{E}\left\{\frac{1}{2}\Big[\exp(\lambda[X - X']) + \exp(-\lambda[X - X'])\Big]\right\} \leq \mathbb{E}\exp(\lambda^2[X - X']^2/2),$$

since $\cosh(x) \leq \exp(x^2/2)$ for all $x \in \mathbb{R}$. The proof is now over since $|X - X'|^2 \leq (b - a)^2$ because $a \leq X, X' \leq b$. □

Proof of Lemma 15.4 (original formulation) – Since $x \mapsto e^{\lambda x}$ is a convex function, for all $x \in [a, b]$ we have:

$$e^{\lambda x} \leq \frac{b - x}{b - a}e^{\lambda a} + \frac{x - a}{b - a}e^{\lambda b}.$$

By taking expectations, we have:

$$\begin{aligned} M_X(\lambda) &\leq \frac{b}{b - a}e^{\lambda a} - \frac{a}{b - a}e^{\lambda b}, \\ &\leq e^{\lambda a}\left[1 + \frac{a}{b - a}(1 - e^{\lambda(b-a)})\right], \\ &\leq e^{F[\lambda(b-a)]}, \end{aligned}$$

in which $F(x) \coloneqq ax/(b - a) + \log[1 + a(1 - e^x)/(b - a)]$. In particular, $F(0) = 0$, $F'(0) = 0$, and $F''(x) = -abe^x/(b - ae^x)^2$. Note that since $\mathbb{E}X = 0$, $a \leq 0$ and $b \geq 0$. Therefore the AMGM inequality[16] yields $b - ae^x \geq 2\sqrt{-abe^x}$, and thus $F''(x) \leq 1/4$. We can then use Taylor's expansion around 0 and bound the remainder, which yields that for all $x \in \mathbb{R}$,

$$F(x) \leq F(0) + xF'(0) + \frac{x^2}{2}\sup_{h\in\mathbb{R}} F''(h) \leq \frac{x^2}{8}.$$

[16] For any $x, y \geq 0$, $\sqrt{xy} \leq (x + y)/2$.

Applying this for $x = \lambda(b-a)$ ends the proof. □

Let Y be a random variable. Denote its moment generating function by M_Y. For any $t \in \mathbb{R}$ and $\lambda > 0$, we have

$$\mathbb{P}(Y \geq t) = \mathbb{P}(\lambda Y \geq \lambda t) = \mathbb{P}(\exp(\lambda Y) \geq \exp(\lambda t)) \leq \exp(-\lambda t) M_Y(\lambda),$$

where the last inequality follows from Markov's inequality (Proposition 15.1). Therefore, we have

$$\mathbb{P}(Y \geq t) \leq \inf_{\lambda > 0} \{\exp(-\lambda t) M_Y(\lambda)\}.$$

This very useful upper bound is usually called the Chernoff method. We are now ready to prove the basic concentration inequality for bounded random variables.

Theorem 15.6 (*Hoeffding's inequality*)

Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ almost surely for $i = 1, \ldots, n$. Then, for any $t \geq 0$, it holds that

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Moreover,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i - \mathbb{E}X_i\right| \geq t\right) \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Proof of Theorem 15.6 – We proceed with the following lines. For any $\lambda \geq 0$, it holds that

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right) &\leq \exp(-\lambda t)\, \mathbb{E}\exp\left(\lambda \sum_{i=1}^n (X_i - \mathbb{E}X_i)\right) \quad \text{(by the Chernoff method)} \\
&= \exp(-\lambda t) \prod_{i=1}^n \mathbb{E}\exp(\lambda(X_i - \mathbb{E}X_i)) \quad \text{(by independence)} \\
&\leq \exp(-\lambda t) \prod_{i=1}^n \exp\left(\frac{\lambda^2}{8}(b_i - a_i)^2\right) \quad \text{(by Hoeffding's lemma)} \\
&= \exp\left(-\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right). \qquad (58)
\end{aligned}$$

Observe that we used that the length of the interval to which $X_i - \mathbb{E}X_i$ belongs is the same as the corresponding length for X_i. We can now choose $\lambda \geq 0$ so as to minimize the right-hand side of eq. (58). One checks easily that the optimal choice is $\lambda = 4t/\sum_{i=1}^n (b_i - a_i)^2$, which proves the first inequality. To prove the second inequality, notice that we can apply the first inequality to $Y_i = -X_i$, which yields

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq -t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Finally, by the union bound (Proposition 14.3)

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{i=1}^n X_i - \mathbb{E}X_i\right| \geq t\right) &\leq \mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right) + \mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq -t\right) \\
&\leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).
\end{aligned}$$

□

Example 15.7. Assume that $a_i = a$ and $b_i = b$ for all $i \in [n]$. Then Hoeffding's inequality gives:

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

In particular, the right-hand side goes to 0 as $n \to \infty$ if $t = \omega(\sqrt{n})$[17]: informally, we see that the sum can not fluctuate by more than $\mathcal{O}(\sqrt{n})$, coherently with the picture given by the central limit theorem!

Further Reading 15.8. The idea of deducing concentration inequalities by using upper bounds on the moment generating function is very fruitful, and is used to prove a large part of classical concentration inequalities. Hoeffding's inequality appears in the foundational work of Hoeffding [Hoe63]. Similar techniques were used in 1920-s by Bernstein and Kolmogorov [Kol29].

15.2 Sidenote: Randomized low-coherence frames

An easy corollary of Hoeffding's inequality 15.6 shows that a set of random vectors with i.i.d. coordinates in $\{\pm 1\}$ has a low worst-case coherence. We will not show the following corollary in the lecture, but it is a classical application of combining a strong concentration inequality with the union bound, a very versatile idea in probability theory, as you will see in the next two sections!

Corollary 15.9 (*Random low-coherence frame*)

Let $d, m \geq 1$. Let $\phi_1, \cdots, \phi_m \in \mathbb{R}^d$ be i.i.d. draws of the vector $X \in \mathbb{R}^d$ drawn with the following distribution: the entries $(X_k)_{k=1}^d$ of X are i.i.d. and $\mathbb{P}[X_k = 1/\sqrt{d}] = \mathbb{P}[X_k = -1/\sqrt{d}] = 1/2$. Notice that $\|\phi_i\|_2 = 1$ almost surely for all $i \in [m]$. Moreover for all $t \geq 0$:

$$\mathbb{P}[\max_{i \neq j} |\langle \phi_i, \phi_j \rangle| \geq t] \leq m(m-1)\exp\Big\{-\frac{dt^2}{2}\Big\}. \tag{59}$$

Therefore, for any $\delta \in (0, 1)$:

$$\mathbb{P}\left[\max_{i \neq j} |\langle \phi_i, \phi_j \rangle| \leq \sqrt{\frac{2}{d}\log\frac{m(m-1)}{\delta}}\right] \geq 1 - \delta. \tag{60}$$

Proof of Corollary 15.9 – Notice that, for any fixed $i \neq j$:

$$\langle \phi_i, \phi_j \rangle = \sum_{k=1}^{d} (\phi_i)_k (\phi_j)_k.$$

Since $i \neq j$, the random variables $Y_k := (\phi_i)_k(\phi_j)_k$ are i.i.d. random variables, with zero mean, and $\mathbb{P}[Y_k = 1/d] = \mathbb{P}[Y_k = -1/d] = 1/2$. We can thus apply Hoeffding's inequality 15.6 to get

$$\mathbb{P}[|\langle \phi_i, \phi_j \rangle| \geq t] \leq 2\exp\Big\{-\frac{dt^2}{2}\Big\}.$$

Notice that there are $m(m-1)/2$ different inner products $|\langle \phi_i, \phi_j \rangle|$. By applying the union bound (Proposition 14.3) we thus get eq. (59):

$$\mathbb{P}[\max_{i \neq j} |\langle \phi_i, \phi_j \rangle| \geq t] \leq m(m-1)\exp\Big\{-\frac{dt^2}{2}\Big\}.$$

Eq. (60) can be easily deduced by letting $t = \sqrt{(2/d)\log[m(m-1)/\delta]}$. □

[17] With the classical notation $x = \omega(y) \Leftrightarrow y = o(x)$.

Remark 15.10. For large d and m, Corollary 15.9 shows that a random frame with $\pm 1/\sqrt{d}$ coordinates has, with a large probability, a worst-case coherence $\mu \sim \sqrt{(4 \log m)/d}$. In particular, by Theorems 10.6 and 10.8, Φ will be a suitable matrix for ℓ_1 recovery of s-sparse vectors as long as $s \lesssim \sqrt{d/\log m}$. Actually (see Further reading 12.9), they are known to work as long as $s \lesssim d/\log(m)$.

Remark 15.11. If, for instance, we pick $m = \alpha d$ (with a fixed $\alpha \geq 1$), then as $d \to \infty$, we have $\mu \sim \sqrt{4(\log d)/d}$, while the Welch bound is $\mu_{\min} \sim \sqrt{(1 - \alpha^{-1})/d}$. Random frames thus satisfy the Welch bound up to a factor $\mathcal{O}(\sqrt{\log d})$ in this setting[18].

Further Reading 15.12. Corollary 15.9 can be shown to hold for much more general distributions than random $\pm 1/\sqrt{d}$ coordinates. In particular, similar results (perhaps up to some multiplicative constants) hold for random vectors on the unit sphere, or i.i.d. vectors with distributions whose tail decay at least as fast as a Gaussian (called *sub-Gaussian* distributions). This is related to concentration results (like Hoeffding's inequality) holding as well for these different cases, see e.g. a famous textbook on high-dimensional probability [Ver18].

[18] More generally, they satisfy it up to a factor $\mathcal{O}(\sqrt{(1 - d/m) \log m})$.

16 Uniform convergence and the Vapnik-Chervonenkis Theorem

In this chapter, we use Hoeffding's inequality to prove the *Vapnik-Chervonenkis theorem*. Fundamentally, it shows the uniform convergence of frequencies of events to their probabilities, but in the context of classification theory, it will allow us to generalize Theorem 14.2 to possibly infinite classes, by union bounding over a set whose cardinality might be much smaller than the whole class (since the union bound over the whole class of eq. (57) fails for infinite classes $\mathcal{F}$). In practice, we will obtain a theorem very close to Theorem 14.2, replacing the size $|\mathcal{F}|$ of the class by a quantity known as the *growth function* of this class.

Finally, in the final Section 17 of the lecture, we will see that the growth function can be bounded by a quantity that is easier to interpret, and that is known as the Vapnik-Chervonenkis (VC) dimension. As many infinite classes have finite VC dimension, this significantly generalizes the results of Section 14 to infinite classes.

16.1 Motivation and statement of the VC theorem

We take again the setup of Section 14, in which we are trying to learn a classifier $f^\star \in \mathcal{F}$, but now we do not assume that $\mathcal{F}$ is finite. Note that a classifier $f : \mathcal{X} \to \{0,1\}$ can be equivalently represented as a set $\{x \in \mathcal{X} : f(x) = 1\}$. A different and more practical (but completely equivalent) representation of f is given by the set $A_f := \{x \in \mathcal{X} : f(x) \neq f^\star(x)\}$[19]. We denote $\mathcal{A} := \{A_f : f \in \mathcal{F}\} \subseteq \{0,1\}^{\mathcal{X}}$. For $A_f \in \mathcal{A}$, the risk of A_f is naturally defined

$$R(A_f) = R(f) = \mathbb{P}[X \in A_f].$$

As in Theorem 14.2 we want to show that any consistent classifier has (with high probability) small risk. That is we want to upper bound

$$\begin{aligned}&\mathbb{P}(\text{there is } f \in \mathcal{F} \text{ with } R(f) \geq \varepsilon \text{ and such that } f(X_i) = f^*(X_i) \text{ for } i = 1, \dots, n),\\ &= \mathbb{P}(\exists A_f \in \mathcal{A} \text{ with } R(A_f) \geq \varepsilon \text{ and such that } X_i \notin A \text{ for all } i = 1, \dots, n).\end{aligned}$$

We now notice that $R(A_f) = \mathbb{P}[x \in A_f] = \mathbb{E}[\mathbb{1}_{x \in A_f}]$. Thus:

$$\begin{aligned}&\mathbb{P}(\text{there is } f \in \mathcal{F} \text{ with } R(f) \geq \varepsilon \text{ and such that } f(X_i) = f^*(X_i) \text{ for } i = 1, \dots, n),\\ &\leq \mathbb{P}\left(\exists A_f \in \mathcal{A} : \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \in A_f} - R(A_f)\right| \geq \varepsilon\right),\\ &\leq \mathbb{P}\left(\sup_{A \in \mathcal{A}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \in A_f} - R(A_f)\right| \geq \varepsilon\right).\end{aligned}$$

As we will refer to this bound several times in the rest of the notes, we state it as a lemma and make a further definition.

Definition 16.1

Given a sample $S_n = (X_1, \dots, X_n)$ of size n, and $A_f \in \mathcal{A}$ we define the *empirical risk* as the risk of f on the sample set

$$R_{S_n}(A_f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \in A_f},$$

where sometimes we drop the subscript f (in A_f) when there is no risk of confusion. We can also alternatively write it as

$$R_{S_n}(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f(X_i) \neq f^*(X_i)}.$$

[19] Check that this is indeed a bijection from the set of classifiers to $\{0,1\}^{\mathcal{X}}$.

We are now ready to state the lemma.

Lemma 16.2

Let $\mathcal{F}$ be a set of classifiers, and $f^\star \in \mathcal{F}$. Let $\mathcal{A} \coloneqq \{A_f : f \in \mathcal{F}\}$, with $A_f \coloneqq \{x \in \mathcal{X} : f(x) \neq f^\star(x)\}$. Then for any $\varepsilon > 0$ and $n \geq 1$:

$$\mathbb{P}(\exists f \in \mathcal{F} : R(f) \geq \varepsilon \text{ and } f(X_i) = f^*(X_i)\ \forall i \in [n]) \leq \mathbb{P}\left(\sup_{A \in \mathcal{A}} \left| R_{S_n}(A) - R(A) \right| \geq \varepsilon \right).$$

Note that the event in LHS in the lemma above corresponds to $\{\exists f \in \mathcal{F} : R(f) \geq \varepsilon \text{ and } R_{S_n}(f) = 0\}$.

Remark 16.3. By the law of large numbers we know that for any given $A \in \mathcal{A}$,

$$R_{S_n}(A) - R(A) = \frac{1}{n} \sum_{i=1}^{n} (\mathbb{1}_{X_i \in A} - \mathbb{E}[\mathbb{1}_{X_i \in A}]) \overset{\text{a.s.}}{\to} 0$$

However in Lemma 16.2, we are interested in the analysis of this convergence *uniformly* over all events $A \in \mathcal{A}$.

Remark 16.4 (Measurability)**.** We require that the random variable appearing in Lemma 16.2 is measurable. As we saw in Section 14, if $\mathcal{F}$ (or equivalently $\mathcal{A}$) is finite or even countable, then no problems of this sort appear. However, for infinite classes of events we need some mild assumptions, which we will not discuss (these are interesting considerations but not the focus of this course). All families of events and classifiers we consider will be be readily seen to verify all needed measurability conditions (note that most objects appearing are finite in the sense that they only depend on the sample of n samples). See Chapter 2 in [VPG15] for a more detailed discussion and relevant references. We additionally refer to Appendix A.1 in [BEHW89].

When $\mathcal{A}$ is finite we can take the union bound in Lemma 16.2, as we did in Section 14. The analysis becomes more complicated when $\mathcal{A}$ is infinite. However, a key remark is that while the set $\mathcal{A}$ is infinite, the set $A \in \mathcal{A}$ only appears in Lemma 16.2 via its *projections* $(\mathbb{1}_{X_1 \in A}, \ldots, \mathbb{1}_{X_n \in A})$. And for any given sample $(X_1, \cdots, X_n)$, the number of such projections (over all possible $A \in \mathcal{A}$) is always smaller than 2^n, and thus finite. It might even be much smaller than 2^n, which motivates the following definition:

Definition 16.5 (*Growth function*)

Given a family of events $\mathcal{A}$, the growth (shatter) function $\mathcal{S}_\mathcal{A}$ is defined by

$$\mathcal{S}_\mathcal{A}(n) \coloneqq \sup_{x_1, \ldots, x_n \in \mathcal{X}} |\{(\mathbb{1}_{x_1 \in A}, \ldots, \mathbb{1}_{x_n \in A}) : A \in \mathcal{A}\}|.$$

That is, the growth function bounds the number of projections of $\mathcal{A}$ on the sample $x_1, \ldots, x_n$.

Observe that $\mathcal{S}_\mathcal{A}(n) \leq 2^n$. Let us give some simple examples (if you prefer to think of classifiers, recall that any $A \in \mathcal{A}$ can be represented as one, for instance by $f(x) = \mathbb{1}_{x \in A}$):

1. The growth function of a finite family of events satisfies $\mathcal{S}_\mathcal{A}(n) \leq |\mathcal{A}|$.
2. Assume that $\mathcal{X} = \mathbb{R}$ and that $\mathcal{A}$ consists of the sets induced by all rays of the form $x \leq t$, $t \in \mathbb{R}$. Then, $\mathcal{S}_\mathcal{A}(n) = n + 1$.
3. Assume that $\mathcal{X} = \mathbb{R}$ and $\mathcal{A}$ consists of all open sets in $\mathbb{R}$. Then, $\mathcal{S}_\mathcal{A}(n) = 2^n$.

Remark 16.6. Recall that when considering a set of classifiers $\mathcal{F}$, we representated it as $\mathcal{A} \coloneqq \{A_f : f \in \mathcal{F}\} \subseteq \{0,1\}^\mathcal{X}$, with $A_f \coloneqq \{x \in \mathcal{X} : f(x) \neq f^\star(x)\}$. While the family $\mathcal{A}$ depends on $f^\star$, its growth function does not, as formalized in the following lemma:

Lemma 16.7 (*Growth function of family of classifiers – it does not depends on f^**)

Let $f^\star \in \mathcal{F}$, with $\mathcal{F}$ a class of classifiers. Let $A_f \coloneqq \{x \in \mathcal{X} : f(x) \neq f^\star(x)\}$ and $A'_f \coloneqq \{x \in \mathcal{X} : f(x) = 1\}$ for $f \in \mathcal{F}$, and we define $\mathcal{A} \coloneqq \{A_f : f \in \mathcal{F}\}$ and $\mathcal{A}' \coloneqq \{A'_f : f \in \mathcal{F}\}$. Then, for all $n \geq 1$:

$$\mathcal{S}_{\mathcal{A}}(n) = \mathcal{S}_{\mathcal{A}'}(n).$$

In particular, we define $\mathcal{S}_{\mathcal{F}}(n) \coloneqq \mathcal{S}_{\mathcal{A}}(n)$, and it does not depend on $f^\star$.

Proof of Lemma 16.7 – One can check that for all $x_1, \cdots, x_n \in \mathcal{X}$ and all $(\varepsilon_i)_{i=1}^n \in \{0,1\}^n$:

$$|\{(\mathbb{1}_{f(x_1)=1}, \cdots, \mathbb{1}_{f(x_n)=1}) : f \in \mathcal{F}\}| = |\{(\mathbb{1}_{f(x_1)=\varepsilon_1}, \cdots, \mathbb{1}_{f(x_n)=\varepsilon_n}) : f \in \mathcal{F}\}|. \tag{61}$$

What we claim follows by taking $\varepsilon_i = 1 - f^\star(x_i)$ and taking the supremum over $x_1, \cdots, x_n$. □

We are now ready to formulate the main result of this lecture. It gives us a guarantee for the uniform convergence of the frequencies of events $A \in \mathcal{A}$ to their probabilities (which appears in Lemma 16.2), depending on the growth function $\mathcal{S}_{\mathcal{A}}(n)$, rather than the size of $\mathcal{A}$.

Theorem 16.8 (*Vapnik-Chervonenkis Theorem*)

Consider a family of events $\mathcal{A}$ (or equivalently a family of classifiers) with the growth function $\mathcal{S}_{\mathcal{A}}$. For any $t \geq 0$, it holds that

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |R_{S_n}(A) - R(A)| \geq t\right) \leq 8\mathcal{S}_{\mathcal{A}}(n) \exp(-nt^2/32).$$

In particular, with probability at least $1 - \delta$, we have

$$\sup_{A \in \mathcal{A}} |R_{S_n}(A) - R(A)| \leq 4\sqrt{\frac{2}{n}\left(\log(8\mathcal{S}_{\mathcal{A}}(n)) + \log\frac{1}{\delta}\right)}.$$

We have the following corollary for our initial classification problem by using Lemma 16.2 (recall the definition of $\mathcal{S}_{\mathcal{F}}(n)$ in Lemma 16.7):

Corollary 16.9 (*VC theorem for classification*)

Let $\mathcal{F}$ be a (possibly infinite) class of classifiers, and let $f^\star \in \mathcal{F}$. Recall that $R(f) \coloneqq \mathbb{P}[f(X) \neq f^\star(X)]$. Then for any $\varepsilon \geq 0$, we have:

$$\mathbb{P}(\exists f \in \mathcal{F} \text{ with } R(f) \geq \varepsilon \text{ and } f(X_i) = f^*(X_i) \text{ for all } i = 1, \ldots, n) \leq 8\mathcal{S}_{\mathcal{F}}(n) \exp(-n\varepsilon^2/32).$$

Corollary 16.9 should be compared with eq. (57): we have managed to get a finite upper bound even for infinite classes of functions, as a function of the growth function rather than the size of the class!

Remark 16.10 (There is no free lunch)**.** Regardless of the family of events/classifiers, we have $S_{\mathcal{F}}(n) \leq 2^n$ but if we use this bound on Corollary 16.9 we would get that the probability of a consistent classifier having risk $\geq \varepsilon$ is upper bounded by $8 \times 2^n \exp(-n\varepsilon^2/32) = 8\exp(n\log(2) - n\varepsilon^2/32) > 1$, which is a vacuous statement. This is not surprising since we are not using any property of the family of classifiers we can expect to be able to learn (to quickly convince yourself of this, just take the classifier that is 1 for sampled data points and 0 for unseen data points).

In Section 17 we will see that $S_{\mathcal{F}}(n)$ can be controlled as a function of an easier to handle quantity, known as the VC dimension. This is at the core of VC learning theory.

16.2 Symmetrization and the proof of the VC Theorem (Theorem 16.8)

One of the main ingredient of the proof is a symmetrization lemma, similarly to what we use to prove a weak form of Hoeffding's inequality in Section 15. To ease readability we split the argument in two lemmas.

Let us first recall the notation.

Recall Notation 16.11. $\mathcal{X}$ is our sample space and samples $X_i \in \mathcal{X}$ are drawn from a probability distribution $\mathbb{P}$. Given n iid samples $S_n = (X_1, \ldots, X_n)$ and an event A (a measurable $A \subseteq \mathcal{X}$) we defined

$$R_{S_n}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in A} \text{ and } R(A) = \mathbb{E}\,\mathbb{1}_{X\in A} = \mathbb{P}(X \in A),$$

we also interchange between $\mathbb{P}_{X_1,\ldots,X_n}$ and $\mathbb{P}_{S_n}$ to denote the probability under the draw of the sample set.

Lemma 16.12 (*Symmetrization lemma I*)

Let $\mathcal{A}$ be a family of measurable events. Let $S_n = (X_1, \ldots, X_n)$ and $S'_n = (X'_1, \ldots, X'_n)$ denote two iid sample sets. Then, for any $t \geq \sqrt{3/n}$

$$\mathbb{P}_{S_n}\left(\sup_{A\in\mathcal{A}} \left|R_{S_n}(A) - R(A)\right| \geq t\right) \leq 2\,\mathbb{P}_{S_n,S'_n}\left(\sup_{A\in\mathcal{A}} \left|R_{S_n}(A) - R_{S'_n}(A)\right| \geq t/2\right).$$

Proof of Lemma 16.12 Let $0 < \varepsilon < \frac{1}{2025}$. Given S_n, let $A^* \in \mathcal{A}$ such that (if there are many such events, we choose one arbitrarily)

$$\left|R_{S_n}(A^*) - R(A^*)\right| \geq (1-\varepsilon)\sup_{A\in\mathcal{A}} \left|R_{S_n}(A) - R(A)\right|.$$

By triangular inequality,

$$\text{LHS} := \mathbb{1}_{\left|R_{S_n}(A^*)-R(A^*)\right|\geq(1-\varepsilon)t}\mathbb{1}_{\left|R_{S'_n}(A^*)-R(A^*)\right|<(1/2-\varepsilon)t} \leq \mathbb{1}_{\left|R_{S_n}(A^*)-R_{S'_n}(A^*)\right|\geq t/2} =: \text{RHS} \tag{62}$$

Since S'_n is independent of S_n, the event A^* does not depend on the sample S'_n. Taking expectations both with respect to S_n and S'_n in (62) we have

$$\mathbb{E}_{S_n,S'_n}\text{LHS} = \mathbb{P}_{S_n}\Big(\left|R_{S_n}(A^*) - R(A^*)\right| \geq (1-\varepsilon)t\Big)\mathbb{P}_{S'_n}\Big(\left|R_{S'_n}(A^*) - R(A^*)\right| < (1/2-\varepsilon)t\Big). \tag{63}$$

By the definition of A^* we have

$$\mathbb{P}_{S_n}\Big(\left|R_{S_n}(A^*) - R(A^*)\right| \geq (1-\varepsilon)t\Big) \geq \mathbb{P}_{S_n}\left(\sup_{A\in\mathcal{A}} \left|R_{S_n}(A) - R(A)\right| \geq t\right). \tag{64}$$

Also, note that $n \times (R_{S'_n}(A^*) - R(A^*)) = \sum_{i=1}^n \left(\mathbb{1}_{X'_i\in A^*} - \mathbb{E}\mathbb{1}_{X'_i\in A^*}\right)$ is sum of n centered Bernoulli random variables with[20] $\text{Var}\left(\mathbb{1}_{X'_i\in A^*} - \mathbb{E}\mathbb{1}_{X'_i\in A^*}\right) \leq \frac{1}{4}$ and so $\text{Var}\left(R_{S'_n}(A^*) - R(A^*)\right) \leq \frac{1}{n^2}\frac{n}{4} = \frac{1}{4n}$. Hence, by Chebyshev's inequality (Proposition 15.2) we have

$$\begin{aligned}\mathbb{P}_{S'_n}\Big(\left|R_{S'_n}(A^*) - R(A^*)\right| < (1/2-\varepsilon)t\Big) &= 1 - \mathbb{P}_{S'_n}\Big(\left|R_{S'_n}(A^*) - R(A^*)\right| \geq (1/2-\varepsilon)t\Big)\\ &\geq 1 - \frac{1/(4n)}{(1/2-\varepsilon)^2t^2} = 1 - \frac{1}{nt^2}\frac{1}{(1-2\varepsilon)^2}.\end{aligned}$$

[20] A Bernoulli random variable $B(p)$ (which takes the value 1 with probability p and the value 0 with probability $1-p$ has variance $p(1-p)$ which is maximized at $p = \frac{1}{2}$.

Since $\varepsilon < \frac{1}{2025}$ we have that $(1-2\varepsilon)^2 \geq \frac{2}{3}$ and so, for any $t \geq \sqrt{\frac{3}{n}}$ we have $\frac{1}{nt^2}\frac{1}{(1-2\varepsilon)^2} \geq 1/2$ and so

$$\mathbb{P}_{S'_n}\left(\left|R_{S'_n}(A^*) - R(A^*)\right| < (1/2-\varepsilon)t\right) \geq \frac{1}{2}, \tag{65}$$

which together with (64) implies

$$\mathbb{E}_{S_n,S'_n}\text{LHS} \geq \frac{1}{2}\mathbb{P}_{S_n}\left(\sup_{A\in\mathcal{A}}\left|R_{S_n}(A) - R(A)\right| \geq t\right). \tag{66}$$

On the other hand,

$$\mathbb{E}_{S_n,S'_n}\text{RHS} = \mathbb{P}_{S_n,S'_n}\left(\left|R_{S_n}(A^*) - R_{S'_n}(A^*)\right| \geq t/2\right) \leq \mathbb{P}_{S_n,S'_n}\left(\sup_{A\in\mathcal{A}}\left|R_{S_n}(A) - R_{S'_n}(A)\right| \geq t/2\right). \tag{67}$$

Combining (66) and (67) with (62) finishes the proof.

□

Lemma 16.13 (*Symmetrization lemma II*)

Let $\varepsilon_1,\ldots,\varepsilon_n$ be independent (from each other and from $X_i,\ i=1,\ldots,n$) random variables taking the values ± 1 each with probability $1/2$. Then, for any $t \geq \sqrt{3/n}$, it holds that

$$\mathbb{P}_{S_n}\left(\sup_{A\in\mathcal{A}}\left|R_{S_n}(A) - R(A)\right| \geq t\right) \leq 4\mathbb{P}_{\substack{X_1,\ldots,X_n,\\ \varepsilon_1,\ldots,\varepsilon_n}}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{X_i\in A}\right| \geq t/4\right).$$

Proof of Lemma 16.13 This follows readily by using Lemma 16.12 for a new set of iid samples S'_n, also independent of $\varepsilon_1,\ldots,\varepsilon_n$, and noting that

$$\begin{aligned}
&\mathbb{P}_{S_n,S'_n}\left(\sup_{A\in\mathcal{A}}\left|R_{S_n}(A) - R_{S'_n}(A)\right| \geq t/2\right)\\
&= \mathbb{P}_{S_n,S'_n}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n\left(\mathbb{1}_{X_i\in A} - \mathbb{1}_{X'_i\in A}\right)\right| \geq t/2\right)\\
&= \mathbb{P}_{\substack{S_n,S'_n,\\ \varepsilon_1,\ldots,\varepsilon_n}}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(\mathbb{1}_{X_i\in A} - \mathbb{1}_{X'_i\in A}\right)\right| \geq t/2\right)\\
&\leq \mathbb{P}_{\substack{S_n,S'_n,\\ \varepsilon_1,\ldots,\varepsilon_n}}\left(\sup_{A\in\mathcal{A}}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{1}_{X_i\in A}\right| + \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{1}_{X'_i\in A}\right|\right] \geq t/2\right)\\
&\leq \mathbb{P}_{\substack{S_n,S'_n,\\ \varepsilon_1,\ldots,\varepsilon_n}}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{1}_{X_i\in A}\right| + \sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{1}_{X'_i\in A}\right| \geq t/2\right)\\
&\leq \mathbb{P}_{\substack{S_n,\\ \varepsilon_1,\ldots,\varepsilon_n}}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{1}_{X_i\in A}\right| \geq t/4\right) + \mathbb{P}_{\substack{S'_n,\\ \varepsilon_1,\ldots,\varepsilon_n}}\left(+\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{1}_{X'_i\in A}\right| \geq t/4\right) \qquad (68)\\
&= 2\mathbb{P}_{\substack{X_1,\ldots,X_n,\\ \varepsilon_1,\ldots,\varepsilon_n}}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{1}_{X_i\in A}\right| \geq t/4\right), \qquad (69)\\
&\qquad (70)
\end{aligned}$$

where we used the union bound in (68) and the fact that S_n and S'_n are identically distributed in (69).
□

Proof of VC Theorem (Theorem 16.8) Since $\mathcal{S}_\mathcal{A}(n)$ is at least 1, if $t < \sqrt{\frac{3}{n}}$ then

$$8\mathcal{S}_\mathcal{A}(n)\exp(-nt^2/32) \geq 8\exp\left(-n\frac{3}{n}/32\right) = 8\exp(-3/32) > 1,$$

and so the RHS in Theorem 16.8 is larger than 1 which means the bound holds trivially.

If $t \geq \sqrt{\frac{3}{n}}$ then, using Lemma 16.13, we consider the term

$$4\mathbb{P}_{\substack{X_1,\dots,X_n,\\ \varepsilon_1,\dots,\varepsilon_n}}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbb{1}_{X_i\in A}\right| \geq t/4\right).$$

As we mentioned above, the key observation is that even though the set of events $\mathcal{A}$ is infinite, there are at most $\mathcal{S}_{\mathcal{A}}(n)$ realizations of $(\mathbb{1}_{X_1\in A},\dots,\mathbb{1}_{X_n\in A})$ for a given sample $X_1,\dots,X_n$. To clarify, let us fix $X_1,\cdots,X_n$, and denote $\mathcal{M}(X_1,\cdots,X_n) \coloneqq \{(\mathbb{1}_{X_1\in A},\dots,\mathbb{1}_{X_n\in A}) : A\in\mathcal{A}\}$. By Definition 16.5, $|\mathcal{M}| \leq \mathcal{S}_{\mathcal{A}}(n)$. Moreover:

$$\begin{aligned}
\mathbb{P}_{\varepsilon_1,\dots,\varepsilon_n}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbb{1}_{X_i\in A}\right| \geq t/4\right) &= \mathbb{P}_{\varepsilon_1,\dots,\varepsilon_n}\left(\sup_{y\in\mathcal{M}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i\right| \geq t/4\right),\\
&\leq \sum_{y\in\mathcal{M}}\mathbb{P}_{\varepsilon_1,\dots,\varepsilon_n}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i\right| \geq t/4\right) \quad \text{(union bound)},\\
&\leq \mathcal{S}_{\mathcal{A}}(n)\sup_{y\in\mathcal{M}}\mathbb{P}_{\varepsilon_1,\dots,\varepsilon_n}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i\right| \geq t/4\right),\\
&= \mathcal{S}_{\mathcal{A}}(n)\sup_{A\in\mathcal{A}}\mathbb{P}_{\varepsilon_1,\dots,\varepsilon_n}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbb{1}_{X_i\in A}\right| \geq t/4\right). \qquad (71)
\end{aligned}$$

We can then apply Hoeffding's inequality in eq. (71) (observe that $\varepsilon_i\mathbb{1}_{X_i\in A}\in[-1,1]$ and are independent), and we have

$$\begin{aligned}
&4\mathbb{P}_{\substack{X_1,\dots,X_n,\\ \varepsilon_1,\dots,\varepsilon_n}}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbb{1}_{X_i\in A}\right| \geq t/4\right)\\
&\leq 4\mathbb{E}_{X_1,\dots,X_n}\left(\mathcal{S}_{\mathcal{A}}(n)\sup_{A\in\mathcal{A}}\mathbb{P}_{\varepsilon_1,\dots,\varepsilon_n}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbb{1}_{X_i\in A}\right| \geq t/4\right)\right) \quad \text{(by eq. (71))}\\
&\leq 4\mathcal{S}_{\mathcal{A}}(n)\mathbb{E}\cdot\left(2\exp(-2nt^2/(4\cdot 16))\right) \quad \text{(Hoeffding's inequality)}\\
&= 8\mathcal{S}_{\mathcal{A}}(n)\exp(-nt^2/32).
\end{aligned}$$

The claim follows. □ □

Challenge 16.1. *You might have noticed that we could have skipped the second symmetrization lemma and analyzed* $\mathbb{P}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{X_i\in A}-\mathbb{1}_{X_i'\in A})\right| \geq t/2\right)$ *directly without introducing random signs* ε_i*. Try to improve the constants in the Theorem by doing this instead (and perhaps also improving the constants in the first symmetrization lemma).*

Further Reading 16.14. The uniform convergence theorem and the growth function appear in the foundational work of Vapnik and Chervonenkis [VC71]. Symmetrization with random signs appears in [GZ84]. A modern presentation of similar results can be found in the textbook [Ver18].

17 The Vapnik-Chervonenkis dimension

17.1 Definition and first examples

We are now ready for the final lecture of this class, in which we will generalize PAC learnability (Corollary 14.7) to infinite classes. To do so, we will upper bound the growth function appearing in Corollary 16.9 using the concept of Vapnik-Chervonenkis (VC) dimension.

Definition 17.1 (*Shattered set*)

Given a family of events $\mathcal{A}$, we say that a finite set $\{x_1, \dots, x_d\} \subset \mathcal{X}$ is *shattered* by $\mathcal{A}$ if the number of projections of $\mathcal{A}$ on $\mathcal{X}$ is equal to 2^d, that is if $\{(\mathbb{1}_{x_1 \in A}, \dots, \mathbb{1}_{x_d \in A}) : A \in \mathcal{A}\} = \{0,1\}^d$, or equivalently

$$|\{(\mathbb{1}_{x_1 \in A}, \dots, \mathbb{1}_{x_d \in A}) : A \in \mathcal{A}\}| = 2^d.$$

Recall the definition of the growth (or shatter) function $\mathcal{S}_{\mathcal{A}}$ in Definition 16.5.

Definition 17.2 (*VC dimension*)

Given a family of events $\mathcal{A}$, the Vapnik Chervonenkis (VC) dimension of $\mathcal{A}$ is the size of the largest subset of $\mathcal{X}$ that is shattered by $\mathcal{A}$. Equivalently, it is the largest integer d such that:

$$\mathcal{S}_{\mathcal{A}}(d) = \sup_{x_1, \dots, x_d \in \mathcal{X}} |\{(\mathbb{1}_{x_1 \in A}, \dots, \mathbb{1}_{x_d \in A}) : A \in \mathcal{A}\}| = 2^d.$$

If $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ for all $n \geq 1$, we set $d = \infty$.

First, we consider several simple examples (try to work them out yourself, to build some intuition). Again here we consider general sets of events $\mathcal{A} \subseteq \{0,1\}^{\mathcal{X}}$, however one can always equivalently build classifiers as $f(x) = \mathbb{1}\{x \in A\}$ for $A \in \mathcal{A}$.

Example 17.3. The VC dimension of the family $\mathcal{A} = \{[a,b], a \leq b\}$ of the closed intervals in $\mathbb{R}$ is equal to 2. This is because a pair of distinct points can be shattered. But there is no interval that contains two points but does not contain a point between them. Thus, the set of three points cannot be shattered. More formally:

$$\begin{cases} \mathcal{S}_{\mathcal{A}}(1) &= \sup_{x \in \mathbb{R}} |\{(\mathbb{1}_{x \in [a,b]}) \, : \, a \leq b\}| = 2, \\ \mathcal{S}_{\mathcal{A}}(2) &= \sup_{x,y \in \mathbb{R}} |\{(\mathbb{1}_{x \in [a,b]}, \mathbb{1}_{y \in [a,b]}) \, : \, a \leq b\}| = 4, \\ \mathcal{S}_{\mathcal{A}}(3) &= \sup_{x,y,z \in \mathbb{R}} |\{(\mathbb{1}_{x \in [a,b]}, \mathbb{1}_{y \in [a,b]}, \mathbb{1}_{z \in [a,b]}) \, : \, a \leq b\}| = 7 < 2^3. \end{cases}$$

Example 17.4. The VC dimension of the family of events induced by halfspaces in $\mathbb{R}^2$ (not necessarily passing through the origin) is equal to 3. Indeed, a set of three distinct points can be shattered in all possible 2^3 ways. At the same time, for a set of 4 points it is impossible to shatter the set in a way such that two diagonals of the corresponding rectangle are in two different halfspaces (draw it!).

Example 17.5. Generalizing abusively from the above examples, one could think that the VC dimension is closely related to the number of parameters. However, there is a classical example of a family of events parametrized by a single parameter such that its VC dimension is infinite. Consider the family of events $\mathcal{A} = \{A_t \, : \, t > 0\}$, with

$$A_t = \{x \in \mathbb{R} \setminus \{0\} : \sin(xt) \geq 0\} = \bigcup_{k \in \mathbb{Z}} \left[\frac{2k\pi}{t}, \frac{(2k+1)\pi}{t} \right] \setminus \{0\}.$$

One can verify that a set of any size can be shattered by this family of sets. Therefore, its VC dimension is infinite.

Example 17.6. The VC dimension of the family of events induced by non-homogeneous half-spaces in $\mathbb{R}^p$ is equal to $p+1$. For a proof of this fact, see the notes of the previous year [BZ22].

17.2 Uniform convergence and the VC dimension

In order to relate the conclusion of VC's Theorem 16.8 (or Corollary 16.9) to the VC dimension, we need to relate it to the growth function $\mathcal{S}_{\mathcal{A}}(n)$ for general values of n. We know that for $n \leq d$, $\mathcal{S}_{\mathcal{A}}(n) = 2^n$. The following theorem gives an upper bound on $\mathcal{S}_{\mathcal{A}}(n)$ for $n \geq d$. Quite surprisingly, it was shown by several authors independently around the same time. While Vapnik-Chervonenkis were motivated by uniform convergence, other authors looked at it from a different perspective. Currently there are several known techniques that can be used to prove this result.

Theorem 17.7 (*Sauer-Shelah-Vapnik-Chervonenkis*)

Assume that the VC dimension of $\mathcal{A}$ is equal to d. Then for any $n \geq d$:

$$\mathcal{S}_{\mathcal{A}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

We present two proofs of this Theorem, one through a tutorial[21] in a challenge below and an alternative proof in Appendix B.

Challenge 17.1 (Proof of the Sauer-Shelah-Vapnik-Chervonenkis Theorem (Theorem 17.7))**.** *Given a set $S = \{x_1, \ldots, x_n\}$ let $\mathcal{A}_S = \{(\mathbb{1}_{x_1 \in A}, \ldots, \mathbb{1}_{x_n \in A}) \colon A \in \mathcal{A}\} \subseteq \{0,1\}^n$. This challenge is a tutorial on how to prove the Sauer-Shelah-Vapnik-Chervonenkis Theorem (Theorem 17.7). We want to prove that, if $\mathcal{A}$ has VC dimension d, then for any n,*

$$\mathcal{S}_{\mathcal{A}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}. \tag{72}$$

In order to prove (72), we will actually prove the following statement by induction: for any set $S = \{x_1, \ldots, x_n\}$:

$$|\mathcal{A}_S| \leq |\{B \subseteq S \colon B \text{ is shattered by } \mathcal{A}|\}, \tag{73}$$

i.e. the number of possible labeling of S is bounded by the number of different subsets of S that can be shattered.

(A) Establish that (73) holds for $S = \emptyset$ (the empty set).

(B) For any set S and any point $x' \notin S$, assume (73) holds for S and for any hypothesis class, and prove that (73) holds for $S' = S \cup \{x'\}$ and any hypothesis class. To this end, for any hypothesis class $\mathcal{A}$, write $\mathcal{A} = \mathcal{A}^0 \cup \mathcal{A}^1$ where:

$$\mathcal{A}^0 = \{A \in \mathcal{A} \colon \mathbb{1}_{x' \in A} = 0\}, \qquad \mathcal{A}^1 = \{A \in \mathcal{A} \colon \mathbb{1}_{x' \in A} = 1\}.$$

(a) Prove that $|\mathcal{A}_{S'}| = |\mathcal{A}^0_S| + |\mathcal{A}^1_S|$.

(b) Prove that (73) holds for S and $\mathcal{A}$ by applying (73) to each of the two terms on the right-hand-side above.

We can now conclude that (73) holds for any (finite) S and any $\mathcal{A}$.

(C) Use (73) to establish (72).

[21] Thanks to Nati Srebro for sharing this tutorial!

□

We can now present a key corollary of Theorem 17.7, which generalizes the conclusion of the uniform convergence Theorem 16.8 to families of events with finite VC dimension.

Theorem 17.8 (*VC Theorem with VC dimension*)

Consider a family of events $\mathcal{A}$ with the VC dimension d. If $n \geq d$, then for any $t > 0$:

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |\nu_n(A) - P(A)| \geq t\right) \leq \exp\left(d\log\frac{en}{d} - \frac{nt^2}{32}\right).$$

In particular, with probability at least $1-\delta$, we have

$$\sup_{A\in\mathcal{A}} |\nu_n(A) - P(A)| \leq 4\sqrt{\frac{2}{n}\left(d\log\left(\frac{8en}{d}\right) + \log\frac{1}{\delta}\right)}.$$

Proof of Theorem 17.8 – The proof uses the uniform convergence Theorem 16.8 together with Theorem 17.7. We use the elementary identity, for $d \leq n$:

$$\sum_{i=0}^{d}\binom{n}{i} \leq \sum_{i=0}^{d}\left(\frac{n}{d}\right)^{d-i}\binom{n}{i} \leq \sum_{i=0}^{n}\left(\frac{n}{d}\right)^{d-i}\binom{n}{i} = \left(1+\frac{d}{n}\right)^{n}\left(\frac{n}{d}\right)^{d} \leq \left(\frac{en}{d}\right)^{d},$$

Therefore we have

$$\log(8\mathcal{S}_{\mathcal{A}}(n)) \leq \log(8\,(en/d)^{d}) \leq d\log(8en/d).$$

The claim follows by Theorem 16.8. □

17.3 Application in classification theory

Definition 17.9 (*VC dimension of a set of classifiers*)

For a class $\mathcal{F}$ of classifiers, we define the VC dimension of $\mathcal{F}$ as the VC dimension of $\mathcal{A}' \subseteq \{0,1\}^{\mathcal{X}}$ whose elements are $A'_f := \{x : f(x) = 1\}$ for $f \in \mathcal{F}$.

Remark 17.10. As a consequence of Lemma 16.7, for any $f^\star \in \mathcal{F}$ the VC dimension of $\mathcal{F}$ is equal to the VC dimension of $\mathcal{A} = \{A_f : f \in \mathcal{F}\}$ with $A_f = \{f(x) \neq f^\star(x)\}$, since $\mathcal{S}_{\mathcal{A}}(n) = \mathcal{S}_{\mathcal{A}'}(n)$.

By using Theorem 17.8 in Lemma 16.2, we can now generalize PAC-learnability of finite classes to any class with finite VC dimension.

Theorem 17.11 (*PAC learnability of classifiers*)

Any class $\mathcal{F}$ with the finite VC dimension d is PAC learnable by any algorithm choosing a consistent classifier in $\mathcal{F}$, with the sample complexity $n = n(\varepsilon,\delta)$ such that:

$$n \geq \frac{32}{\varepsilon^2}\left[d\log\frac{8en}{d} + \log\frac{1}{\delta}\right]. \tag{74}$$

Proof of Theorem 17.11 – Let $f^\star \in \mathcal{F}$. We can apply Theorem 17.8 in Lemma (16.2) (using Remark 17.10), we get for any $\varepsilon > 0$:

$$\mathbb{P}(\text{there is } f \in \mathcal{F} \text{ with } R(f) \geq \varepsilon \text{ and such that } f(X_i) = f^*(X_i) \text{ for } i = 1,\ldots,n),$$

$$\leq \exp\Big(d\log\frac{en}{d} - \frac{n\varepsilon^2}{32}\Big).$$

Equivalently, with probability at least $1-\delta$ (the sup being taken over consistent classifiers $\widehat{f}$):

$$\sup_{\widehat{f}} R(\widehat{f}) \leq 4\sqrt{\frac{2}{n}\left(d\log\left(\frac{8en}{d}\right) + \log\frac{1}{\delta}\right)}.$$

Hence if the sample size $n(\varepsilon,\delta)$ is such that $4\sqrt{\frac{2}{n}\left(d\log\left(\frac{8en}{d}\right)+\log\frac{1}{\delta}\right)} \leq \varepsilon$, then $\sup_{\widehat{f}} R(\widehat{f}) \leq \varepsilon$. □

Example 17.12. The classes $\mathcal{F}$ of halfspaces in $\mathbb{R}^p$, intervals and rays in $\mathbb{R}$ are PAC learnable.

Further Reading 17.13. The Sauer-Shelah-Vapnik-Chervonenkis lemma appears independently and in different contexts in [VC71, Sau72, She72]. Further relations between PAC learning and the VC dimensions were made in [BEHW89]. So far we observed that the finiteness of the VC dimension imply the PAC-learnability. In the notes of the previous year [BZ22], another sufficient condition for PAC-learnability is discussed, namely the existence of a finite sample compression scheme. Relating the existence of finite compression schemes to the VC dimension leads to important conjectures in learning theory.

18 Fourier Transform and Bochner's theorem

In this section we will introduce Bochner's Theorem, which characterizes translation-invariant Positive Definite Kernels, but before we need to take a small detour to introduce the Fourier Transform.

18.1 Fourier Transform

This is a brief introduction to Fourier Transform. Math BSc students at ETH have a detailed and rigorous introduction in Analysis IV, others have many options for books on the subject ([SS03] is an excellent one). In this subsection, functions are functions from $\mathbb{R}$ (or $\mathbb{R}^p$) to $\mathbb{C}$.

18.1.1 Fourier Transform in $\mathbb{R}$

Given $f \in L^1(\mathbb{R})$, a complex-valued integrable function, we can define its Fourier transform as:

$$\hat{f}(\xi) \coloneqq \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t)\, e^{-i\xi t}\, \mathrm{d}t \qquad (\xi \in \mathbb{R}). \tag{75}$$

If $\hat{f} \in L^1(\mathbb{R})$, then we have a *Fourier inversion theorem*: for all $t \in \mathbb{R}$ which are continuity points of f we have:

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{f}(\xi)\, e^{i\xi t}\, \mathrm{d}\xi. \tag{76}$$

Furthermore, the definition (75) can be extended to square-integrable functions $f \in L^2(\mathbb{R})$.[22] An essential result in this context is *Plancherel's Theorem*, which states that the Fourier transform is an isometry of $L^2(\mathbb{R})$, i.e.

$$\int_{\mathbb{R}} |f(t)|^2\, \mathrm{d}t = \int_{\mathbb{R}} |\hat{f}(\xi)|^2\, \mathrm{d}\xi. \tag{77}$$

Challenge 18.1. *Prove these properties. You can assume $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ (or even $f \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ if it makes it easier), and the same for $\hat{f}$. There are fascinating connections between the regularity (smoothness, integrability, decay, etc...) of f and $\hat{f}$, but for a first introduction, try to prove the properties above assuming whatever regularity you need. You can find out more in e.g. [SS03].*

Challenge 18.2. *Similarly to Plancherel's Theorem (77), can you show that the Fourier Transform also perserves inner-products in $L^2(\mathbb{R})$?*

One reason the Fourier Transform is a central object in so many areas of Mathematics and beyond is that it effectively diagonalizes translations (and so also differentiation, which is in a sense the reason why it is such a useful tool when studying differential equations). This will be more clear once we talk about the Discrete Fourier Transform, and in a more abstract sense once you study some representation theory of groups.

For now, we observe an important property of the Fourier Transform that illustrates this fact: for $f \in L^1(\mathbb{R})$ and $t_0 \in \mathbb{R}$, let $h(t) \coloneqq T_{t_0} f(t) = f(t - t_0)$. Then

$$\hat{h}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t - t_0) e^{-it\xi} \mathrm{d}t = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t - t_0) e^{-i(t-t_0)\xi} e^{-it_0\xi} \mathrm{d}t = e^{-it_0\xi} \hat{f}(\xi). \tag{78}$$

The transformation $\mathcal{M}_{t_0} : g(\xi) \mapsto e^{-it_0\xi} g(\xi)$ is known as a modulation. It is "diagonal" in the sense that the value of $g(\xi)$ depends only on the value of $g(\xi)$ and not on the value of g at other arguments.

[22] There are very interesting question about how much regularity is needed for what follows, and how much regularity do certain properties on f enforce on $\hat{f}$, and vice-versa. This, however, is not the focus of this course. Our approach here is: Let us focus on functions that have however much regularity (and integrability) we need so regularity and integrability are not a major concern; afterwards we could go back and try to ask ourselves what are the minimal assumptions needed.

Challenge 18.3. *Derive a formula for the Fourier Transform of the derivative of f in terms of the Fourier Transform of f. Is it also "diagonal"? (in the same sense as above)*

Challenge 18.4. *Derive formulas for the Fourier Transform of:*

1. *A dilation of a function f, i.e. $h(x) \coloneqq f(\alpha x)$, for $\alpha \in \mathbb{R}$.*
2. *A modulation of a function f, i.e. $h(x) \coloneqq e^{i\beta x} f(x)$, for $\beta \in \mathbb{R}$.*
3. *Note that above we derived formulas for the Fourier Transform of translations $h(x) = f(x - x_0)$ and derivatives $h(t) = f'(t)$.*

18.1.2 Fourier Transform in $\mathbb{R}^p$

The Fourier Transform can be analogously defined in $\mathbb{R}^p$. Indeed, given $f \in L^1(\mathbb{R}^p)$, we can define its Fourier transform as:

$$\hat{f}(u) \coloneqq \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} f(x)\, e^{-iu^\top x}\, \mathrm{d}x \qquad (u \in \mathbb{R}^p). \tag{79}$$

The properties showed above for $p = 1$ have direct analogues in this setting, we include here the inverse formula. If $\hat{f} \in L^1(\mathbb{R}^p)$ then, for all $x \in \mathbb{R}^p$ which are continuity points of f:

$$f(x) = \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \hat{f}(u)\, e^{iu^\top x}\, \mathrm{d}u. \tag{80}$$

Challenge 18.5. *Show analogues in this setting of the properties described above for the one dimensional Fourier Transform.*

18.2 Bochner's theorem

In this section, we focus on the special case of *translation-invariant* kernels. They are kernels that are a function only of the difference between the points, i.e. such that $K(x_i, x_j) = q(x_i - x_j)$, for some function $q : \mathbb{R}^p \mapsto \mathbb{R}$. Note that this is the case e.g. for the Gaussian Kernel $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\varepsilon^2)$.

In this specific setting, Bochner's theorem relates a kernel being positive with properties of its Fourier Transform. This theorem can be used to solve Challenge 6.1 (but there are other ways). We are now ready to state Bochner's theorem:

Theorem 18.1 (*Bochner*)

Let $K(x, y) = q(x - y)$ be a translation invariant kernel, real-valued and symmetric. Assume that q is continuous. Then the two following are equivalent:

(*i*) K is positive definite.

(*ii*) There exists a positive and finite measure μ on $\mathbb{R}^p$ such that q is the Fourier Transform of μ, i.e. for all $x \in \mathbb{R}^p$:

$$q(x) = \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} e^{-iu^\top x} \mathrm{d}\mu(u).$$

The proof of Bochner's theorem 18.1 can be found in several textbooks on harmonic analysis, cf. for instance [Kat04]. In these notes we show a weaker version of Bochner's Theorem.

Theorem 18.2 (*Bochner, weak version*)

Let $K(x,y) = q(x-y)$ be a translation invariant kernel, real-valued and symmetric. Assume that q is continuous, and that $q, \hat{q} \in L^1(\mathbb{R}^p)$. Then the two following are equivalent:

(i) K is positive definite.

(ii) For all $u \in \mathbb{R}^p$, $\hat{q}(u) \geq 0$.

Remark – Actually the hypothesis that $\hat{q} \in L^1(\mathbb{R}^p)$ is not necessary in Theorem 18.2: one can show that either (i) or (ii) imply that $\hat{q} \in L^1(\mathbb{R}^p)$, cf e.g. the notes [Gub18].

In the lecture and in this section we prove only the easier implication $(ii) \Rightarrow (i)$, see Appendix A for the other implication. Note that the proof of $(ii) \Rightarrow (i)$ is exactly the same in both Theorem 18.2 and Theorem 18.1.

Proof of $(ii) \Rightarrow (i)$ in Theorem 18.2 – Since $\hat{q} \in L^1(\mathbb{R}^p)$, we can use the Fourier inversion formula for all $x \in \mathbb{R}^p$[23]

$$q(x) = \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} e^{iu^\top x}\, \hat{q}(u)\, \mathrm{d}u. \tag{81}$$

We fix $n \geq 0$ and $x_1, \cdots, x_n \in \mathbb{R}^p$. Let $M_{ij} := q(x_i - x_j)$. Since q is even (since the Kernel is symmetric), M is symmetric. Let us show that M is positive semidefinite. We fix any $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n$, and show that $\alpha^\top M \alpha \geq 0$. We have:

$$\begin{aligned}
\alpha^\top M \alpha &= \sum_{j,k=1}^{n} \alpha_j \alpha_k q(x_j - x_k), \\
&= \frac{1}{(2\pi)^{p/2}} \sum_{j,k=1}^{n} \alpha_j \alpha_k \int_{\mathbb{R}^p} e^{iu^\top (x_j - x_k)} \hat{q}(u) \mathrm{d}u, \\
&= \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \sum_{j,k=1}^{n} \alpha_j \alpha_k e^{iu^\top (x_j - x_k)} \hat{q}(u) \mathrm{d}u, \\
&= \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \left| \sum_{j=1}^{n} \alpha_j e^{iu^\top x_j} \right|^2 \hat{q}(u) \mathrm{d}u \geq 0.
\end{aligned}$$

□

[23] Since q is continuous, the formula is valid for all x.

19 Fourier Series and Shannon Sampling

In this section we will shift gears somewhat and show an important application of Fourier theory in Signal Processing.

19.1 Fourier Series and $L^2([-\pi,\pi])$

Let us start by recalling[24] some properties of $L^2([-\pi,\pi])$. All of the sequel can be analogously done for $L^2[-\Omega\pi,\Omega\pi]$ for any $\Omega > 0$ by appropriately scaling quantities. To ease notation, we set $\Omega = 1$. $L^2([-\pi,\pi])$ is the Hilbert space of square-integrable complex-valued functions in $[-\pi,\pi]$ with the inner-product given by[25]

$$\langle f, g\rangle \coloneqq \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx,$$

and the associated norm

$$\|f\|^2 = \int_{-\pi}^{\pi} |f(x)|^2 dx.$$

A remarkable property of $L^2([-\pi,\pi])$ is that the harmonic functions

$$\left\{x \mapsto \frac{1}{\sqrt{2\pi}} e^{inx}\right\}_{n\in\mathbb{Z}} \tag{82}$$

are an orthonormal basis for $L^2([-\pi,\pi])$.

Challenge 19.1. *Show that the functions (82) are orthonormal, i.e.*

$$\left\langle \frac{1}{\sqrt{2\pi}} e^{inx}, \frac{1}{\sqrt{2\pi}} e^{imx} \right\rangle = \delta_{n,m}.$$

The fact that this is a basis means that for every function $f \in L^2([-\pi,\pi])$, there exists a sequence $\{a_n\}_{n\in\mathbb{Z}}$ such that

$$f(x) = \sum_{n=-\infty}^{\infty} a_n \frac{1}{\sqrt{2\pi}} e^{inx},$$

with equality in the sense of $L^2([-\pi,\pi])$. Since the basis (82) is orthonormal, the coefficients are given by

$$a_n = \left\langle f(x), \frac{1}{\sqrt{2\pi}} e^{inx} \right\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(x) e^{-inx} = \hat{f}(n).$$

This expansion is known as Fourier Series.

Definition 19.1 (*Fourier Series*)

Given $f \in L^2([-\pi,\pi])$ we define its Fourier Series as

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{inx}, \tag{83}$$

Challenge 19.2. *Try to show Parseval's theorem: For $f \in L^2([-\pi,\pi])$,*

$$\|f\|^2 = \sum_{k\in\mathbb{Z}} |\hat{f}(n)|^2.$$

Note that we will identify a function $f \in L^2([-\pi,\pi])$ with the function in $L^2(\mathbb{R})$ that is equal to f in $[-\pi,\pi]$ and zero elsewhere.

[24] The ETH Math BSc students see the proof of this in Analysis IV, others can see it in any of several excellent books on Theory of Hilbert Spaces, Functional Analysis, or Fourier Theory, a very good example is [SS03].

[25] Warning: In Physics it is more common to use the convention $\langle f, g\rangle = \int_{-\pi}^{\pi} \overline{f(x)} g(x) dx$. We use the classical convention in Mathematics $\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx$.

19.2 Shannon Sampling Theorem

The Shannon Sampling Theorem is a key result in Signal Processing. In this section, we will consider functions $f : t \in \mathbb{R} \mapsto f(t) \in \mathbb{C}$, which we interpret as a *signal*, e.g. the sound of a music, that is a function of the *time* $t \in \mathbb{R}$. Sometimes the functions are real-valued, although the theory below is naturally presented in the more general case of complex-valued functions.

Recall definition (75). Plancherel's Theorem (77) states that for $f \in L^2(\mathbb{R})$,

$$\int_{\mathbb{R}} |f(t)|^2 \, \mathrm{d}t = \int_{\mathbb{R}} |\hat{f}(\xi)|^2 \, \mathrm{d}\xi. \tag{84}$$

When talking about signals, the quantity $\int |f(t)|^2$ is sometimes called the *energy* of the signal, while the integrand $|\hat{f}(\xi)|^2$ on the right-hand-side of eq. (77) is sometimes referred to as the *spectral density* of the signal. By eq. (77), the spectral density in ξ represents how the energy of the signal f is distributed across frequencies (i.e. what is the "contribution" of the frequency $t \mapsto e^{-i\xi t}$).

Bandlimited functions – We can often limit the range of frequencies $\xi \in \mathbb{R}$ that we consider. This can be motivated by two observations

- In physical signals, the large majority of the energy is usually spread out over a finite range of frequencies, that we call its *bandwith*. Physically, the spectral density of any $f \in L^2(\mathbb{R})$ has to decrease simply because it is integrable by eq. (77). One can then put a cut-off on values of $|f(\xi)|$ smaller than some threshold: this effectively creates a signal whose frequencies should lie in a finite range $[-\Lambda, \Lambda]$, for some $\Lambda > 0$.
- The observations of the signal we can make also effectively limit its bandwith. Think for instance of the human ear or eye, which can only see light in the wavelength range of (approximately) 380 to 750 nanometers: effectively, we are only observing a cut-off of the signal with a finite bandwith.

Definition 19.2 (*Bandlimited function*)

For any $\Omega > 0$, the space of Ω-bandlimited functions $\mathcal{B}_\Omega$ is the set of all $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ such that f is continuous and $\hat{f}(\xi) = 0$ for all $\xi \notin [-\Omega\pi, \Omega\pi]$.

Remark 19.3. An important theorem in Fourier analysis is the Paley-Wiener theorem, which relates decay properties of $f(t)$ when $|t| \to \infty$ with the analyticity of $\hat{f}(\xi)$[26]. In this context, bandlimited functions possess strong regularity properties (in particular they are $\mathcal{C}^\infty$), since they are the inverse Fourier transform of compactly-supported functions.

Whittaker-Kotelnikov-Shannon Sampling Theorem – We can now state the main theorem of this section:

Theorem 19.4 (*Whittaker-Kotelnikov-Shannon*)

Let $\Omega > 0$ and $f \in \mathcal{B}_\Omega$. Then for all $t \in \mathbb{R}$:

$$f(t) = \sum_{k \in \mathbb{Z}} f\left(\frac{k}{\Omega}\right) \frac{\sin(\pi(\Omega t - k))}{\pi(\Omega t - k)}. \tag{85}$$

Moreover we have:

$$\int_{\mathbb{R}} |f(t)|^2 dt = \frac{1}{\Omega} \sum_{k \in \mathbb{Z}} \left| f\left(\frac{k}{\Omega}\right) \right|^2.$$

[26] For example, if $tf(t)$ is integrable, then one can show that $\hat{f}'(\xi) = (2\pi)^{-1/2} \int (-it) f(t) e^{-it\xi} \mathrm{d}\xi$.

Proof of Theorem 19.4 – We prove here the theorem for $\Omega = 1$ for clarity of exposition. The proof for any $\Omega > 0$ follows with straightforward adaptations (which involve factors of Ω in many places and make the mathematical expressions less elegant, while being conceptually the same). Since $f \in \mathcal{B}$, $\hat{f} \in L^2([-\pi, \pi])$. The idea is to consider the Fourier Series of $\hat{f}$. We have

$$\hat{f}(\xi) = \mathbb{1}_{\{\xi \in [-\pi,\pi]\}} \frac{1}{\sqrt{2\pi}} \sum_{n \in \mathbb{Z}} a_n e^{in\xi}. \tag{86}$$

Moreover we have:

$$a_n = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \hat{f}(\xi)\, \overline{e^{in\xi}}\, \mathrm{d}\xi = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{f}(\xi)\, e^{i(-n)\xi} \mathrm{d}\xi = f(-n), \tag{87}$$

where we used the Fourier inversion formula in the last step. Indeed, since $\hat{f}$ is continuous (it is easy to see since $f \in L^1$) and is compactly supported, we also have $\hat{f} \in L^1(\mathbb{R})$. Furthermore, by the Fourier inversion formula, we have, for any $t \in \mathbb{R}$,

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{f}(\xi) e^{it\xi} d\xi,$$

using (86) and (87) we have

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \left(\frac{1}{\sqrt{2\pi}} \sum_{n \in \mathbb{Z}} f(-n) e^{in\xi} \right) e^{it\xi} d\xi.$$

Using Fubini and the change of indexing $n \leftrightarrow -n$, we have

$$f(t) = \frac{1}{2\pi} \sum_{n \in \mathbb{Z}} f(n) \int_{-\pi}^{\pi} e^{-in\xi} e^{it\xi} d\xi.$$

A simple calculation gives

$$\int_{-\pi}^{\pi} e^{-in\xi} e^{it\xi} d\xi = \frac{\sin(\pi(t-n))}{\pi(t-n)},$$

which completes the proof of the first part of the Theorem.

In Signal Processing this function is usually referred to as $\mathrm{sinc}(x) \coloneqq \frac{\sin(x)}{x}$ (and $\mathrm{sinc}(0) = 1$).

The second part of the theorem is obtained using that $\sum |a_k|^2 = \sum |f(k)|^2 = \int |\hat{f}(\xi)|^2 \mathrm{d}\xi$ and the Plancherel theorem. □

Remark 19.5. Theorem 19.4 shows that when f is Ω-bandlimited, it is uniquely determined (and can be reconstructed) using discrete samples taken at the frequency Ω (i.e. samples in $\mathbb{Z}/\Omega$). This is known as the *Nyquist rate*. Conversely, Theorem 19.4 also shows that one can transmit a square-summable sequence of numbers (a_k) at a frequency Ω by representing them as the samples of a Ω-bandlimited function for which we have an explicit form.

Further Reading 19.6. The Nyquist rate is actually optimal, in the sense that a stable reconstruction of $f \in \mathcal{B}_\Omega$ from samples taken with frequency $\omega < \Omega$ is in general impossible. This was shown by Landau in a beautiful paper [Lan67]. This is an area where the notion of Frame (in infinite dimensions) is important. Asking whether one can reconstruct a bandlimited signal f with samples $\{a_n\}_{n \in \mathbb{N}}$ essentially corresponds to asking whether $\left\{ x \mapsto \frac{1}{\sqrt{2\pi}} e^{ia_n x} \right\}_{n \in \mathbb{Z}}$ forms a spanning set. Asking for stable reconstruction corresponds to asking that it forms a frame (in the sense of Definition 11.1, but in infinite dimensions). It turns out that it is possible to have spanning sets with less frequency of samples, but not frames; this illustrats why the frame condition generalizing orthogonal basis in Hilbert spaces, in this context.

20 The Discrete Fourier Transform

We will now (re)introduce a related object, the Discrete Fourier Transform.

Let us consider Fourier Series ((83) and Definition 19.1) in a grid $x = \frac{k}{N}2\pi$ for an integer $N > 0$ and $k = 0, \cdots, N-1$. In this section it eases notation to identify f with a function in $[0, 2\pi]$ (both can be identified with a 2π-periodic function in $\mathbb{R}$).

$$f\left(k\frac{2\pi}{N}\right) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{\left(in\frac{k}{N}2\pi\right)} = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{\left(i\frac{nk}{N}2\pi\right)}. \tag{88}$$

Since, for any integer a, $e^{\left(i\frac{nk}{N}2\pi\right)} = e^{\left(i\frac{(n+aN)k}{N}2\pi\right)}$, there are only N different exponentials in the sum. We can rewrite the sum as

$$f\left(k\frac{2\pi}{N}\right) = \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{N-1} \left[e^{\left(i\frac{nk}{N}2\pi\right)} \left(\sum_{a=-\infty}^{\infty} \hat{f}(n+aN) \right) \right]. \tag{89}$$

Let us ease notation by taking $\omega_N = e^{-\frac{2\pi i}{N}}$ to be the N-root of unity, $x \in \mathbb{C}^n$ given by $x_k = f\left(k\frac{2\pi}{N}\right)$ and $y \in \mathbb{C}^N$ given by $y_k = \frac{1}{\sqrt{2\pi}} \sum_{a=-\infty}^{\infty} \hat{f}(k+aN)$. Then

$$x = Ty, \tag{90}$$

where $T \in \mathbb{C}^{N\times N}$ is given by $T_{ab} \coloneqq \omega_N^{-(a-1)(b-1)}$. A multiple of the Hermitian conjugate (also referred to as adjoint) is the celebrated Discrete Fourier Transform.

Definition 20.1 (*Discrete Fourier Transform*)

The matrix $F \in \mathbb{C}^{N\times N}$ given by $F_{ab} = N^{-1/2}\omega_N^{(a-1)(b-1)} = N^{-1/2}e^{-\frac{2\pi i(a-1)(b-1)}{N}}$ is known as the Discrete Fourier Transform (DFT) Matrix.

It shares many of the properties of the objects described above (Fourier Transform and Fourier Series). Since it is a linear transformation in finite dimensions (a matrix) many of its properties are easily described in terms of classical matrix properties. For example, the fact that F is a Unitary matrix immediately implies that F is an isometry (a Parseval/Plancherel-style Theorem) and that its inverse if F^* (the analogue to the Fourier inverse formula). Notice also that T above, in (90), is given by $T = \sqrt{N}F^*$.

Challenge 20.1. *Show that the matrix F defined above is a Unitary matrix, i.e. $F^*F = I$.*

Remark 20.2. As we mentioned a couple of lectures ago, one of the reasons Fourier theory is so ubiquitous is that the (Discrete) Fourier Transform essentially corresponds to the change of basis (of the space of functions) that diagonalizes translations. This can be readily viewed in the discrete setting. The unitary matrix F simultaneously diagonalizes shift matrices (and thus all circulant matrices — see the challenge below). This observation allows one to develop Fourier Theory to other groups, which is tightly connected to "Representation Theory of Groups" (in particular, "characters of Abelian groups"). Viewed in this abstract algebraic light, the Fourier Transform, Fourier Series, and the Discrete Fourier Transform correspond to different groups (translations in $\mathbb{R}$, cyclic translations in $\mathbb{S}^1$ (the torus $[-\pi, \pi]$)), and cyclic translations in $\mathbb{Z}/n\mathbb{Z}$). If you are interested in learning more, look up also "Harmonic Analysis", "Pontryagin Duality", "Spherical Harmonics", and "Peter–Weyl Theorem".

Challenge 20.2. *A matrix M is circulant if $M_{ij} = M_{kl}$ whenever $i - j = k - l$.*

1. *Show that for any circulant M we have that FMF^* is a diagonal matrix.*
2. *What are the diagonal entries of FMF^*?*

21 The Paley ETF and some Number Theory

The goal os this section is to continue in the spirit of these lecture notes and show a connection with yet another field of Mathematics, Number Theory.

This exposition on the Paley ETF and an introduction to Number Theory was adapted from a lecture written by Antoine Maillard in [BM22]. The errors/typos are all mine.

21.1 A bit of number theory

Recall that for any $p \geq 2$, $\mathbb{Z}_p$ (also denoted $\mathbb{Z}/p\mathbb{Z}$) is the cyclic group of order p (under addition) of integers modulo p. Moreover, if p is prime then $\mathbb{Z}_p$ is a field, and we denote $\mathbb{Z}_p^\times := \mathbb{Z}_p \backslash \{0\}$ the multiplicative group. In Appendix C we recall (and show) some basics of number theory, in particular the classical result that $\mathbb{Z}_p^\times$ is a cyclic group, i.e. there is an element $g \in \mathbb{Z}_p^\times$ (called a *generator*) such that $\mathbb{Z}_p^\times = \{g^k,\, 1 \leq k \leq p-1\}$. For a reference on number theory, you can check out the introductory Number Theory course at ETH!

Definition 21.1 (*Quadratic residue*)

Let $p \geq 3$ be a prime number. We say that an integer $x \in \mathbb{Z}$ is a quadratic residue mod p if there exists $q \in \mathbb{Z}$ such that $x \equiv q^2 \bmod p$. Otherwise, we say that x is a quadratic non-residue (mod p).

Definition 21.2 (*Legendre symbol*)

Let $p \geq 3$ be a prime number, and $a \in \mathbb{Z}$. We define the Legendre symbol as:

$$\left(\frac{a}{p}\right) := \begin{cases} 1 & \text{if } a \text{ is a quadratic residue mod } p \text{ and } a \not\equiv 0 \bmod p, \\ -1 & \text{if } a \text{ is a non quadratic residue mod } p, \\ 0 & \text{if } a \equiv 0 \bmod p. \end{cases} \tag{91}$$

We will need the following properties of the Legendre symbol (or of the quadratic residues).

Proposition 21.3 (*Properties of quadratic residues*)

Let $p \geq 3$ be a prime number. Then:

(*i*) (*Euler's criterion*) For all $a \in \mathbb{Z}$,

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \bmod p. \tag{92}$$

(*ii*) (*Multiplicativity*) For all $a, b \in \mathbb{Z}$,

$$\left(\frac{a}{p}\right)\left(\frac{b}{p}\right) = \left(\frac{ab}{p}\right). \tag{93}$$

(*iii*) Quadratic residues form exactly half the elements of $\mathbb{Z}_p^\times$:

$$\sum_{a \in \mathbb{Z}_p^\times} \left(\frac{a}{p}\right) = 0. \tag{94}$$

(*iv*) We have

$$\left(\frac{-1}{p}\right) = \begin{cases} 1 & \text{if } p \equiv 1 \bmod 4, \\ -1 & \text{if } p \equiv 3 \bmod 4. \end{cases}$$

Remark 21.4. Recall that Fermat's little theorem states that for all $a \in \mathbb{Z}$, if $a \not\equiv 0 \bmod p$ then $a^{p-1} \equiv 1 \bmod p$, which we recover by Euler's criterion (i).

Proof of Proposition 21.3 – We start by proving (i). If $a \equiv 0 \bmod p$ the statement is clear, so we assume $a \not\equiv 0 \bmod p$. Assume first that a is a quadratic non-residue mod p. Then the collection of pairs $\{\{x, ax^{-1}\}\}_{x\in\mathbb{Z}_p^\times}$ partitions $\mathbb{Z}_p^\times$ into $(p-1)/2$ pairs (since a is a non-residue, the two elements in the pair are always distinct). Moreover, the product of the elements in every pair is always given by a. Therefore we have $1 \times 2 \times \cdots \times (p-1) \equiv a^{(p-1)/2} \bmod p$. Since $(p-1)! \equiv -1 \bmod p$ when p is prime (this is called Wilson's theorem, see Theorem C.5), we reach $-1 \equiv a^{(p-1)/2} \bmod p$.

Let us now assume that a is a quadratic residue, so $a \equiv r^2 \bmod p$ for some $r \in \mathbb{Z}_p^\times$. Since the equation $a \equiv x^2 \bmod p$ has only two solutions $x = \pm r$ in the field $\mathbb{Z}_p^\times$ (see Theorem C.4)[27], we can now partition $\mathbb{Z}_p^\times \backslash \{-r, r\}$ by using $(p-3)/2$ pairs: $\{\{x, ax^{-1}\}\}_{x\in\mathbb{Z}_p^\times\backslash\{-r,r\}}$. Multiplying all elements of $\mathbb{Z}_p^\times$, we reach $(p-1)! \equiv a^{(p-3)/2} \times (-r^2) \bmod p$. Applying again Wilson's theorem, we reach $1 \equiv a^{(p-1)/2} \bmod p$.

Given (i), statement (ii) is immediate, so we now prove (iii). Let $g \in \mathbb{Z}_p^\times$ be a generator of $\mathbb{Z}_p^\times$, i.e. $\mathbb{Z}_p^\times = \{g^k, k \in \{0, \cdots, p-2\}\}$ (see Lemma C.8). Since all g^k for $0 \le k \le p-2$ are distinct and $(p-1)/2 \le p-2$, we can not have $g^{(p-1)/2} \equiv 1 \bmod p$. Therefore, by (i), the only possibility is that $g^{(p-1)/2} \equiv -1 \bmod p$, i.e. g is a quadratic non-residue. Moreover, we have

$$\sum_{a\in\mathbb{Z}_p^\times} \left(\frac{a}{p}\right) \in [-(p-1), p-1] \qquad \text{and} \qquad \begin{aligned} \sum_{a\in\mathbb{Z}_p^\times} \left(\frac{a}{p}\right) &= \sum_{k=0}^{p-2} \left(\frac{g^k}{p}\right), \\ &\equiv \sum_{k=0}^{p-2} g^{\frac{k(p-1)}{2}} \bmod p, \\ &\equiv \sum_{k=0}^{p-2} (-1)^k \bmod p, \\ &\equiv 0 \bmod p. \end{aligned}$$

Point (iv) is easy to check. For $p \equiv 1 \bmod 4$ we have $p = 4q+1$ (for some non-zero integer q) and so, by Euler's criterion, $\left(\frac{-1}{p}\right) \equiv (-1)^{2q} \equiv 1 \bmod p$. On the other hand, for $p \equiv 3 \bmod 4$ we have $p = 4q+3$ and so, by Euler's criterion, $\left(\frac{-1}{p}\right) \equiv (-1)^{2q+1} \equiv -1 \bmod p$. □

Proposition 21.3 allows us to deduce an important property of a quantity which is a called the *quadratic Gauss sum*:

Theorem 21.5 (*Gauss sums*)

Let $p \ge 3$ be a prime number, and let $\omega := e^{2i\pi l/p}$, for some $l \not\equiv 0 \bmod p$. Then

$$\left[\sum_{k=0}^{p-1} \left(\frac{k}{p}\right)\omega^k\right]^2 = p\left(\frac{-1}{p}\right) = \begin{cases} p & \text{if } p \equiv 1 \bmod 4, \\ -p & \text{if } p \equiv 3 \bmod 4. \end{cases}$$

Remark 21.6. Since

$$\sum_{k=0}^{p-1} \left(\frac{k}{p}\right)\omega^k = \sum_{\substack{1\le k\le p-1 \\ \left(\frac{k}{p}\right)=1}} \omega^k - \sum_{\substack{1\le k\le p-1 \\ \left(\frac{k}{p}\right)=-1}} \omega^k = 2\sum_{\substack{1\le k\le p-1 \\ \left(\frac{k}{p}\right)=1}} \omega^k - \left(\sum_{k=0}^{p-1} \omega^k\right) + \omega^0 = \sum_{r=1}^{p-1} \omega^{r^2} + \omega^0 = \sum_{r=0}^{p-1} e^{2i\pi l r^2/p},$$

this equality sometimes appears written in terms of sums of roots of unity with quadratic exponents.

[27] In this case it is very easy to see, since $x^2 \equiv r^2 \bmod p \Leftrightarrow (x-r)(x+r) \equiv 0 \bmod p$, and $\mathbb{Z}_p$ is a field.

Proof of Theorem 21.5 – Let $g_p(x) \coloneqq \sum_{k=0}^{p-1} \left(\frac{k}{p}\right) x^k$. We have

$$g_p(x)^2 = \sum_{j,k=0}^{p-1} \left(\frac{k}{p}\right)\left(\frac{j}{p}\right) x^{j+k}.$$

For any u such that $u^p = 1$ (in particular this hold for $u = \omega$ and $u = 1$), we have $u^{j+k} = u^{j+k(\bmod p)}$, thus we can group terms:

$$g_p(u)^2 = \sum_{n=0}^{p-1} \left[\sum_{\substack{0\leq j,k\leq p-1 \\ j+k\equiv n \bmod p}} \left(\frac{k}{p}\right)\left(\frac{j}{p}\right) \right] u^n.$$

Let us denote a_n the element in front of u^n in this sum, i.e. $g_p(u)^2 = \sum_{n=0}^{p-1} a_n u^n$. By (iii) of Property 21.3, $g_p(1) = 0$, thus $\sum_{n=0}^{p-1} a_n = 0$. Moreover, we can compute, using (ii) of Proposition 21.3:

$$a_0 = \sum_{j=0}^{p-1} \left(\frac{j}{p}\right)\left(\frac{-j}{p}\right) = \sum_{j=0}^{p-1} \left(\frac{j^2}{p}\right)\left(\frac{-1}{p}\right) = \sum_{j=1}^{p-1} \left(\frac{-1}{p}\right) = (p-1)\left(\frac{-1}{p}\right).$$

Finally, let $1 \leq n \leq p-1$. Letting $j = nj'$ and $k = nk'$ with $j', k' \in \mathbb{Z}_p^\times$ (and identifying the integers j, k, n with the corresponding element of $\mathbb{Z}_p$), we have

$$\begin{aligned} a_n &= \sum_{\substack{0\leq j',k'\leq p-1 \\ j'+k'\equiv 1 \bmod p}} \left(\frac{nk'}{p}\right)\left(\frac{nj'}{p}\right) \\ &= \sum_{\substack{0\leq j',k'\leq p-1 \\ j'+k'\equiv 1 \bmod p}} \left(\frac{n^2}{p}\right)\left(\frac{k'}{p}\right)\left(\frac{j'}{p}\right) \\ &= a_1. \end{aligned}$$

Therefore $a_1 = a_2 = \cdots = a_{p-1}$. Combining the different results above, we reach that

$$\begin{cases} a_0 &= (p-1)\left(\frac{-1}{p}\right), \\ a_n &= -\left(\frac{-1}{p}\right) \quad \forall n \in \{1, \cdots, p-1\}. \end{cases}$$

Thus

$$g_p(\omega)^2 = p\left(\frac{-1}{p}\right) - \left(\frac{-1}{p}\right)\sum_{n=0}^{p-1} \omega^n = p\left(\frac{-1}{p}\right),$$

since $\sum_{n=0}^{p-1} \omega^n = 0$ is a basic property of p-th roots of unity. □

Remark 21.7. A beautiful fact about quadratic residues (which we will neither prove nor use) and the Legendre symbol is *quadratic reciprocity*, which states that for any two distinct odd primes p and q

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}.$$

This fact is known to have many proofs, in fact there is a one page arxiv paper [Vek21] with a (very short) proof!

21.2 Definition of the Paley ETF

We now introduce the Paley ETF. Other brief descriptions of its construction (with slightly different conventions) can be found e.g. in [Ban16, BFMW13].

Its construction is not mathematically complicated, but involves several steps. Let $p \geq 1$ be a prime number such that $p \equiv 1 \bmod 4$, and let $M := (p+1)/2$ and $N := 2M = p+1$. Recall the definition of the Discrete Fourier Transform matrix in Definition 20.1, and we let F be the $p \times p$ DFT matrix. To lighten some notations, we use $\{0, \cdots, p-1\}$ rather than $\{1, \cdots, p\}$ to index its rows and columns. We have for $0 \leq a, b \leq p-1$:

$$F_{ab} = \frac{1}{\sqrt{p}} e^{-\frac{2i\pi ab}{p}}.$$

Let $S = \{0\} \cup Q$, with $Q \subseteq \{1, \cdots, p-1\}$ the subset of quadratic residues modulo p, cf. Definition 21.1. By Proposition 21.3-(iii), $|S| = M$. We define G as the $M \times p$ matrix formed by picking rows of F whose index is in S, i.e. if we denote $S = \{i_0, \cdots, i_{M-1}\} \subseteq [p-1]$, we have for $0 \leq k \leq M-1$ and $0 \leq l \leq p-1$:

$$G_{kl} := F_{i_k, l} = \frac{1}{\sqrt{p}} e^{-\frac{2i\pi i_k l}{p}}. \tag{95}$$

We end the construction by two steps:

- (i) We let $H := DG$, with $D \in \mathbb{R}^{M \times M}$ the diagonal matrix whose elements are $D_{00} = 1$, and $D_{kk} = \sqrt{2}$ for $1 \leq k \geq M-1$ (recall that we are indexing rows with $\{0, \cdots, p-1\}$). Effectively, this multiplies all elements of G by $\sqrt{2}$, except the ones in the first row.
- (ii) We build $\Phi \in \mathbb{C}^{M \times N}$ (i.e. a $M \times 2M$ matrix) by concatenating the columns of H with the canonical basis element $(1, 0, \cdots, 0) \in \mathbb{R}^M$.

As we will see below, these two steps are necessary: step (i) ensures that the columns of Φ have unit norm and satisfy $\mu = |\langle \phi_i, \phi_j \rangle|$ for all $i \neq j$, and step (ii) ensures that the frame is tight. Anticipating on what we will later prove, we call this construction the Paley ETF [BFMW13].

Definition 21.8 (*Paley ETF*)

For any prime $p \geq 5$ such that $p \equiv 1 \bmod 4$, the columns of the matrix Φ built by the procedure above are called the *Paley Equiangular Tight Frame*.

Example 21.9. Let $p = 5$. It is easy to check that $S = \{0, 1, 4\}$. Therefore we have:

$$G = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & e^{-\frac{2i\pi}{5}} & e^{-\frac{4i\pi}{5}} & e^{-\frac{6i\pi}{5}} & e^{-\frac{8i\pi}{5}} \\ 1 & e^{-\frac{8i\pi}{5}} & e^{-\frac{6i\pi}{5}} & e^{-\frac{4i\pi}{5}} & e^{-\frac{2i\pi}{5}} \end{pmatrix}.$$

This leads to the construction:

$$\Phi = \begin{pmatrix} \sqrt{\frac{1}{5}} & \sqrt{\frac{1}{5}} & \sqrt{\frac{1}{5}} & \sqrt{\frac{1}{5}} & \sqrt{\frac{1}{5}} & 1 \\ \sqrt{\frac{2}{5}} & \sqrt{\frac{2}{5}} e^{-\frac{2i\pi}{5}} & \sqrt{\frac{2}{5}} e^{-\frac{4i\pi}{5}} & \sqrt{\frac{2}{5}} e^{-\frac{6i\pi}{5}} & \sqrt{\frac{2}{5}} e^{-\frac{8i\pi}{5}} & 0 \\ \sqrt{\frac{2}{5}} & \sqrt{\frac{2}{5}} e^{-\frac{8i\pi}{5}} & \sqrt{\frac{2}{5}} e^{-\frac{6i\pi}{5}} & \sqrt{\frac{2}{5}} e^{-\frac{4i\pi}{5}} & \sqrt{\frac{2}{5}} e^{-\frac{2i\pi}{5}} & 0 \end{pmatrix}.$$

Challenge 21.1. *Check that the matrix Φ built above for the case $p = 5$ is indeed an ETF.*

We now prove the main theorem of this section, which is that the Paley ETF is indeed an equiangular tight frame.

Theorem 21.10

For any prime $p \geq 5$, the Paley ETF is a complex equiangular tight frame.

Remark 21.11. Theorem 21.10 does not actually require $p \equiv 1 \bmod 4$ to hold. On the other hand, it is only in this case that the Paley ETF can be mapped to a real ETF, see below.

Proof of Theorem 21.10 – We denote $\phi_1, \cdots, \phi_N$ the columns of Φ. Trivially, $\|\phi_N\| = 1$. For any $i \in [N-1]$, we have

$$\|\phi_i\|^2 = \frac{1}{p} + \sum_{k=1}^{M-1} \frac{2}{p} = \frac{1 + 2(M-1)}{p} = 1.$$

We now show that Φ is tight. One checks easily from the properties of the DFT matrix that the rows of Φ are pairwise orthogonal, and that they all have squared norm 2. Therefore we have $\Phi\Phi^\dagger = 2\mathrm{I}_M$, i.e. Φ is a tight frame.

To prove that Φ is an ETF, it is thus sufficient show that for all $i, j \in [N]$, $|\langle \phi_i, \phi_j \rangle|$ takes the same value. Let us denote $\mu = 1/\sqrt{2M-1} = 1/\sqrt{p}$, i.e. the minimal worst-case coherence given by the Welch bound 11.6 for a unit-norm frame of $N = 2M$ vectors in $\mathbb{C}^M$. In what follows, we show successively:

(*i*) For all $j \in [N-1]$, $|\langle \phi_j, \phi_N \rangle| = \mu$.

(*ii*) For all $j, j' \in [N-1]$ with $j \neq j'$, $|\langle \phi_j, \phi_{j'} \rangle| = \mu$.

The propery (*i*) is trivial, since $|\langle \phi_j, \phi_N \rangle| = |(\phi_j)_1| = 1/\sqrt{p}$. We focus on (*ii*). We have

$$\begin{aligned}
\langle \phi_j, \phi_{j'} \rangle &= \frac{1}{p} + \frac{2}{p} \sum_{k=1}^{p-1} \mathbb{1}\left\{ \left(\frac{k}{p}\right) = 1 \right\} e^{-\frac{2i\pi k(j-j')}{p}}, \\
&\overset{(a)}{=} \frac{1}{p} + \frac{1}{p} \sum_{k=1}^{p-1} \left[1 + \left(\frac{k}{p}\right) \right] e^{-\frac{2i\pi k(j-j')}{p}}, \\
&= \frac{1}{p} \sum_{k=0}^{p-1} e^{-\frac{2i\pi k(j-j')}{p}} + \frac{1}{p} \sum_{k=1}^{p-1} \left(\frac{k}{p}\right) e^{-\frac{2i\pi k(j-j')}{p}}, \\
&= \frac{1}{p} \sum_{k=0}^{p-1} \left(\frac{k}{p}\right) e^{-\frac{2i\pi k(j-j')}{p}},
\end{aligned}$$

where we used in (a) that $\left(\frac{k}{p}\right) \in \{\pm 1\}$. Therefore by Theorem 21.5, we have $\langle \phi_j, \phi_{j'} \rangle^2 = \pm 1/p$, and thus $|\langle \phi_j, \phi_{j'} \rangle| = 1/\sqrt{p}$. □

Remark 21.12 (The real Paley ETF)**.** When $p \equiv 1 \bmod 4$, the proof above shows that the inner product between any two elements of the frame is real, i.e. that $\Phi^\dagger\Phi$ is a real matrix (and recall it is positive semidefinite). We can thus write $\Phi^\dagger\Phi = \Psi^\top\Psi$, in which Ψ is a real matrix given by the Cholesky decomposition of $\Phi^\dagger\Phi$. Since Φ form an ETF, one can then check easily that the columns of Ψ form a real Equiangular Tight Frame. For this reason, the Paley ETF is often studied in the case $p \equiv 1 \bmod 4$.

Further Reading 21.13 (Real ETFs and strongly regular graphs)**.** There is a fascinating and general connection between real ETFs and a class of regular graphs known as *strongly regular graphs*, see e.g. Theorem 19 in [BFMW13]. The latter are defined as d-regular graphs, such that the common number of neighbors of any two vertices i, j only depends on whether i and j are adjacent or not. In particular, the Paley ETF is mapped to a graph on p vertices, known as the Paley graph, such that vertices i and

j are connected iff $i - j$ is a quadratic residue mod p. One can check that this definition is consistent when $p \equiv 1 \bmod 4$ (i.e. $i \sim j \Leftrightarrow j \sim i$), and then prove that this is indeed a strongly regular graph. For more details on the Paley graph and strongly regular graphs, see e.g. this excellent book draft by Daniel Spielman [Spi25] (available at http://cs-www.cs.yale.edu/homes/spielman/sagt/) which is also a great reference for more notions of spectral and algebraic graph theory.

Further Reading 21.14. The Paley ETF is related to several fascinating questions. For example, it is tantalizing to conjecture that the Paley ETF satisfies the Restricted Isometry Property (RIP) for sparsity patterns beyond what the incoherence μ would imply (see [BFMW13] and Open Problem 6.4. in [Ban16]). In a sense, we expect quadratic residues to behave (pseudo)randomly, and we know that random subsets of F are good RIP matrices (although determining precisely "how good" they are is also open: Open Problem 6.1 in [Ban16]). In this particular case, (pseudo)randomness of the Paley ETF is tightly connected to a fascinating open question in Number Theory: the clique number of the Paley Graph (defined in Further Reading 21.13) [BMM17]. It is conjectured that the clique number is of poly-logarithmic size (see Open Problem 8.4. in [Ban16]) while the current best bound [HP21] is $\sqrt{p/2}$ (in fact, one of my most recent papers [BBD+25] is inspired by this question [Kun24]). From these references you can find your way to lots of fascinating work on this question. I would also recommend a course here at ETH on Probabilistic Number Theory (see this book [Kow21], available at https://people.math.ethz.ch/~kowalski/probabilistic-number-theory.pdf).

22 Group Testing

Remark 22.1. Part of this Section is borrowed from my earlier notes [Ban16] which were written before the COVID-19 pandemic. The pandemic motivated new interest in this problem.

During the Second World War the United States was interested in determining which soldiers called up for the army had syphilis. However, syphilis testing back then was expensive and testing every soldier individually would have been very costly and inefficient. A basic breakdown of a test is: 1) Draw sample from a given individual, 2) Perform required tests, and 3) Determine presence or absence of syphilis.[28]

If there are n soldiers, this method of testing leads to n tests. If a significant portion of the soldiers were infected then the method of individual testing would be reasonable. The goal however, is to achieve effective testing in the more likely scenario where it does not make sense to test n (say $n = 100,000$) people to get k (say $k = 10$) positives.

Let's say that it was believed that there is only one soldier infected, then one could mix the samples of half of the soldiers and with a single test determined in which half the infected soldier is, proceeding with a binary search we could pinpoint the infected individual in $\log n$ tests. If instead of one, one believes that there are at most k infected people, then one could simply run k consecutive binary searches and detect all of the infected individuals in $k \log n$ tests. Which would still be potentially much less than n.

For this method to work one would need to observe the outcome of the previous tests before designing the next test, meaning that the samples have to be prepared adaptively. This is often not practical, if each test takes time to run, then it is much more efficient to run them in parallel (at the same time). This means that one has to non-adaptively design T tests (meaning subsets of the n individuals) from which it is possible to detect the infected individuals, provided there are at most k of them. Constructing these sets is the main problem in (Combinatorial) Group testing, introduced by Robert Dorfman [Dor43] with essentially the motivation described above.[29]

Let A_i be a subset of $[T] = \{1, \ldots, T\}$ that indicates the tests for which soldier i participates. Consider $\mathcal{A}$ the family of n such sets $\mathcal{A} = \{A_1, \ldots, A_n\}$.

Definition 22.2 (*k-disjunct*)

We say that a family $\mathcal{A}$ of subsets of $[T]$ satisfies the k-disjunct property if no set in $\mathcal{A}$ is contained in the union of k other sets in $\mathcal{A}$.

A test set designed so that $\mathcal{A}$ is k-disjunct will succeed at identifying the (at most k) infected individuals – the set of infected tests is also a subset of $[T]$ and it will be the union of the A_i's that correspond to the infected soldiers. If the set of infected tests contains a certain A_i then this can only be explained by the soldier i being infected (provided that there are at most k infected people).

Theorem 22.3

Given n and k, there exists a family $\mathcal{A}$ satisfying the k-disjunct property for a number of tests

$$T = \mathcal{O}\left(k^2 \log n\right).$$

We will use the probabilistic method. We will show that, for $T = Ck^2 \log n$ (where C is a universal constant), by drawing the family $\mathcal{A}$ from a (well-chosen) distribution gives a $k-$disjunct family with positive probability, meaning that such a family must exist (otherwise the probability would be zero).

[28] A more recent motivation/application is COVID testing; mathematically the problem is the same. There are also many applications not related with disease testing.

[29] in fact, our description for the motivation of Group Testing very much follows the description in [Dor43].

Let $0 \leq p \leq 1$ and let $\mathcal{A}$ be a collection of n random (independently drawn) subsets of $[T]$. The distribution for a random set A is such that each $t \in [T]$ belongs to A with probability p (and independently of the other elements).

Consider $k+1$ independent draws of this random variable, $A_0, \ldots, A_k$. The probability that A_0 is contained in the union of A_1 through A_k is given by

$$\Pr\left[A_0 \subseteq (A_1 \cup \cdots \cup A_k)\right] = \left(1 - p(1-p)^k\right)^T.$$

This is minimized for $p = \frac{1}{k+1}$. For this choice of p, we have

$$1 - p(1-p)^k = 1 - \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right)^k$$

Given that there are n such sets, there are $(k+1)\binom{n}{k+1}$ different ways of picking a set and k others to test whether the first is contained in the union of the other k. Hence, using a union bound argument, the probability that $\mathcal{A}$ is k-disjunct can be bounded as

$$\Pr[k-\text{disjunct}] \geq 1 - (k+1)\binom{n}{k+1}\left(1 - \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right)^k\right)^T.$$

In order to show that one of the elements in $\mathcal{A}$ is k-disjunct we show that this probability is strictly positive. That is equivalent to

$$\left(1 - \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right)^k\right)^T \leq \frac{1}{(k+1)\binom{n}{k+1}}.$$

Note that $\left(1 - \frac{1}{k+1}\right)^k \to e^{-1}\frac{1}{1-\frac{1}{k+1}} = e^{-1}\frac{k+1}{k}$, as $k \to \infty$. Thus, we only need

$$T > \frac{\log\left((k+1)\binom{n}{k+1}\right)}{-\log\left(1 - \frac{1}{k+1}e^{-1}\frac{k+1}{k}\right)} \geq \frac{\log\left(k\binom{n}{k+1}\right)}{-\log\left(1 - (ek)^{-1}\right)} = \mathcal{O}(k^2 \log(n/k)),$$

where the last inequality uses the fact that $\log\left(\binom{n}{k+1}\right) = \mathcal{O}\left(k \log\left(\frac{n}{k}\right)\right)$ due to Stirling's formula and the Taylor expansion $-\log(1-x^{-1})^{-1} = \mathcal{O}(x)$ □

This argument simply shows the existence of a family satisfying the k-disjunt property. However, it is easy to see that by having T slightly larger one can ensure that the probability that the random family satisfies the desired property can be made very close to 1.

Remarkably, the existence proof presented here is actually very close to the best known lower bound.

Theorem 22.4

Given n and k, if there exists a family $\mathcal{A}$ of subsets of $[T]$ satisfying the k-disjunct property, then

$$T = \Omega\left(\frac{k^2 \log n}{\log k}\right).$$

We prove Theorem 22.4 in Appendix D (the proof is a neat combinatorial argument).

There is another upper bound, incomparable to the one in Theorem 22.3 that is known.

Theorem 22.5

Given n and k, there exists a family $\mathcal{A}$ satisfying the k-disjunct property for a number of tests

$$T = \mathcal{O}\left(k^2 \left(\frac{\log n}{\log k}\right)^2\right).$$

The proof of this Theorem uses ideas of Coding Theory (in particular Reed-Solomon codes), which we will not cover in this course (despite being another beautiful area of Mathematics with important applications), we defer the reader to [Ban16] for a proof of this upper bound, together with a crash course on coding theory, in the same notation as style as this section.

The following Corollary follows immediately.

Corollary 22.6

Given n and k, there exists a family $\mathcal{A}$ satisfying the k-disjunct property for a number of tests

$$T = \mathcal{O}\left(\frac{k^2 \log n}{\log k} \min\left\{\log k, \frac{\log n}{\log k}\right\}\right).$$

While the upper bound in Corollary 22.6 and the lower bound in Theorem 22.4 are quite close, there was still a gap. This gap was recently closed and Theorem 22.4 was shown to be optimal [DVPS14].

Remark 22.7. We note that the lower bounds established in Theorem 22.4 are not an artifact of the requirement of the sets being k-disjunct. For the measurements taken in Group Testing to uniquely determine a group of k infected individuals it must be that the there are no two subfamilies of at most k sets in $\mathcal{A}$ that have the same union. If $\mathcal{A}$ is not $k-1$-disjunct then there exists a subfamily of $k-1$ sets that contains another set A, which implies that the union of that subfamily is the same as the union of the same subfamily together with A. This means that a measurement system that is able to uniquely determine a group of k infected individuals must be $k-1$-disjunct.

22.1 In terms of ~~linear~~ Bernoulli algebra

We can describe the process above in terms of something similar to a sparse linear system. let 1_{A_i} be the $t-$*dimensional* indicator vector of A_i, $1_{i:n}$ be the (unknown) $n-$dimensional vector of infected soldiers and $1_{t:T}$ the $T-$dimensional vector of infected (positive) tests. Then

$$\begin{bmatrix} | & & | \\ 1_{A_1} & \cdots & 1_{A_n} \\ | & & | \end{bmatrix} \otimes \begin{bmatrix} | \\ | \\ 1_{i:n} \\ | \\ | \end{bmatrix} = \begin{bmatrix} | \\ 1_{t:T} \\ | \end{bmatrix},$$

where $\otimes$ is matrix-vector multiplication in the Bernoulli algebra, basically the only thing that is different from the standard matrix-vector multiplications is that the addition operation is replaced by binary "or", meaning $1 \oplus 1 = 1$.

This means that we are essentially solving a linear system (with this non-standard multiplication). Since the number of rows is $T = \mathcal{O}(k^2 \log(n/k))$ and the number or columns $n \gg T$ the system is underdetermined. Note that the unknown vector, $1_{i:n}$ has only k non-zero components, meaning it is $k-$sparse. Interestingly, despite the similarities with the setting of sparse recovery discussed in a previous lecture, in this case, $\tilde{\mathcal{O}}(k^2)$ measurements are needed, instead of $\tilde{\mathcal{O}}(k)$ as in the setting of Compressed Sensing.

A Rest of Proof of Bochner's Theorem

Proof of $(i) \Rightarrow (ii)$ in Theorem 18.2 – We prove this statement when $p = 1$, as the notations are lighter and the principle is exactly the same. Here, we give the proof for all dimensions $p \geq 1$. We will show that for all $T > 0$, we have

$$H_T(u) := \int_{[-T,T]^p} q(x)\, e^{-iu^\top x} \prod_{j=1}^{p} \left(1 - \frac{|x_j|}{T}\right) \mathrm{d}x \geq 0. \tag{96}$$

Let us describe how it allows to end the proof. Note that for all $x \in \mathbb{R}^p$, we have $q(x)e^{-iu^\top x} \prod_j (1 - |x_j|/T) \to q(x)e^{-iu^\top x}$ as $T \to \infty$. We can use the dominated convergence theorem (check the domination hypothesis!) to take the limit $T \to \infty$ in eq. (96). This yields that $\hat{q}(u) \geq 0$. Let us now prove eq. (96).

It is easy to see (prove it!) that for all $x \in \mathbb{R}$, one has

$$\begin{aligned}\left(1 - \frac{|x|}{T}\right)\mathbb{1}\Big\{|x| \leq T\Big\} &= \frac{1}{T}\int_{-T/2}^{T/2} \mathbb{1}\Big\{-\frac{T}{2} - x \leq \theta \leq \frac{T}{2} - x\Big\}\,\mathrm{d}\theta, \\ &= \frac{1}{T}\int_{-T/2}^{T/2} \mathbb{1}\Big\{-\frac{T}{2} - \theta \leq x \leq \frac{T}{2} - \theta\Big\}\,\mathrm{d}\theta.\end{aligned}$$

Therefore

$$\begin{aligned} H_T(u) &= \frac{1}{T^p}\int_{\mathbb{R}^p} q(x)\, e^{-iu^\top x}\left[\int_{[-T/2,T/2]^p} \prod_{j=1}^{p} \mathbb{1}\Big\{-\frac{T}{2} - \theta_j \leq x_j \leq \frac{T}{2} - \theta_j\Big\}\,\mathrm{d}\theta\right]\mathrm{d}x, \\ &\overset{(a)}{=} \frac{1}{T^p}\int_{[-T/2,T/2]^p}\left[\int_{\mathbb{R}^p} q(x)\, e^{-iu^\top x} \prod_{j=1}^{p} \mathbb{1}\Big\{-\frac{T}{2} - \theta_j \leq x_j \leq \frac{T}{2} - \theta_j\Big\}\,\mathrm{d}x\right]\mathrm{d}\theta, \\ &\overset{(b)}{=} \frac{1}{T^p}\int_{[-T/2,T/2]^p}\int_{[-T/2,T/2]^p} q(y-\theta)\, e^{-iu^\top(y-\theta)}\mathrm{d}y\,\mathrm{d}\theta. \end{aligned} \tag{97}$$

In (a) we used Fubini's theorem to change the order of the integrals, and in (b) we changed variables $x = y - \theta$. Since q is continuous, we can approximate the integral in eq. (97) by Riemann sums. For any $N \geq 1$, we partition the set $[-T/2, T/2]^p$ in N cells $C_1, \cdots, C_N$, such that each cell has volume $V(C_k) = T^p/N$. For each $k \in [N]$, we fix an arbitrary point $r_k \in C_k$. Riemann sums theory yields that we have:

$$\begin{aligned} H_T(u) &= \lim_{N\to\infty} \frac{T^p}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} q(r_k - r_l)e^{-iu^\top(r_k - r_l)}, \\ &= \lim_{N\to\infty} \frac{T^p}{N^2}\sum_{k,l=1}^{N} \overline{e^{iu^\top r_k}}\, q(r_k - r_l)e^{iu^\top r_l}. \end{aligned} \tag{98}$$

Since K is positive definite, the matrix $(q(r_k - r_l))_{k,l=1}^N$ is positive semi-definite. Since any real symmetric matrix is also Hermitian, for all $z \in \mathbb{C}^N$, we have $\sum_{k,l} \overline{z_k} q(r_k - r_l) z_l \geq 0$. Applying it for $z_k = e^{iu^\top r_k}$ in eq. (98) shows that $H_T(u) \geq 0$. □

B Alternative proof of the SSVC Theorem

In this Section we show an alternative proof of Theorem 17.7 based on the so-called *shifting* technique.
(Alternative) proof of Theorem 17.7 We use the approach based on the *shifting* technique. Fix any set of points $x_1, \ldots, x_n$ in $\mathcal{X}$. Set $V = \{(\mathbb{1}_{x_1 \in A}, \ldots, \mathbb{1}_{x_n \in A}) : A \in \mathcal{A}\} \subseteq \{0,1\}^n$. For $i = 1, \ldots, n$ consider the shifting operator $S_{i,V}$ acting on $(v_1, \ldots, v_n) \in V$ as follows:

$$S_{i,V}((v_1, \ldots, v_n)) = \begin{cases} (v_1, \ldots, v_{i-1}, 0, v_{i+1}, \ldots, v_n), & \text{if } (v_1, \ldots, v_{i-1}, 0, v_{i+1}, \ldots, v_n) \notin V; \\ (v_1, \ldots, v_n), & \text{otherwise.} \end{cases}$$

In words, $S_{i,V}$ changes the i-th coordinate 1 with 0 if this does not yields a copy of a vector that is already in V. Define $S_i(V) = \{S_{i,V}(v) : v \in V\}$. This means that we apply the shifting operator to all vectors in V. By our construction we have $|S_i(V)| = V$. Moreover, note that since $V \subseteq \{0,1\}^n$, it can be seen as a collection of sets of $\{1, \cdots, n\}$ (identifying $(v_1, \cdots, v_n)$ with the set $\{j \in [n] : v_j = 1\}$). With this view in mind, we have the following lemma.

Lemma B.1

Any set $I \subset \{1, \ldots, n\}$ shattered by $S_i(V)$ is also shattered by V.

Proof of Lemma B.1 – Take any set I shattered by $S_i(V)$. If $i \notin I$, then the claim follows immediately since the shifting operator does not affect this index. Otherwise, without loss of generality assume that $i = 1$ and $I = \{1, \ldots, k\}$. Since I is shattered by $S_1(V)$, for any $u \in \{0,1\}^k$ there is $v \in S_1(V)$ such that $v_i = u_i$ for $i = 1, \ldots, k$. If $u_1 = 1$, then both v and $v' = (0, v_2, \ldots, v_n)$ belong to V since otherwise v would have been shifted. Thus, for any $u \in \{0,1\}^k$ there is $w \in V$ such that $w_i = u_i$ for $i = 1, \ldots, k$. This means that I is also shattered by V. □

Starting from the set V, we apply shifting repeatedly to all $i \in \{1, \cdots, n\}$ until no shifts are possible. That is, we reach the set V' such that $S_i(V') = V'$ for all $i = 1, \ldots, n$. This happens because whenever a nontrivial shift happens, the total number of 1-s in V decreases, so this procedure has to stop.

Finally, we prove that V' contains no vector with more than d 1-s. Indeed, let us assume that there is a vector $v \in V'$ with $k > d$ 1-s. Then the set of these k coordinates is shattered by V': it is easy to see that otherwise shifting would have reduced the number of 1-s in v. By Lemma B.1, this implies that the same subset of size $k > d$ is also shattered by V. We obtain a contradiction with the fact that the VC dimension of $\mathcal{A}$ is equal to d.

Since V' is included in the set of vectors with at most d 1-s, we have:

$$|V'| \leq \sum_{i=0}^{d} \binom{n}{i}.$$

The claim follows since $|V| = |V'|$. □

C Some elements of number theory

We first recall some basic definitions of group theory, here specified to the case of the group $\mathbb{Z}_p^\times$.[30] This appendix culminates with the proof that $\mathbb{Z}_p^\times$ is cyclic (meaning that there is an element whose powers give all elements in $\mathbb{Z}_p^\times$), if you have seen the proof of this fact, feel free to skip this appendix.

C.1 Order of a group element

Definition C.1 (*Order of an element*)

Let $a \in \mathbb{Z}_p^\times$. The *order* of a, denoted $|a|$, is the smallest $k \geq 1$ such that $a^k \equiv 1 \bmod p$.

By Fermat's Little Theorem,[31] we know that the order of any element can not be higher than $p-1$:

Theorem C.2 (*Fermat's little theorem*)

Let $p \geq 2$ be a prime, and $a \in \mathbb{Z}_p^\times$. Then $a^{p-1} \equiv 1 \bmod p$.

This yields the easy corollary, a particular case of Lagrange's theorem:

Corollary C.3

Let $p \geq 2$ be a prime, and $a \in \mathbb{Z}_p^\times$. Then $|a|$ divides $p-1$.

Note that this fact is a general result in group theory, a corollary of Lagrange's theorem: the order of each element must divide the cardinality of the group, here $p-1$.

Proof of Corollary C.3 – We know that $|a| \leq p-1$. We denote $p-1 = k|a|+r$ the Euclidean division of $p-1$ by $|a|$, with $0 \leq r < |a|$. Then $a^{p-1} \equiv a^{k|a|+r} \bmod p \equiv a^r \bmod p$. By Fermat's little theorem, we thus have $1 \equiv a^r \bmod p$. But since $r < |a|$ we must have $r = 0$. □

C.2 Polynomials on $\mathbb{Z}_p$

Theorem C.4 (*Roots of a polynomial*)

Let $p \geq 2$ be prime, and let f be a polynomial function over $\mathbb{Z}_p$ (i.e. the coefficients of f are in $\mathbb{Z}_p$) of degree $n \geq 1$. Then the equation $f(x) \equiv 0 \bmod p$ has at most n solutions in $\mathbb{Z}_p$.

Proof of Theorem C.4 – The proof is by induction over the degree n. If $n = 1$, then $f(x) = ax + b$ with $a \not\equiv 0 \bmod p$ and it has a unique root $x = -a^{-1}b$. Assume that $n \geq 2$ and that the claim holds for $n-1$. Let f be a polynomial over $\mathbb{Z}_p$ of degree n. Assume that f has at least one root $a \in \mathbb{Z}_p$ (otherwise the claim holds). Then we can write $f(x) = (x-a)g(x)$ with g a polynomial over $\mathbb{Z}_p$ of degree $n-1$[32]. Since $\mathbb{Z}_p$ is a field (because p is prime), the roots of f are thus exactly a and the roots of g, making at most $n-1+1 = n$ solutions by the induction hypothesis. □

[30] If you don't remember how to show that $\mathbb{Z}_p^\times$ is a group, try to do it without looking it up!

[31] There are several beautiful proofs of this theorem, two of my favorites are: (i) The number of words with p letters from an alphabet of size a such that not all letters are the same is $a^p - a$; these can be partitioned in subsets of size p corresponding to words that are cyclic shifts of one another, and thus p divides $a^p - a$. (ii) Since $(x+y)^p \equiv x^p + y^p \bmod p$ (you can see this via the Binomial Theorem), Fermat's little Theorem can be shown by indiction on a and the observation that $(a+1)^p \equiv a^p + 1^p \equiv a^p + 1 \bmod p$.

[32] This follows by the Euclidean division of polynomials.

C.3 Wilson's theorem

Theorem C.5 (*Wilson's theorem*)

Let $p \geq 2$ be prime. Then

$$(p-1)! \equiv -1 \bmod p.$$

Note that this is identity is actually equivalent to p being prime.

Proof of Theorem C.5 – By Theorem C.4, the only solutions to $x^2 \equiv 1 \bmod p$ are $x \equiv \pm 1 \bmod p$. Therefore, we can form the $(p-3)/2$ pairs $\{a, a^{-1}\}$ for $a \in \mathbb{Z}_p^\times \backslash \{-1, 1\}$, which are pairwise disjoint. Since all elements of $\mathbb{Z}_p^\times \backslash \{-1, 1\}$ fall into such a pair, we have

$$(p-1)! = \prod_{a \in \mathbb{Z}_p^\times} a \equiv 1 \times (-1) \times (1)^{(p-3)/2} \bmod p \equiv -1 \bmod p.$$

□

C.4 $\mathbb{Z}_p^\times$ is a cyclic group

We now completely characterize the orders of the elements of $\mathbb{Z}_p^\times$. We need to introduce Euler's function:

Definition C.6 (*Euler's totient function*)

Euler's function ϕ is the function from $\mathbb{N}_{>0}$ to $\mathbb{N}_{>0}$ that maps each integer $n \geq 1$ with the number of $m \in \{1, \cdots, n\}$ such that m and n are coprime.

In particular $\phi(n) \geq 1$, and by definition $\phi(p) = p - 1$ if and only if p is prime. We moreover have the following property of Euler's function:

Proposition C.7

For any $n \geq 1$ we have

$$\sum_{d|n} \phi(d) = n.$$

Proof of Proposition C.7 – Fix $n \geq 1$. For any $d|n$, we denote $A(d) \coloneqq \{k \in [1, n] : \gcd(k, n) = d\}$. Note that $k \in A(d) \Leftrightarrow k = dl$ for some $l \in [1, n/d]$ which is coprime with n/d. Therefore $|A(d)| = \phi(n/d)$. Moreover, the sets $\{A(d) \subseteq [1, n] : d|n\}$ are pairwise disjoint and their union is $[1, n]$. Therefore we have

$$\sum_{d|n} \phi(n/d) = \sum_{d|n} \phi(d) = n,$$

since if d ranges over the divisors of n, so does n/d. □

We can now prove the main result of this appendix:

Lemma C.8 (*Order of elements of* $\mathbb{Z}_p^\times$)

Let $p \geq 2$ be a prime number, and $d \geq 1$ such that $d|(p-1)$. There are exactly $\phi(d)$ elements in $\mathbb{Z}_p^\times$ with order d.

In particular $\mathbb{Z}_p^\times$ is what we call a *cyclic group*, i.e. there is an element with order $p-1$ (actually $\phi(p-1)$ of them), that is an element whose powers generate the whole group!

Proof of Lemma C.8 – If $a \in \mathbb{Z}_p^\times$ has order $d|(p-1)$, then it is a root of the polynomial

$$x^d - 1 \equiv 0 \bmod p.$$

By Theorem C.4, there are at most d solutions to this equation, and since $a, a^2, \cdots, a^d = 1$ are all distinct solutions, they form all the solutions. Therefore the set $\{a^k, 1 \leq k \leq d\}$ must contain all the elements of order d. However, one checks easily that for all $k \in [1, d]$, $|a^k| = d \Leftrightarrow d$ and k are coprime. Therefore we have shown that if there is at least one element of order d, then there must be exactly $\phi(d)$ elements of order d. Moreover, since all elements of $\mathbb{Z}_p^\times$ have an order:

$$p - 1 = \sum_{d|(p-1)} \#\{a \in \mathbb{Z}_p^\times \text{ such that } |a| = d\}. \tag{99}$$

We have thus shown that the element inside the sum in the right hand side of eq. (99) can only be 0 or $\phi(d)$. However by Proposition C.7, we know

$$\sum_{d|(p-1)} \phi(d) = p - 1. \tag{100}$$

Therefore, the only possibility is that $\#\{a \in \mathbb{Z}_p^\times \text{ such that } |a| = d\} = \phi(d)$ for all $d|(p-1)$. □

D Group Testing lower bounds

Proof of Theorem 22.4 Fix a u such that $0 < u < \frac{T}{2}$; later it will be fixed to $u := \left\lceil (T-k)/\binom{k-1}{2} \right\rceil$. We start by constructing a few auxiliary family of sets. Let

$$\mathcal{A}_0 = \{A \in \mathcal{A} : |A| < u\},$$

and let $\mathcal{A}_1 \subseteq \mathcal{A}$ denote the family of sets in $\mathcal{A}$ that contain their own unique u-subset,

$$\mathcal{A}_1 := \left\{A \in \mathcal{A} : \exists F \subseteq A : |F| = u \text{ and, for all other } A' \in \mathcal{A},\ F \not\subseteq A'\right\}.$$

We will procede by giving an upper bound to $\mathcal{A}_0 \cup \mathcal{A}_1$. For that, we will need a couple of auxiliary family of sets. Let $\mathbb{F}$ denote the family of sets F in the definition of $\mathcal{A}_1$. More precisely,

$$\mathbb{F} := \left\{F \in [T] : |F| = u \text{ and } \exists! A \in \mathcal{A} : F \subseteq A\right\}.$$

By construction $|\mathcal{A}_1| \leq |\mathbb{F}|$

Also, let $\mathbb{B}$ be the family of subsets of $[T]$ of size u that contain an element of $\mathcal{A}_0$,

$$\mathbb{B} = \left\{B \subseteq [T] : |B| = u \text{ and } \exists A \in \mathcal{A}_0 \text{ such that } A \subseteq B\right\}.$$

We now prove that $|\mathcal{A}_0| \leq |\mathbb{B}|$. Let $\mathbb{B}'$ denote the family of subsets of $[T]$ of size u that are not in $\mathbb{B}$,

$$\mathbb{B}' = \left\{B' \subseteq [T] : |B'| = u \text{ and } B' \notin \mathbb{B}\right\}.$$

By construction of $\mathcal{A}_0$ and $\mathbb{B}$, no set in $\mathbb{B}'$ contains a set in $\mathcal{A}_0$ nor does a set in $\mathcal{A}_0$ contain a set in $\mathbb{B}'$. Also, both $\mathcal{A}_0$ and $\mathbb{B}'$ are antichains (or Sperner family), meaning that no pair of sets in each family contains each other. This implies that $\mathcal{A}_0 \cup \mathbb{B}'$ is an antichain containing only sets with u or less elements. The Lubell-Yamamoto-Meshalkin inequality [Yam54] directly implies that (as long as $u < \frac{T}{2}$) the largest antichain whose sets contain at most u elements is the family of subsets of $[T]$ of size u. This means that

$$|\mathcal{A}_0| + |\mathbb{B}'| = |\mathcal{A}_0 \cup \mathbb{B}'| \leq \binom{T}{u} = |\mathbb{B} \cup \mathbb{B}'| = |\mathbb{B}| + |\mathbb{B}'|.$$

This implies that $|\mathcal{A}_0| \leq |\mathbb{B}|$.

Because $\mathcal{A}$ satisfies the k-disjunct property, no two sets in $\mathcal{A}$ can contain eachother. This implies that the families $\mathbb{B}$ and $\mathbb{F}$ of sets of size u are disjoint which implies that

$$|\mathcal{A}_0 \cup \mathcal{A}_1| = |\mathcal{A}_0| + |\mathcal{A}_1| \leq |\mathbb{B}| + |\mathbb{F}| \leq \binom{T}{u}.$$

Let $\mathcal{A}_2 := \mathcal{A} \setminus (\mathcal{A}_0 \cup \mathcal{A}_1)$. We want to show that if $A \in \mathcal{A}_2$ and $A_1, \ldots, A_j \in \mathcal{A}$ we have

$$\left|A \setminus \bigcup_{i=1}^{j} A_i\right| > u(k-j). \tag{101}$$

This is readily shown by noting that if (101) did not hold then one could find $B_{j+1}, \ldots, B_k$ subsets of A of size t such that $A \setminus \bigcup_{i=1}^{j} A_i \subseteq \bigcup_{i=j+1}^{k} B_i$. Since A as no unique subsets of size t there must exist $A_{j+1}, \ldots, A_k \in \mathcal{A}$ such that $B_i \subseteq A_i$ for $i = j+1, \ldots, k$. This would imply that $A \subseteq \bigcup_{i=1}^{k} A_i$ which would contradict the k-disjunct property.

If $|\mathcal{A}_2| > k$ then we can take $A_0, A_1, \ldots, A_k$ distinct elements of $\mathcal{A}_2$. For this choice and any $j = 0, \ldots, k$

$$\left|A_j \setminus \bigcup_{0 \leq i < j} A_i\right| \geq 1 + u(k-j).$$

This means that

$$\left|\bigcup_{j=0}^{k} A_j\right| = \sum_{j=0,\ldots,k} \left|A_j \setminus \bigcup_{0\le i<j} A_i\right| \ge \sum_{j=0,\ldots,k} [1 + u(k-j)] = 1 + k + u\binom{k+1}{2}.$$

Since all sets in $\mathcal{A}$ are subsets of $[T]$ we must have $1 + k + u\binom{k+1}{2} \le \left|\bigcup_{j=0}^{k} A_j\right| \le T$. On the other hand, taking

$$u := \left\lceil (T-k)/\binom{k+1}{2} \right\rceil$$

gives a contradition (note that this choice of u is smaller than $\frac{T}{2}$ as long as $k > 2$). This implies that $|\mathcal{A}_2| \le k$ which means that

$$n = |\mathcal{A}| = |\mathcal{A}_0| + |\mathcal{A}_1| + |\mathcal{A}_2| \le k + \binom{T}{u} = k + \binom{T}{\left\lceil (T-k)/\binom{k+1}{2} \right\rceil}.$$

This means that

$$\log n \le \log\left(k + \binom{T}{\left\lceil (T-k)/\binom{k+1}{2} \right\rceil}\right) = O\left(\frac{T}{k^2}\log k\right),$$

which concludes the proof of the theorem. □

We essentially borrowed the proof of Theorem 22.4 from [Fur96]. We warn the reader however that the notation in [Fur96] is drasticly different than ours, T corresponds to the number of people and n to the number of tests.

E Some Coding Theory and the proof of Theorem 22.5

In this section we (very) briefly introduce error-correcting codes and use Reed-Solomon codes to prove Theorem 22.5. We direct the reader to [GRS15] for more on the subject.

Lets say Alice wants to send a message to Bob but they can only communicate through a channel that erases or replaces some of the letters in Alice's message. If Alice and Bob are communicating with an alphabet Σ and can send messages with lenght N they can pre-decide a set of allowed messages (or codewords) such that even if a certain number of elements of the codeword gets erased or replaced there is no risk for the codeword sent to be confused with another codeword. The set C of codewords (which is a subset of Σ^N) is called the codebook and N is the blocklenght.

If every two codewords in the codebook differs in at least d coordinates, then there is no risk of confusion with either up to $d-1$ erasures or up to $\lfloor \frac{d-1}{2} \rfloor$ replacements. We will be interested in codebooks that are a subset of a finite field, meanign that we will take Σ to be $\mathbb{F}_q$ for q a prime power and C to be a linear subspace of $\mathbb{F}_q$.

The dimension of the code is given by

$$m = \log_q |C|,$$

and the rate of the code by

$$R = \frac{m}{N}.$$

Given two code words c_1, c_2 the Hamming distance $\Delta(c_1, c_2)$ is the number of entries where they differ. The distance of a code is defined as

$$d = \min_{c_1 \neq c_2 \in C} \Delta(c_1, c_2).$$

For linear codes, it is the same as the minimum weight

$$\omega(C) = \min_{c \in C \setminus \{0\}} \Delta(c).$$

We say that a linear code C is a $[N, m, d]_q$ code (where N is the blocklenght, m the dimension, d the distance, and $\mathbb{F}^q$ the alphabet.

One of the main goals of the theory of error-correcting codes is to understand the possible values of rates, distance, and q for which codes exist. We simply briefly mention a few of the bounds and refer the reader to [GRS15]. An important parameter is given by the entropy function:

$$H_q(x) = x \frac{\log(q-1)}{\log q} - x \frac{\log x}{\log q} - (1-x) \frac{\log(1-x)}{\log q}.$$

- Hamming bound follows essentially by noting that if a code has distance d then balls of radius $\lfloor \frac{d-1}{2} \rfloor$ centered at codewords cannot intersect. It says that

$$R \leq 1 - H_q\left(\frac{1}{2}\frac{d}{N}\right) + o(1)$$

- Another particularly simple bound is Singleton bound (it can be easily proven by noting that the first $n+d+2$ of two codewords need to differ in at least 2 coordinates)

$$R \leq 1 - \frac{d}{N} + o(1).$$

There are probabilistic constructions of codes that, for any $\epsilon > 0$, satisfy

$$R \geq 1 - H_q\left(\frac{d}{N}\right) - \epsilon.$$

This means that R^* the best rate achievable satisties

$$R^* \geq 1 - H_q\left(\frac{d}{N}\right), \tag{102}$$

known as the Gilbert–Varshamov (GV) bound [Gil52, Var57]. Even for $q = 2$ (corresponding to binary codes) it is not known whether this bound is tight or not, nor are there deterministic constructions achieving this Rate.

E.0.1 The proof of Theorem 22.5

Reed-Solomon codes [RS60] are $[n, m, n - m + 1]_q$ codes, for $m \leq n \leq q$. They meet the Singleton bound, the drawback is that they have very large q ($q > n$). We'll use their existence below.

Proof of Theorem 22.5

We will construct a family $\mathcal{A}$ of sets achieving the upper bound in Theorem 22.5. We will do this by using a Reed-Solomon code $[q, m, q-m+1]_q$. This code has q^m codewords. To each codework c we will correspond a binary vector a of length q^2 where the i-th q-block of a is the indicator of the value of $c(i)$. This means that a is a vector with exactly q ones (and a total of q^2 entries)[33]. We construct the family $\mathcal{A}$ for $T = q^2$ and $n = q^m$ (meaning q^m subsets of $[q^2]$) by constructing, for each codeword c, the set of non-zero entries of the corresponding binary vector a.

These sets have the following properties,

$$\min_{j \in [n]} |A_j| = q,$$

and

$$\max_{j_1 \neq j_2 \in [n]} |A_{j_1} \cap A_{j_2}| = q - \min_{c_1 \neq c_2 \in C} \Delta(c_1, c_2) \leq q - (q - m + 1) = m - 1.$$

This readily implies that $\mathcal{A}$ is k-disjunct for

$$k = \left\lfloor \frac{q-1}{m-1} \right\rfloor,$$

because the union of $\left\lfloor \frac{q-1}{m-1} \right\rfloor$ sets can only contain $(m-1)\left\lfloor \frac{q-1}{m-1} \right\rfloor < q$ elements of another set.

Now we pick $q \approx 2k\frac{\log n}{\log k}$ (q has to be a prime but there is always a prime between this number and its double by Bertrand's postulate (see [**?**] for a particularly nice proof)). Then $m = \frac{\log n}{\log q}$ (it can be taken to be the ceiling of this quantity and then n gets updated accordingly by adding dummy sets).

This would give us a family (for large enough parameters) that is k-disjunct for

$$\begin{aligned}
\left\lfloor \frac{q-1}{m-1} \right\rfloor &\geq \left\lfloor \frac{2k\frac{\log n}{\log k} - 1}{\frac{\log n}{\log q} + 1 - 1} \right\rfloor \\
&= \left\lfloor 2k\frac{\log q}{\log k} - \frac{\log q}{\log n} \right\rfloor \\
&\geq k.
\end{aligned}$$

Noting that

$$T \approx \left(2k\frac{\log n}{\log k}\right)^2.$$

concludes the proof. □

[33] This is precisely the idea of code concatenation [GRS15]

References

[ALMT14] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference, available online*, 2014.

[Alo86] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6:83–96, 1986.

[AM85] N. Alon and V. Milman. Isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory*, 38:73–88, 1985.

[Bac21] Francis Bach. Learning theory from first principles. *Online version*, 2021.

[Ban16] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. *Available online at: http://www.cims.nyu.edu/~bandeira/TenLecturesFortyTwoProblems.pdf*, 2016.

[BBD+25] Afonso S. Bandeira, Jarosław Błasiok, Daniil Dmitriev, Ulysse Faure, Anastasia Kireeva, and Dmitriy Kunisky. The lovász number of random circulant graphs. *SAMPTA 2025*, 2025.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[BFMW13] Afonso S Bandeira, Matthew Fickus, Dustin G Mixon, and Percy Wong. The road to deterministic matrices with the restricted isometry property. *Journal of Fourier Analysis and Applications*, 19(6):1123–1149, 2013.

[BKMR25] Afonso S. Bandeira, Anastasia Kireeva, Antoine Maillard, and Almut Rödder. Randomstrasse101: Open problems of 2024, 2025.

[BM22] Afonso S. Bandeira and Antoine Maillard. Mathematics of signals, networks, and learning learning, spring 2023. Available at: https://people.math.ethz.ch/~abandeira//msnl_spring_2023.pdf, 2022.

[BMM17] Afonso S. Bandeira, Dustin G. Mixon, and Joel Moreira. A conditional construction of restricted isometries. *International Mathematics Research Notices*, 2017(2):372–381, 2017.

[BSS23] A. S. Bandeira, A. Singer, and T. Strohmer. Mathematics of data science. Available at: https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf, 2023.

[BZ22] Afonso S. Bandeira and Nikita Zhivotovskiy. Mathematics of machine learning. Available at: https://people.math.ethz.ch/~abandeira/Math_of_ML_Lecture_Notes2021.pdf, 2022.

[CFM+13] Peter G. Casazza, Matthew Fickus, Dustin G. Mixon, Jesse Peterson, and Ihar Smalyanau. Every hilbert space frame has a naimark complement. *Journal of Mathematical Analysis and Applications*, 406(1):111–119, 2013.

[Che70] J. Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. *Problems in analysis (Papers dedicated to Salomon Bochner, 1969), pp. 195–199. Princeton Univ. Press*, 1970.

[Chr16] Ole Christensen. *An Introduction to Frames and Riesz Bases*. Birkhäuser, 2016.

[Chu10] F. Chung. Four proofs for the cheeger inequality and graph partition algorithms. *Fourth International Congress of Chinese Mathematicians, pp. 331–349*, 2010.

[CRPW12] V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[CT05] E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[Dor43] R. Dorfman. The detection of defective members of large populations. 1943.

[DVPS14] A. G. D'yachkov, I. V. Vorob'ev, N. A. Polyansky, and V. Y. Shchukin. Bounds on the rate of disjunctive codes. *Problems of Information Transmission*, 2014.

[FM15] Matthew Fickus and Dustin G Mixon. Tables of the existence of equiangular tight frames. *arXiv preprint arXiv:1504.00253*, 2015.

[FR13] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhauser, 2013.

[Fur96] Z. Furedia. On r-cover-free families. *Journal of Combinatorial Theory, Series A*, 1996.

[Gil52] E. N. Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31:504–522, 1952.

[GRS15] V. Guruswami, A. Rudra, and M. Sudan. *Essential Coding Theory*. Available at: http://www.cse.buffalo.edu/faculty/atri/courses/coding-theory/book/, 2015.

[Gub18] John A Gubner. Derivation of the Fourier inversion formula, Bochner's theorem, and Herglotz's theorem, 2018.

[GZ84] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[HP21] Brandon Hanson and Giorgis Petridis. Refined estimates concerning sumsets contained in the roots of unity. *Proceedings of the London Mathematical Society*, 122(3):353–358, 2021.

[Kat04] Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.

[Kol29] A Kolmogoroff. Über das gesetz des iterierten logarithmus. *Mathematische Annalen*, 101(1):126–135, 1929.

[Kow21] Emmanuel Kowalski. *An Introduction to Probabilistic Number Theory*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2021.

[Kun24] D. Kunisky. Spectral pseudorandomness and the road to improved clique number bounds for paley graphs. *Experimental Mathematics*, 2024.

[Lan67] HJ Landau. Sampling, data transmission, and the nyquist rate. *Proceedings of the IEEE*, 55(10):1701–1706, 1967.

[Llo82] S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theor.*, 28(2):129–137, 1982.

[Mix] D. G. Mixon. Short, Fat matrices BLOG.

[Mix12] Dustin G. Mixon. Sparse signal processing with frame theory. *PhD Thesis, Princeton University, also available at arXiv:1204.5958[math.FA]*, 2012.

[RS60] I. S. Reed and G. Solomon. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics (SIAM)*, 8(2):300–304, 1960.

[Sau72] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[She72] Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

[Spi25] Daniel A. Spielman. Spectral and algebraic graph theory. *available at:* http://cs-www.cs.yale.edu/homes/spielman/sagt/sagt.pdf, 2025.

[SS03] Elias M. Stein and Rami Shakarchi. *Fourier Analysis: An Introduction*. Princeton Lectures in Analysis, Princeton University Press, 2003.

[Var57] R. R. Varshamov. Estimate of the number of signals in error correcting codes. *Dokl. Acad. Nauk SSSR*, 117:739–741, 1957.

[VB04] L. Vanderberghe and S. Boyd. *Convex Optimization*. Cambridge University Press, 2004.

[VC71] Vladimir Naumovich Vapnik and Aleksei Yakovlevich Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279, 1971.

[Vek21] Bogdan Veklych. A minimalist proof of the law of quadratic reciprocity. *arXiv preprint arXiv:2106.08121*, 2021.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[VPG15] Vladimir Vovk, Harris Papadopoulos, and Alexander Gammerman. *Measures of Complexity*. Springer, 2015.

[Wel74] Lloyd Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information theory*, 20(3):397–399, 1974.

[Yam54] K. Yamamoto. Logarithmic order of free distributive lattice. *Journal of the Mathematical Society of Japan*, 6:343–353, 1954.